NRC G/T

THE NATIONAL RESEARCH CENTER ON THE GIFTED AND TALENTED

The University of Connecticut
The University of Georgia
The University of Virginia
Yale University

An Analysis of the Research on Ability Grouping: Historical and Contemporary Perspectives

James A. Kulik, Ph.D.
The University of Michigan
Ann Arbor, Michigan

February 1992
Number 9204

ABILITY GROUPING

RESEARCH-BASED DECISION MAKING SERIES

An Analysis of the Research on Ability Grouping: Historical and Contemporary Perspectives

James A. Kulik, Ph.D.
The University of Michigan
Ann Arbor, Michigan

February 1992
Number 9204

ABILITY GROUPING

RESEARCH-BASED DECISION MAKING SERIES

THE NATIONAL RESEARCH CENTER ON THE GIFTED AND TALENTED

The National Research Center on the Gifted and Talented (NRC/GT) is funded under the Jacob K. Javits Gifted and Talented Students Education Act, Office of Educational Research and Improvement, United States Department of Education.

The Directorate of the NRC/GT serves as the administrative unit and is located at The University of Connecticut.

The participating universities include The University of Georgia, The University of Virginia, and Yale University, as well as a research unit at The University of Connecticut.

The University of Connecticut
Dr. Joseph S. Renzulli, Director
Dr. E. Jean Gubbins, Assistant Director

The University of Connecticut
Dr. Francis X. Archambault, Associate Director

The University of Georgia
Dr. Mary M. Frasier, Associate Director

The University of Virginia
Dr. Carolyn M. Callahan, Associate Director

Yale University
Dr. Robert J. Sternberg, Associate Director

Copies of this report are available from:
NRC/GT
The University of Connecticut
362 Fairfield Road, U-7
Storrs, CT 06269-2007

Research for this report was supported under the Javits Act Program (Grant No. R206R00001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. Grantees undertaking such projects are encouraged to express freely their professional judgement. This report, therefore, does not necessarily represent positions or policies of the Government, and no official endorsement should be inferred.

Note to Readers...

All papers that are commissioned by The National Research Center on the Gifted and Talented for the Research-Based Decision Making Series may be reproduced in their entirety or in sections. All reproductions, whether in part or whole, should include the following statement:

Research for this report was supported under the Javits Act Program (Grant No. R206R00001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. Grantees undertaking such projects are encouraged to express freely their professional judgement. This report, therefore, does not necessarily represent positions or policies of the Government, and no official endorsement should be inferred.

This document has been reproduced with the permission of The National Research Center on the Gifted and Talented.

If sections of the papers are printed in other publications, please forward a copy to:

The National Research Center on the Gifted and Talented
The University of Connecticut
362 Fairfield Road, U-7
Storrs, CT 06269-2007

About the Author....

Dr. James A. Kulik is a Research Scientist at the University of Michigan Center for Research on Learning and Teaching. At Michigan he has taught psychology courses, directed university-wide programs of teacher evaluation and student testing and placement, and conducted research in education. He is the author of the book *Undergraduate Education in Psychology* and of numerous articles. His major interest in recent years has been the quantitative integration of research findings in education, a topic covered in his coauthored monograph *Meta-Analysis in Educational Research*, published in the *International Journal of Educational Research*.

An Analysis of the Research on Ability Grouping: Historical and Contemporary Perspectives

James A. Kulik
The University of Michigan
Ann Arbor, Michigan

ABSTRACT

Researchers have struggled for decades to find answers to questions about ability grouping. Does anyone benefit from it? Who benefits most? Does grouping harm anyone? How? How much? Why? Reviewers of the research still disagree about the answers. For every reviewer who has concluded that grouping is helpful, another has concluded that it is harmful.

Today, however, reviewers are using statistical methods to organize and interpret the research literature on grouping, and they are more hopeful than ever before of coming to a consensus on what the research says. They have painstakingly catalogued the features and results of hundreds of studies, and with the help of new statistical methods, they are now drawing a composite picture of the studies and findings on grouping. In his 1976 presidential address to the American Educational Research Association, Glass coined the term meta-analysis to describe this statistical approach to reviewing research literature.

Meta-analytic reviews have already shown that the effects of grouping programs depend on their features. Some grouping programs have little or no effect on students; other programs have moderate effects; and still other programs have large effects. The key distinction is among (a) programs in which all ability groups follow the same curriculum; (b) programs in which all groups follow curricula adjusted to their ability; and (c) programs that make curricular and other adjustments for the special needs of highly talented learners.

Programs that entail only minor adjustment of course content for ability groups usually have little or no effect on student achievement. In some grouping programs, for example, school administrators assign students by test scores and school records to high, middle, and low classes, and they expect all groups to follow the same basic curriculum. The traditional name for this approach is XYZ grouping. Pupils in middle and lower classes in XYZ programs learn the same amount as equivalent pupils do in mixed classes. Students in the top classes in XYZ programs outperform equivalent pupils from mixed classes by about one month on a grade-equivalent scale. Self-esteem of lower aptitude students rises slightly and self-esteem of higher aptitude students drops slightly in XYZ classes.

Grouping programs that entail more substantial adjustment of curriculum to ability have clear positive effects on children. Cross-grade and within-class programs, for example, provide both grouping and curricular adjustment in reading and arithmetic for elementary school pupils. Pupils in such grouping programs outperform equivalent control students from mixed-ability classes by two to three months on a grade-equivalent scale.

Programs of enrichment and acceleration, which usually involve the greatest amount of curricular adjustment, have the largest effects on student learning. In typical evaluation

studies, talented students from accelerated classes outperform nonaccelerates of the same age and IQ by almost one full year on achievement tests. Talented students from enriched classes outperform initially equivalent students from conventional classes by 4 to 5 months on grade equivalent scales.

An Analysis of the Research on Ability Grouping: Historical and Contemporary Perspectives

James A. Kulik
The University of Michigan
Ann Arbor, Michigan

EXECUTIVE SUMMARY

Research literature on ability grouping used to be like the Bible. You could quote from it to support almost any view. Both advocates and opponents of grouping cited it to back their positions. Now, reviewers are using new statistical methods to organize and summarize the literature on grouping, and its message has become clearer.

The reviewers have painstakingly catalogued the features and results of hundreds of studies, and with the help of new statistical methods, they are drawing a composite picture of the studies and their findings. Their reviews have already shown that certain approaches to grouping consistently produce positive effects on children while other programs seldom produce measurable effects.

These scientific analyses of the research literature could hardly be more timely. In school systems around the country, parents, teachers, and school administrators are wrestling as never before with questions about ability grouping. They have read Jeannie Oakes's book *Keeping Track,* and they know the arguments against grouping. They also know the arguments in favor of the practice. Now, they want dependable answers. What does the research say?

What Is Ability Grouping?

Ability grouping, or homogeneous grouping, is the separation of same-grade school children into groups or classes that differ markedly in school aptitude. School personnel usually separate children into ability groups on the basis of test scores and school records. Ability grouping plays a key role in a number of school programs: separate classes in elementary schools for children of high, middle, and low aptitude; single-subject grouping in high school; cross-grade grouping for reading or arithmetic; special classes for the gifted and talented; and within-class grouping.

Writers on educational issues usually distinguish between ability grouping and tracking. They reserve the term *tracking,* or *curricular tracking,* for high school programs in which students choose, on the basis of their educational and job goals, either college-preparatory, general, or vocational classes in English, mathematics, and other subjects. Such tracking differs from ability grouping in several respects. First, curricular

tracking occurs only in high schools, whereas ability grouping can and does occur at all levels of education. Second, students themselves make course decisions in tracking programs, whereas preferences of pupils and their parents seldom play a role in placement into ability groups. Third, same-grade courses in different curricular tracks have different curricular objectives, whereas all ability-grouped classes in the same grade may have the same objectives.

The Art of Research Reviews

Researchers have been conducting controlled experiments on ability grouping for more than a half century. One of the first of these experiments took place in 1927 in Salt Lake City. At the beginning of the school year, a researcher identified two equivalent groups of elementary school children. Pupils in one group were separated by ability into homogeneous classes; the other group was assigned to mixed-ability classes. At the end of the school year, the researcher found that children from the homogeneous classes outperformed those from the mixed classes by about 2 months on a grade-equivalent scale. In the years that followed, hundreds of other researchers carried out similar experiments, and dozens of reviewers attempted to make sense of their findings.

The research reviewers, however, have painted at least four different pictures of the experimental results. Each of the pictures comes from a different era, and each reflects the educational concerns of its times. Each of the pictures also clashes with the other pictures in the set. Viewed together, the four portraits show that research reviewers sometimes see different things in the same studies. Although research experimentation is a science, research reviewing is too often a subjective art.

The original picture of the research comes down to us from the late 1920s when the mental testing movement was at its height in American education. Mental tests had just proven their value in the evaluation of recruits during World War I, and many mental testers expected even greater benefits from the use of tests for selection and placement of children in schools. Reviewers of the time shared the optimism about testing, and not surprisingly, they had positive things to say about ability grouping. Their most important conclusions, repeated in review after review in the early 1930s, was that grouping led to better school outcomes only when ability groups worked with methods and materials that suited their aptitude levels. The reviewers also noted that grouping programs had little or no effect when groups at all levels used the same methods and materials.

In the 1930s, John Dewey's philosophy of progressive education became an important influence on American schools, and with its rise, enthusiasm about grouping began to fade. Progressive educators held that the social spirit of the classroom did as much for children as formal instruction did, and they criticized grouping programs for fostering undemocratic feeling and traditional content teaching. Their reviews of the research on ability grouping focused on negative effects. Reviews of the time reported that students learned less and also declined in self-concept and leadership skills in grouped classes.

During the 1950s, the pendulum of opinion about grouping began to swing back. The United States and Russia were fighting a cold war for scientific and technological supremacy, and American schools were expected to contribute to the struggle by emphasizing academic and scientific excellence. Reviewers did their part by re-examining research results on grouping. The new reviews reported that higher aptitude youngsters made notable gains when taught in special enriched and accelerated classes. The reviewers reported that accelerated and enriched classes helped talented children academically and also seemed to have no detrimental effects on their social and emotional adjustment.

The civil rights movement of the 1960s inspired researchers to think more deeply about questions of educational equity, and it led ultimately to still another re-evaluation of grouping research. After the 1960s many reviewers reported seeing a different pattern in the research results on grouping. In *Keeping Track*, Jeannie Oakes expressed this newer point of view when she wrote that no one benefits from ability grouping and that children who are in the middle and lower groups clearly suffer a loss in achievement, academic motivation, and self-esteem.

Are any of these portraits accurate? Or do they each contain a bit of the truth? Until recently, there was no scientific way to answer such questions. Research reviews were the last word in research interpretation. When research reviewers disagreed, appeal to a higher authority was impossible.

Scientific Reviews of Research

The situation changed dramatically during the 1970s. In his 1976 presidential address to the American Educational Research Association, Gene V. Glass urged reviewers to abandon their subjective approach and to adopt instead rigorously scientific standards for research reviews. Glass's address had a powerful impact. It helped transform the art of research reviewing into a science.

Glass used the term *meta-analysis* to describe the new approach. Reviewers who use meta-analytic methods first locate studies of an issue by clearly specified, objective procedures. They then characterize the outcomes and features of these studies in quantitative terms. Finally, they use statistics to describe findings and to relate characteristics of the studies to outcomes. This meta-analytic approach helps reviewers to maintain objectivity and to describe precisely the benefits and losses associated with various educational alternatives.

Several research groups have carried out meta-analyses on grouping findings. Among the most comprehensive analyses are those carried out by Robert Slavin at Johns Hopkins University and those conducted by my research group at the University of Michigan. These meta-analyses show that different grouping programs produce different effects. Some programs have little or no effect on students, other programs have moderate effects, and still other programs have large effects. The key distinction is

among (a) programs in which all ability groups follow the same curriculum; (b) programs in which all groups follow curricula adjusted to their ability; and (c) programs that make curricular and other adjustments for the special needs of highly talented learners.

Grouping Without Curricular Adjustment

Some school administrators think that it is easier for teachers to teach and for learners to learn in classes where students resemble one another in learning rate. They therefore assign same-grade students to classes by aptitude. The high, middle, and low classes in many of the programs use the same text materials and follow the same basic course of study. The traditional name for this approach is *XYZ grouping*, but XYZ classes have also been called *multilevel*, *multitrack*, and *homogeneous classes*. Robert Slavin of Johns Hopkins University calls the approach *ability-grouped class-assignment*.

Although small school systems were experimenting with XYZ classes at the turn of the century, Detroit in 1919 became the first large city to introduce a formal XYZ plan. Teachers in the Detroit schools tested all children at the start of Grade 1 and placed them by test results into X, Y, and Z groups. The top 20 per cent went to the X classes, the middle 60 per cent to Y classes, and the bottom 20 per cent to Z classes. The X, Y, and Z groups studied from the same texts and followed the same course of study. This model became popular throughout the country both for all-day programs of grouping in elementary schools and for single-subject grouping in high schools. No other approach to grouping has been the subject of more research scrutiny over the years.

Our meta-analyses at Michigan covered 51 separate studies of XYZ classes, and the Johns Hopkins analyses covered 47 studies. Both analyses reached the same conclusion about lower and middle ability students: These students learn the same amount in XYZ and mixed classes. The evidence from the higher aptitude groups was less clear. Our meta-analyses at Michigan found that higher aptitude learners make slightly larger gains in XYZ programs. A higher aptitude student who gained 1.0 years on a grade-equivalent scale after a year in a mixed class would gain 1.1 years in an XYZ class. The Johns Hopkins meta-analysis suggested that gains for higher aptitude students were equal in XYZ and mixed classes.

Some of the studies of XYZ classes examined student self-concepts. Our meta-analysis showed that the average scores on self-esteem scales were nearly identical for students from XYZ and mixed classes. Nonetheless, XYZ classes had a small effect on student self-esteem. We found that self-esteem went up slightly for low-aptitude learners in XYZ programs, and it went down slightly for high-aptitude learners. Brighter children lost a little of their self-assurance when they were put into classes with equally talented children. Slower children gained a little in self-confidence when they were taught in classes with other slower learners.

Why were the effects of XYZ classes so small? The main problem with XYZ classes is probably their curricular uniformity. School personnel are usually careful in

placing children into high, middle, and low classes, but they seldom adjust the curriculum to the ability levels of the classes. For example, children in the high group in a Grade 5 program may be ready for work at the sixth grade level; children in the middle group are usually ready for work at the fifth grade level; and children in the low group may need remedial help to cover fifth grade material. But all groups work with the same materials and follow the same course of study in most XYZ classes. XYZ programs are thus programs of differential placement but not differential treatment.

Grouping With Curricular Adjustment

Unlike XYZ plans, programs of cross-grade and within-class grouping provide different curricula for children at different ability levels. Both group placement and curricula vary with student aptitude in these programs.

The best known approach to cross-grade grouping is the Joplin plan, which was first used during the 1950s for reading instruction in the Joplin, Missouri, elementary schools. During the hour reserved for reading in the Joplin schools, children in Grades 4, 5, and 6 broke into nine different groups that were reading at anything from the Grade 2 to Grade 9 level. The children went to their reading classes without regard to their regular grade placement but returned to their regular age-graded classrooms at the end of the hour. Almost all formal evaluations of cross-grade grouping involve the Joplin plan for reading instruction in elementary schools.

A popular model for within-class grouping of children in arithmetic was also developed in the 1950s. A teacher following the model would use test scores and school records to divide her class into three groups for their arithmetic lessons, and she would use textbook material from several grade levels to instruct the groups. The high group in Grade 6, for example, would use texts from Grades 6, 7, and 8; the middle group would use texts from Grades 5, 6, and 7; and the low group would use texts from Grades 4, 5, and 6. The teacher would present material to one group for approximately 15 minutes before moving on to another group. Other approaches to within-class grouping are possible, but almost all controlled evaluations examine within-class programs that follow this model.

Both the Michigan and Johns Hopkins meta-analyses found that cross-grade and within-class programs usually produce positive results. The Michigan analysis, for example, covered 14 studies of cross-grade grouping and 11 studies of within-class grouping. More than 80 per cent of the studies of each type reported positive results. The average gain attributable to cross-grade or within-class grouping was between 2 and 3 months on a grade equivalent scale. The typical pupil in a mixed-ability class might gain 1.0 years on a grade-equivalent scale in a year, whereas the typical pupil in a cross-grade or within-class program would gain 1.2 to 1.3 years. Effects were similar for high, middle, and low aptitude pupils.

Cross-grade and within-class programs appear to work because they provide different curricula for pupils with different aptitude. In cross-grade programs, students move up or down grades to ensure a match between their reading ability and their reading instruction. In within-class programs, teachers divide students into ability groups so that they can work on different materials with children of differing ability levels. Curriculum varies with student aptitude in these programs. The programs thus differ in an important respect from multilevel classes.

Special Accelerated and Enriched Classes

American education has a long tradition of offering special classes for students with special needs. Schools offer special classes for children who are physically handicapped, emotionally or socially maladjusted, lacking in proficiency in English, and so on. Many educators also look on gifted and talented children as learners with special needs. Schools have traditionally used two different approaches with such children: acceleration and enrichment.

The first classes devised especially for gifted and talented children were accelerated ones. The Cambridge Double Track Plan of 1891, for example, put bright children into special classes that covered the work of six years in four, and the special-progress classes of New York City, established in 1900, allowed bright pupils to complete the work of three years in two. Other school systems introduced other forms of acceleration early in the century, and by the 1920s accelerated instruction seemed to be established as the basic method for dealing with gifted school children.

By the 1920s, however, some educators began to question the wisdom of accelerating children through their school work. Their main concern was that accelerated programs might not meet children's emotional and social needs, whereas programs of enriched instruction might meet such needs. In a program of enrichment that Leta Hollingworth set up in the New York City schools in 1916, for example, gifted and talented children did not simply follow a telescoped regular curriculum. Instead, they spent about half of their school hours working on the prescribed curriculum, and about half pursuing enriching activities. In Hollingworth's class for seven- to nine-year olds, enriching activities included conversational French, biography, history of civilization, and a good deal of extra work in science, mathematics, English composition, and music.

Our meta-analysis covered 23 studies of acceleration. The studies compared the achievement of equivalent students in accelerated classes and nonaccelerated control classes. All of the studies examined moderate acceleration of a whole class of students rather than acceleration of individual children. In each of the comparisons involving students who were initially equivalent in age and intelligence, the accelerates outperformed the nonaccelerates. In the typical study, the average superiority for the accelerates was nearly one year on a grade-equivalent scale of a standardized achievement test.

Our meta-analysis also covered 25 studies of enriched classes for highly talented students. Twenty-two of the 25 studies found that talented students achieved more when they were taught in enriched rather than regular mixed-ability classes. In the average study, students in the enriched classes outperformed equivalent students in mixed classes by about 4 to 5 months. Children receiving enriched instruction gained 1.4 to 1.5 years on a grade-equivalent scale in the same period during which equivalent control children gained only 1.0 year.

Why do these classes have such strong effects? First, the adjustment in curriculum in accelerated and enriched classes is substantial because the children in these classes are unusually talented academically. Second, special resources are usually available for enriched and accelerated classes. The teachers of enriched and accelerated classes often have special training for work with gifted and talented students. Parents of youngsters in these classes sometimes band together in formal or informal networks to support their children. Special funding is sometimes available for these classes. Any of these resources could add to the success of accelerated and enriched classes.

What About Tracking?

Research reviewers have not conducted meta-analyses of findings on curricular tracking because almost no experimental studies are available on the topic. Instead of comparing tracked versus untracked high schools, researchers interested in tracking have compared student performance or teacher behaviors in high and low tracks. Although not without interest, such comparisons shed no light on the relative effectiveness of tracked versus untracked high schools.

Jeannie Oakes, in her book *Keeping Track*, uses research on ability grouping in her critique of tracking. Unfortunately, the findings that she cites come from studies of XYZ classes. Studies of XYZ classes are not directly relevant to the question of curricular tracking because XYZ classes follow a common curriculum whereas curricular tracks by definition do not. To evaluate adequately the effectiveness of high schools with tracks, we need controlled studies comparing the performance of initially equivalent students who were taught in tracked and untracked classes.

Conclusion

The questions that people ask about grouping are not easy to answer. Do children benefit from it? Who benefits most? Does grouping harm anyone? How? Why? The answers depend on the type of grouping program. Results differ in programs that (a) group students by aptitude but prescribe a common curriculum for all groups; (b) group students by aptitude and prescribe different curricula for the groups; and (c) place highly talented students into special enriched and accelerated classes that differ from other classes in both curricula and other resources. Benefits from the first type of program are

positive but very small. Benefits from the second type are positive and larger. Benefits from the third type of program are positive, large, and important.

These results are relevant to Jeannie Oakes's call for the elimination of all forms of ability grouping from American schools. Meta-analytic evidence suggests that this proposed reform could greatly damage American education. Teachers, counselors, administrators, and parents should be aware that student achievement would suffer with the total elimination of all school programs that group students by aptitude.

The harm would be relatively small from the simple elimination of XYZ programs in which high, middle, and low classes cover the same basic curriculum. If schools replaced all their XYZ classes with mixed ones, the achievement level of higher aptitude students would fall slightly, but the achievement level of other students would remain about the same. If schools eliminated grouping programs in which all groups follow curricula adjusted to their ability, the damage would be greater, and it would be felt more broadly. Bright, average, and slow students would suffer academically from elimination of such programs. The damage would be greatest, however, if schools, in the name of de-tracking, eliminated enriched and accelerated classes for their brightest learners. The achievement level of such students falls dramatically when they are required to do routine work at a routine pace. No one can be certain that there would be a way to repair the harm that would be done if schools eliminated all programs of acceleration and enrichment.

Guidelines From Meta-analytic Studies of Ability Grouping

Guideline 1: Although some school programs that group children by ability have only small effects, other grouping programs help children a great deal. Schools should therefore resist calls for the wholesale elimination of ability grouping.

Research support: The effect of a grouping program depends on its features. It is important to distinguish among programs that (a) make curricular and other adjustments for the special needs of highly talented learners, (b) make curricular adjustments for several ability groups at a grade level, and (c) provide the same curriculum for all ability groups in a grade.

Guideline 2: Highly talented youngsters profit greatly from work in accelerated classes. Schools should therefore try to maintain programs of accelerated work.

Research support: Talented students from accelerated classes outperform nonaccelerates of the same age and IQ by almost one full year on the grade-equivalent scales of standardized achievement tests.

Guideline 3: Highly talented youngsters also profit greatly from an enriched curriculum designed to broaden and deepen their learning. Schools should therefore try to maintain programs of enrichment.

Research support: Talented students from enriched classes outperform control students from conventional classes by 4 to 5 months on grade-equivalent scales.

Guideline 4: Bright, average, and slow youngsters profit from grouping programs that adjust the curriculum to the aptitude levels of the groups. Schools should try to use ability grouping in this way.

Research support: Cross-grade and within-class programs are examples of programs that provide both grouping and curricular adjustment. Children from such grouping programs outperform control children from mixed classes by 2 to 3 months on grade-equivalent scales.

Guideline 5: Benefits are slight from programs that group children by ability but prescribe common curricular experiences for all ability groups. Schools should not expect student achievement to change dramatically with either establishment or elimination of such programs.

Research support: In XYZ grouping, all ability groups follow the same course of study. Middle and lower ability students learn the same amount in schools with and without XYZ classes. Higher ability students in schools with XYZ classes outperform equivalent students from mixed classes by about one month on a grade-equivalent scale.

References

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3-8.

Kulik, J. A., & Kulik, C.-L. C. (1991). Ability grouping and gifted students. In N. Colonel and G. A. Davis (Eds.), *Handbook of gifted education* (pp. 178-196). Boston: Allyn & Bacon.

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

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research, 57,* 293-336.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60,* 471-499.

Table of Contents

List of Tables

An Analysis of the Research on Ability Grouping: Historical and Contemporary Perspectives

James A. Kulik
The University of Michigan
Ann Arbor, Michigan

Introduction

In 1930 two researchers at the University of Minnesota, W. S. Miller and Henry J. Otto, published the first comprehensive review of experimental research on ability grouping. The 20 studies that they examined varied in quality and results, but the reviewers saw enough consistency in the findings to draw a positive conclusion. When grouping is done properly, they reported, it benefits children. One year later, however, in 1931, a researcher at Columbia University, Alice Keliher, reviewed the same experimental evidence and came to a different conclusion. Grouping, she wrote, is more likely to harm than help students. The evidence convinced her that grouping has negative effects on children's school achievement, self-concepts, and social outlook.

The controversy that thus began has gone on now for six decades. During the past 60 years, hundreds of researchers have carried out evaluation studies on the topic of ability grouping, and dozens of reviewers have tried to make sense of the findings. What did the reviewers conclude? Like Miller and Otto, some concluded that the evidence supports ability grouping. Like Keliher, others argued the opposite position. Some judged the research to be too contradictory for any conclusions.

Recently, educational researchers have developed objective, scientific methods for reviewing research findings, and reviewers are today applying these new methods to the accumulated findings on ability grouping. There is reason to believe, therefore, that reviewers may finally come to a consensus about what the research says. But the day of consensus has not yet arrived, and the debate continues. Basic questions still trouble educators. Does anyone benefit from ability grouping? Who benefits most? Is anyone harmed? How? Why?

The *Dictionary of Education* (1959) defines homogeneous ability grouping as "the classification of pupils for the purpose of forming instructional groups having a relatively high degree of similarity in regard to certain factors that affect learning" (p. 269). The definition is broad enough to encompass a variety of programs used in elementary and secondary schools: separate classes in elementary schools for children of high, middle, and low aptitude; cross-grade grouping; single-subject grouping; separate curricular tracks for high school students; special classes for the gifted and talented; and numerous methods of within-class grouping.

Writers disagree about the best names to give to such programs. Some older reviews refer to grouping plans by their originators or place of origin. Otto (1941), for example, describes such methods as the following: the Cambridge plan of 1893 of separating students into slower or faster tracks; the Santa Barbara concentric plan from the turn of the century of dividing students into A, B, and C sections; and such additional schemes as the Pueblo, Portland, Batavia, and North Denver plans. More recent reviews have attempted to classify grouping methods more systematically and to provide more descriptive names.

No one has yet produced an adequate taxonomy of schemes for grouping, however. Mort (1928) distinguished between *homogeneous grouping* and *ability grouping*. He used the first term to refer to formation of groups of children of similar aptitude or achievement, and the second to refer to separation of students for the purpose of adjusting the curriculum to the level of their abilities. Many reviewers consider this distinction to be an important one, but few use the terms *homogeneous grouping* and *ability grouping* in Mort's sense. Other reviewers have distinguished flexible from inflexible programs, between-class from within-class programs, comprehensive from single-subject programs. Because no one knows which characteristics are the key ones for understanding grouping effects, the creation of a Mendeleev's table of grouping schemes remains a task for the future.

Among the many grouping plans that have been proposed, a few stand out because they have been studied frequently by researchers:

1. **XYZ classes.** Students at a single grade level are divided into groups - often high, middle, and low groups - on the basis of ability level, and the groups are instructed in separate classrooms. Separation may be for the full day or for a single subject only. Reviewers have referred to this approach by such names as *ability-grouped class assignment* (Slavin, 1987), *multitrack grouping* (Miles, 1954), and *multilevel grouping* (J. Kulik & Kulik, in press).
2. **Cross-grade grouping.** Children from several grades who are at the same level of achievement in a subject are formed into groups, and the groups are then taught the subject in separate classrooms, without regard to the children's regular grade placement or age. Most cross-grade programs are elementary school programs in reading. Researchers and reviewers sometimes refer to this grouping method as the Joplin plan (Slavin, 1987).
3. **Within-class grouping.** A teacher forms ability groups within a single classroom and provides each group with instruction appropriate to its level of aptitude. This type of grouping has been used frequently for reading and arithmetic instruction in elementary schools. It is sometimes referred to as *intraclass grouping* (Petty, 1953).
4. **Accelerated classes for the gifted and talented.** Students who are high in aptitude in a subject receive instruction that allows them to proceed more rapidly through their schooling or to finish schooling at an earlier age than other students.
5. **Special enriched classes for the gifted and talented.** Students who are high in academic aptitude receive richer, more varied educational experiences than would be available to them in the regular curriculum for their age level. The instruction is usually, but not always, provided in a separate classroom. The special grouping may be for a full day or a single subject.

When used in the literature, the term *ability grouping* may refer to one, several, or all of these practices.

Reviewers of the literature have listed the pros and cons of ability grouping many times (e.g., Passow, 1962; Slavin, 1987; Turney, 1931). Most proponents claim that it benefits both teachers and students. Teachers do not have to contend with a wide range of individual differences in ability-grouped classes, and pupils do not have to cope with instruction that is above or beneath their level of comprehension. Proponents hold that teachers, therefore, find it easier to teach and students to learn in ability-grouped classes. Opponents argue, however, that just the opposite is true. They contend that teaching fast

or slow students requires special skills that many teachers lack, and they also claim that homogeneous grouping harms many students, especially middle and lower aptitude students, who may suffer a loss in self-esteem, academic motivation, and overall accomplishment when placed in the slower groups.

One of the first attempts to settle the controversy by something approximating controlled experimentation took place in the schools of Urbana, Illinois, in 1916, when Whipple compared the accomplishments of a mixed-ability class and a class consisting of gifted elementary school pupils (Whipple, 1919). In the 1920s and 1930s, researchers began carrying out more sophisticated studies of the effects of ability grouping, and reviewers began writing reviews of the accumulating literature. By the 1960s, however, a note of futility could be detected in the reviews. Thus, Passow (1962) gave his influential review the title "The maze of research on ability grouping." Heathers (1969) wrote in the fourth edition of the influential *Encyclopedia of Educational Research* that he was looking forward to reading an epitaph for research on ability grouping in the fifth edition. The reason for the disenchantment was not hard to find. Despite all the studies, reviewers were unable to agree on what conclusions to draw.

Nonetheless, there are good reasons for returning to the research evidence once again today. For one thing, social scientists have developed objective, scientific methods for research reviewing during the past decade (e.g., Cooper, 1984; Glass, McGaw, & Smith, 1981; J. Kulik & Kulik, 1989; Light & Pillemer, 1989; Rosenthal, 1984). These scientific methods were not available during the heyday of grouping research, and most reviewers, therefore, used unreliable, impressionistic methods in reviewing literature on the topic. With the new scientific methods now available, reviewers are finally in a good position to determine what the research actually says.

There is another good reason for once again turning our attention to research findings on ability grouping. Some educational researchers and reformers are today advocating unprecedented changes in school grouping practices. Oakes, for example, has called for the elimination of all forms of grouping, or complete *de-tracking*, of American schools (Oakes, 1985; Oakes & Lipton, 1990). She wants schools to eliminate not only XYZ grouping but also special classes for the gifted and talented, advanced placement classes, and every other arrangement that provides special instructional opportunities for students on the basis of achievement, aptitude, or interest. School systems throughout the country are responding to such calls, and programs that provide special educational opportunities for special groups are under threat everywhere. It has never seemed more important, therefore, to know what the research actually says about the effects of ability grouping.

My purpose here is to set out the evaluation evidence on grouping as clearly as I can. I present this evidence in the next three sections of this report. In the first section, I examine earlier reviews of research on ability grouping. Most of these reviews were written before objective methods had been developed for summarizing large bodies of research literature, and most of the reviews are, therefore, subjective in their interpretation of the research evidence. The reviews are important to examine, however, because they raise important questions about possible consequences of grouping. In the second section of the report, I describe how new scientific review procedures are giving us a better understanding of the effects of grouping programs. In the third and final section of the report, I present the detailed findings from the new analyses.

My major conclusion is that ability grouping usually helps higher aptitude students, but the amount of benefit depends on the design of the grouping program. Higher aptitude students benefit a small amount from XYZ classes in which there is little

formal adjustment of curriculum to ability level; they gain a moderate amount in cross-grade and within-class programs, in which the curriculum is typically adjusted to group ability; and they gain moderate-to-large amounts in special enriched and accelerated classes. Another major conclusion is that grouping programs have smaller effects on middle and lower aptitude learners, but the size of the effect again depends on program type. XYZ grouping has no real effect on the school achievement of middle and lower aptitude learners, but programs of cross-grade and within-class grouping have moderate positive effects on such students.

Reviews of Research During Four Eras

During the 100 years that schools have practiced ability grouping, educational priorities have shifted back and forth. In some eras, schools have stressed learning of basic skills; in others, emotional and social growth. In some periods, they have emphasized education of the gifted and talented; in others, education of the disadvantaged. Such shifts in emphases have affected educational research. Just as each era has its own educational philosophy, each era also has its own research studies and reviews.

Research and reviews on ability grouping have been influenced by four currents in education:

1. **The mental testing movement.** The earliest reviews of research on ability grouping were written in the 1920s and early 1930s when schools were first capitalizing on psychometric advances in measurement of intelligence and school achievement.
2. **Progressive education.** In the late 1930s and early 1940s, reviewers were influenced by John Dewey's ideas of progressive education. Reviews of this era question the value of standardized tests and instructional approaches that rely heavily on test results.
3. **Educational excellence.** Reviews in the 1950s reflected the intense interest of the times in the development of scientific potential in gifted and talented youngsters.
4. **Educational equity**. The civil rights movement of the 1960s led researchers to focus on educational equity and educational practices that contribute to the achievement of equity.

In the sections that follow, I examine some of the conclusions that were reached in representative reviews of these four periods.

Mental Testing Movement

Mental testing is now generally recognized as one of the influential scientific inventions of the 19th century, but mental tests did not begin to play a prominent role in American society until the Army Alpha test was developed during World War I. With Terman's revision of the Binet-Simon intelligence scales in 1916, the testing movement started to develop momentum in the schools. By the 1920s many schools were using tests not only to measure intelligence and achievement but also for selection and placement of children into classes.

Even before formal testing programs were set up in schools, educators were aware of how greatly children of the same age varied in their accomplishments. Ayres's

nationwide study of 1909, for example, showed that in a typical American classroom about one-third of the children were retarded, or too old for their grade level. But standardized testing provided even more graphic proof of pupil variability. The tests showed, for example, that some children in a typical middle-school class were capable of work at the twelfth-grade level, whereas, others were still performing at the third-grade level (Burr, 1931). No problem seemed more pressing to teachers in the 1920s than meeting the needs of such diverse children, and no solution seemed more appealing than grouping together for instruction students who were similar in ability. The new mental tests were the vehicle that made such grouping possible.

Miller and Otto's (1930) article provided one of the earliest comprehensive reviews of the results of such grouping. Their review listed results from 20 studies of grouping carried out in the years 1920 through 1929, but Miller and Otto felt that only 7 of the studies met basic criteria of scientific adequacy. Although the results of the studies varied, two of them (Burtt, Chassel, & Hatch, 1923; Dvorak & Rae, 1929) suggested that ability grouping leads to improved outcomes only when the grouping is accompanied by adaptation of methods and materials. Without such adaptation, Miller and Otto noted, performance of students from ability-grouped and mixed-ability classes is indistinguishable. Thus, Miller and Otto concluded that homogeneous classification by itself does not ensure improved performance, but it does lead to better performance when it is used properly as a means of adapting instruction to student aptitude. Miller and Otto noted that schools all too often used the same methods and materials at all levels of grouping and thus defeated its basic purpose.

Turney's (1931) review also distinguished sharply between studies in which suitable adaptations of method and materials were made and studies without such adaptations. Like Miller and Otto, Turney focused especially on the studies by Burtt et al. (1923) and Dvorak and Rae (1929). Both studies suggested that achievement is greater in homogeneous groups only when grouping is accompanied by suitable curricular adaptation. Turney also found 10 studies in which no special attempt was made to adjust method, content, or time. These studies reported a total of 29 comparisons of the effectiveness of homogeneous and mixed-ability classes. In 15 of the cases, the pupils in the homogeneous groups excelled those in control groups; in 4 cases, the homogeneous groups were inferior to the heterogeneous groups; and in 10 cases, results were indecisive. Turney felt that the homogeneous classes made a good showing even in these poor circumstances. In spite of the fact that no conscious attempt was made in any of the 10 studies to capitalize on the potential of ability grouping, the value of grouping seemed evident in half of the comparisons.

Cornell (1936) examined evidence on both cognitive and emotional effects of ability grouping. She concurred with Turney on cognitive effects. Where an effort is made to adapt instructional materials and methods to the needs of different levels, she concluded, achievement is better in homogeneous groups than in heterogeneous groups. She also noted that the evidence was less conclusive on social, emotional, and personality adjustment of pupils in ability-grouped classes. She found that educational researchers held a variety of opinions on whether grouping stigmatizes pupils, whether it makes children happier, whether it affects school citizenship, and on other related questions. Cornell's review of survey results from teachers, principals, parents, and children, however, led her to conclude that the burden of such evidence was in favor of ability grouping.

Miller and Otto, Cornell, Turney, and other reviewers of this period were writing during an age of faith in testing, and ability grouping was clearly a practice that fit the spirit of the age. The reviewers must have been tempted to gloss over the negative

findings on grouping, but to their credit they resisted the temptation. They noted correctly that studies of grouping produced both positive and negative findings. Their challenge was to find some common characteristic in the studies that produced the positive results. The reviewers met the challenge by drawing a distinction between studies of grouping per se and studies of grouping with differentiated instruction. Findings on grouping per se were variable, they concluded, but findings on grouping with differentiated instruction seemed to be consistently positive.

It is clear today that this conclusion, although plausible, was largely speculative. The reviewers could cite only a few studies to support their notion that differentiated instruction was the key ingredient that made grouping programs work. More and better studies were needed to confirm the hypothesis. Decades would pass, however, before researchers would carry out such studies because by the mid-1930s ability grouping was losing much of its attraction as a research area. By that time John Dewey and his colleagues had mounted a successful campaign against ability grouping, and research progress on the topic became temporarily stalled.

Influence of Progressive Education

A major emphasis of John Dewey's educational philosophy was on the need for social living in the classroom. Dewey thought that the social spirit of the classroom did as much for children as formal instruction did and that children learned as much from contact with one another as they did from textbooks and lectures. Schools could be improved, Dewey argued, only if they abandoned traditional concepts of content teaching and undemocratic social structures. Grouping plans could not make schools better, Dewey noted, because they were inextricably tied to traditional and undemocratic notions of teaching.

Dewey also thought that grouping plans were a real threat to his notion of educational reform. The grouping plans could divert attention away from the concepts of progressive education. Dewey was explicit about the effect that grouping plans might have on his own reform agenda:

> Instead of mixing up together a lot of pupils of different abilities we can divide them into a superior, a middle, and an inferior section. . . . It may turn out that the net result will be to postpone the day of a reform of education which will get us away from inferior, mean and superior mediocrities so as to deal with individualized mind and character. The movement is on a par with the movements to make instruction more efficient while retaining the notion of teaching which emphasizes the receptively docile mind instead of an inquiring and pioneering purpose. (Dewey, 1929, p. 482)

According to Dewey, ability grouping was a step in the wrong direction, and Dewey and his followers, therefore, marshaled their arguments and evidence against it.

One of Dewey's arguments was that the intelligence tests that played a large role in all grouping schemes were inherently inadequate. Dewey argued that the tests should not be used for grouping because they reflected only a limited conception of intelligence. The tests measured only formal classroom learning, and in Dewey's view, that was only a small part of what intelligence actually was. Dewey viewed intelligence not as a single thing but rather as a characteristic of many different behaviors. A child could be intelligent in reciting lessons, in fitting into a school administration, in influencing companions, and in a hundred other ways. It was pointless, therefore, to identify some

children as more intelligent than others on the basis of scores on intelligence tests and achievement batteries.

Burr (1931) provided a statistical analysis of differences in children's accomplishments that supported Dewey's contentions. The analysis covered achievement and intelligence test scores from approximately 3400 pupils in six cities where ability grouping was practiced. Burr found a great overlap in the achievement of students in different ability groups. About 80 per cent of the total grade-range of achievement was found in each group. Burr also noted that when groups were made non-overlapping in achievement in one subject, such as reading, they overlapped greatly in other subjects, such as arithmetic. Even when groups were made non-overlapping in one phase of a subject, such as arithmetic reasoning, they overlapped in other phases of the same subject, such as arithmetic computation. Burr's conclusion was unequivocal: Homogeneous grouping was impossible. Children were too variable and too inconsistent for grouping to work.

Keliher (1931) also based her analysis of grouping on Dewey's educational ideas, but she supported her arguments with evidence from a variety of fields. She used historical, statistical, and empirical analyses to build her case against grouping. Although her monograph was based on a Ph.D. dissertation, it was no small scholarly achievement. In his own article on homogeneous grouping, Symonds (1931) referred to Keliher's dissertation as one of the finest pieces of educational argumentation yet produced.

Like Burr, Keliher examined evidence on student variation in performance from subject to subject, and like him she reported that variation was great enough to make truly homogeneous grouping impossible. But she also argued that even if academic performance were more consistent, grouping children by rank on tests of academic ability would still be undesirable. Intelligence tests, she argued, measured only capacity for academic learning, not the capacity for social judgment, moral decision-making, or artistic contributions. Grouping was fundamentally impossible, Keliher concluded, because there was no way to measure the unique characteristics of children, and there was no way to classify children on the basis of these characteristics.

Keliher also examined the empirical evaluations that had accumulated on effects of ability grouping. She reviewed 14 studies of homogeneous grouping, 13 of which overlapped with Miller and Otto's study pool. Keliher reported that only 5 of the studies were experiments with control groups. She noted that in 4 of the studies, there was no advantage to grouping, and in 2 of the studies, there was a definite disadvantage to homogeneous grouping. Keliher concluded, therefore, that there was no advantage and some disadvantage for learning results in homogeneous groups. She suggested that the disadvantages might, in fact, be larger than they appeared to be because positive teacher expectations about grouping might have masked some of its unfavorable effects.

Keliher did not examine carefully the evidence that suggested to others that grouping produced positive effects when it was accompanied by curricular differentiation. For Keliher, the idea of providing differentiated instruction on the basis of test scores was repugnant. It smacked of unacceptable "Prussian class distinctions." She challenged the right of anyone to restrict a pupil's educational opportunities by placing the child into a group that received instruction different from that given to others. In addition, Keliher suggested that pupil attitudes of inferiority or superiority would follow on the heels of homogeneous grouping because homogeneous classes would foster a competitive rather than cooperative spirit.

Interpretations like those of Burr and Keliher influenced mainstream opinions, and by the late 1930s interest in grouping began to wane. Looking back at the history of ability grouping, Otto (1950) identified the 1920s and early 1930s as the years of peak interest. After 1935, researchers stopped studying the effects of ability grouping, Otto reported, and demographers stopped trying to determine the extent of its use. The interest of teachers and administrators, he wrote, "changed from the rather narrow issues involved in ability grouping to broader concerns for well-rounded development in which emotional, social, character, and personality development receive as much attention as scholastic development" (p. 367). He recommended that future research focus on what had long been Dewey's concern: The whole child. Thus, by 1950, thanks in large part to the efforts of Dewey's followers, ability grouping seemed to be an idea whose time had passed.

Era of Educational Excellence

In 1957 Russia launched the satellite Sputnik, and the pendulum of opinion about ability grouping began to swing back. The Soviet satellite cast a long shadow over American education. Ordinary citizens looked up, scratched their heads, and wondered how the Russians could have beaten the United States into space. School officials began to wonder whether American education might be to blame, and government officials asked what the government could do to help. Studies were begun, commissions were established, and gradually a consensus emerged. This country would have to do more to cultivate its pool of young intellectual talent. Schools would have to provide special opportunities for the gifted and talented. *Excellence* became the byword in education.

Special programs for the gifted and talented were, of course, nothing new in American education. Schools had been providing such programs since the turn of the century. Special rapid-advancement classes, for example, had been established in the New York City schools in 1900, and during the 1920s Leta Hollingworth had set up exemplary enriched classes for the gifted and talented (Tannenbaum, 1958). But such programs had played only a supporting role in American education. With Sputnik circling overhead, the programs moved to center stage. Educators scrambled to provide special opportunities for children of high ability, researchers measured program effects, and reviewers began to draw a new picture of the consolidated results.

Miles (1954) provided one of the first glimpses of the new picture. She first examined results of 4 studies in which children were accelerated in their school work either individually or in groups. The studies looked at effects of acceleration on school achievement, personality, and school attitudes, and each of the studies reported positive results. Miles also examined results from six studies of special enrichment classes for the gifted. Again, results of all studies were favorable. Miles cautioned, however, that too few studies were available for her to conclude that separate accelerated or enriched classes for the gifted were superior to separate provisions within regular classrooms. She concluded:

> Gifted children, like others, require adequate opportunity and stimulation, and this can be more or less successfully given with due planfulness under various systems, so long as diverse rates of advancement are permitted and an adequately enriched curriculum is maintained. If segregation is used, selection in terms of total personality assets and needs is certainly desirable. (p. 1032)

Passow (1958) also reviewed literature on effects of acceleration and enrichment. Passow's review covered 16 studies of special enriched classes for the gifted and talented.

Of the 16 studies, 5 were carried out at the elementary level, and 11 were conducted at the secondary and college levels. In addition, Passow reviewed 18 studies of the use of acceleration with the gifted and talented. Of the 18 studies, 5 were conducted at the elementary level, 4 at the secondary level, and 9 at the college level.

Passow found a good deal of consistency in the outcomes of the special enriched classes. He pointed out that comparative studies in both elementary and secondary schools demonstrated the beneficial effects of such classes on the academic, personal, and social growth of the gifted and talented. Passow noted that the near unanimity in findings on special classes contrasted strongly with the lack of unanimity in studies of XYZ classes. Passow's conclusions about programs of acceleration were equally favorable. He pointed out that the experimental evidence at all levels of education showed that gifted and talented students gained academically by acceleration. In addition, research into the effects of acceleration generally demonstrated no detrimental effects on the social and emotional adjustment of students.

Ekstrom's review (1961) covered studies of both ability-grouped classes and special classes for the gifted and talented. Ekstrom divided the studies of ability-grouped classes into two rough categories: those in which little or no effort was made to provide a differentiated curriculum and those in which there seemed to be an effort to differentiate instruction. Of 9 studies of ability-grouped classes without a differentiated curriculum, 2 favored the ability-grouped classes, 2 favored the mixed-ability classes, and 5 produced no differences or mixed results. Results of studies of ability-grouped classes with curricular differentiation were more consistent. Of 8 such studies, 6 favored the grouped classes, none favored mixed-ability classes, and 2 were inconclusive. Results of studies of special classes for the gifted and talented produced even more favorable results for homogeneous grouping. Of 14 such studies, 10 produced positive results, none produced negative results, and 4 produced mixed results. Ekstrom thought that the success of special classes was due to the greater amount of curricular adjustment in the classes.

Thus, three decades after Miller and Otto (1930) had suggested that the effectiveness of ability grouping depends on the amount of curricular adaptation that accompanies it, Ekstrom found enough data to give the hypothesis an adequate test. In studies of grouping per se, in which no clear effort was made to provide differentiated instruction, she found only small and inconsistent effects. In studies of programs in which grouping was accompanied by curricular differentiation, she found clearer effects. In studies of special classes for the gifted and talented, where curriculum was clearly adjusted to meet the special needs of the students, she found the clearest effects of all. Miller and Otto's speculation seemed to be completely supported.

Begle's (1975) review of grouping in math education provided some additional insights. Like virtually every other reviewer, Begle noted the inconsistency in overall findings on XYZ classes. Although some investigators reported positive findings, almost as many reported negative results, and most reported results that could be classified as neither positive nor negative. Begle found little difference in results of XYZ programs in which curriculum was meant to be adjusted and results of programs in which curriculum was not adjusted. Begle pointed out, however, that it was hard to determine how much curricular differentiation took place in any of the XYZ classes.

Begle's observations would have been unremarkable if they had stopped there, but unlike other reviewers, Begle went on to look systematically at the possibility that XYZ grouping had different effects on learners in different groups. He concluded that it did. Effects on students in the lower and middle tracks were negligible, but effects on those in

the upper tracks were positive. Begle concluded that the evidence is clear that XYZ classes can benefit higher ability students.

Begle also observed that results were positive for programs of within-class grouping in arithmetic in elementary schools. Three of the 8 studies of within-class grouping located by Begle reported significantly positive results, 5 reported no significant effects, and no study reported negative or mixed results. Begle noted that in all studies of within-class programs, the curriculum was clearly adapted to the ability level of the groups. Begle, like Ekstrom (1961), had found good evidence in support of Miller and Otto's (1930) differentiated curriculum hypothesis.

Begle's (1976) review of studies of special accelerated classes in mathematics for the gifted and talented fills out the picture of grouping effects. Begle's review asked two questions. First, do accelerated students learn more than comparable students who are not accelerated? Begle concluded that they did. In each of 7 controlled studies, accelerated students scored higher than comparable students who were not accelerated. Second, Begle asked whether the young accelerates who move to higher grades are able to keep up with older, equally talented nonaccelerates already in those grades. Begle found 12 studies on the issue. He concluded that accelerates usually do as well as talented, older students and when they do not, they seldom lag far behind.

Overall, the main findings of reviewers like Miles, Passow, Ekstrom, and Begle have stood up very well over time. No one has ever seriously challenged their conclusions about enriched and accelerated classes. These reviewers established once and for all that academically talented students who are placed into such classes benefit intellectually and do not suffer emotionally from the experience. Even opponents of ability grouping usually concede this point today. Where proponents and opponents of programs for talented students differ is on the importance of the point. Opponents of ability grouping argue that the facts established by reviewers like Passow, Ekstrom, and Begle are not as important as other more recently established findings on grouping. More recent findings are presented in the next section of this paper.

Emphasis on Educational Equity

The 1960s were watershed years in American education. The Supreme Court decision in Brown vs. the Board of Education in 1954 outlawed school segregation and ushered in an era of intense interest in the civil rights of all Americans. Equity became a major theme in American education, and researchers in education who had been preoccupied with the average child or the high ability child began to pay attention to disadvantaged and underprepared children. What effects did educational programs have on such children? Were some teaching practices unfair to them? What effects, for example, did grouping programs have on minority and disadvantaged children?

The possibility that ability grouping might have damaging effects on children from economically and culturally disadvantaged homes had been raised in the earliest reviews of grouping practices. Keliher (1931), for example, was very clear about the potential dangers. One of her major concerns was that grouping might further stigmatize children who had already been treated badly by society. But these concerns did not become central to reviews of grouping until the 1960s.

Eash (1961) expressed the new point of view in an article on generalizations from research on ability grouping. He first noted the "resurrection" of interest in ability grouping during the 1950s. He attributed this revival to endorsements given to ability

grouping by distinguished scholars concerned about excellence in the schools. Eash reported that he could not second their endorsements, however, because to him grouping seemed discriminatory and antidemocratic.

In his review of studies, Eash distinguished sharply between earlier and recent evidence on grouping. He indicated that both types of evidence suggested that grouping per se does not produce improved achievement in children. But Eash reported that earlier and later studies told a different story about the effects of grouping on average and lower ability children. The earlier evidence suggested that there were only minimal effects from grouping per se. According to Eash, more recent evidence suggested that even such grouping programs had negative effects on average and low ability students.

He cited only one recent study to support his interpretation, however, and examination of the actual report of the study shows that it did not find what Eash reported it did. The study was carried out by Husen and Svensson (1960) in Stockholm in 1955 and 1956. Husen and Svensson compared the achievement of students in two kinds of homogeneous schools: upper and lower track. Husen and Svensson were especially interested in how children of low, middle, and high socioeconomic status fared in the two types of schools. They found many nonsignificant and a few significant effects of the two settings. Almost all the significant effects were in favor of the homogeneous schools. Husen and Svensson's key finding, however, was that homogeneous schools were especially beneficial for high aptitude students from the lower socioeconomic groups.

Although the findings of Husen and Svensson were complex, they were consistent with the conclusions drawn by reviewers like Ekstrom (1961) and Passow (1958). These reviewers reported that separate classes with more challenging curricula usually raise the performance of higher aptitude students. Husen and Svensson also found that higher aptitude students benefit especially from separate instruction. What was new was their finding that the benefit to high-aptitude children was especially clear when the children were from the lower socioeconomic level. Eash may have been mistaken in his interpretation of Husen and Svensson's findings, but his conclusion had a powerful effect on later reviewers. After Eash's review appeared, reviewer after reviewer mentioned the difference between earlier and later evidence on ability grouping.

Findley and Bryan (1971) reviewed conclusions in earlier reviews of grouping research, as well as newer literature on grouping. Their method of presenting the new literature was straightforward. They simply presented abstracts of relevant studies in chronological order. Each of the abstracts presented the findings of an individual study, and Findley and Bryan made little or no attempt to classify the studies, interrelate findings, or even to tally them. To readers, the abstracts may have seemed contradictory and confusing, but after presenting a string of abstracts, Findley and Bryan presented clear and unequivocal conclusions, without apologies and without any indication of how the conclusions were drawn from the studies.

Findley and Bryan's review of achievement results, for example, covered 14 studies that appeared in the literature between 1960 and 1966. Most of the studies reported only negligible effects from ability grouping, but a few reported significant effects. One study (Morgenstern, 1963) reported especially large achievement gains for lower ability students; one (Provus, 1960) reported especially large gains for higher ability students; one study (Tobin, 1966) reported significant gains for students at all ability levels; and another (Fick, 1963) reported gains for higher ability and losses for lower ability students.

The pattern in such results is not obvious, but Findley and Bryan drew the conclusion that the slight tendency for grouping to improve achievement in higher level groups is offset by a substantial loss by average and low groups:

> One special footnote is a trend in the results of ability grouping nowadays as contrasted with findings in the 1920s and 1930s. The earlier studies more often than not reported gains by the low groups and losses by the high groups when compared with similar students taught in heterogeneous classes. Today, the trends are just the opposite: any advantages are shown by high level groups: disadvantages are shown quite commonly for the low groups. (p. 30)

Findley and Bryan's treatment of affective findings is similar. The authors reviewed a wide range of both controlled and uncontrolled studies. Some of these studies reported positive and some reported negative effects from grouping. Findley and Bryan's conclusion about the affective consequences of grouping plans was negative:

> As with studies of impact on achievement, the earlier studies show more benefits to the low achievers than now when the low achievers and the high achievers have ethnic and socioeconomic overtones. . . . On the current scene, then, the impact of ability grouping on the affective development of children is to build (inflate?) the egos of the high groups and reduce the self-esteem of average and low groups in the total school population. (p. 40)

Heathers (1969) contributed the review of research on grouping to the fourth edition of the *Encyclopedia of Educational Research*. His conclusions about ability grouping are similar to those that were reached by other reviewers of the era. He reported, first of all, that no consistent effects are found in comparisons of mean scores of students taught in ability-grouped and mixed ability classes. Second, Heathers reported that recent evidence showed that ability grouping has significantly different effects on the achievement of students of high and low ability. Higher aptitude students may benefit, but low aptitude students lose in ability-grouped classes. According to Heathers, the rich get richer and the poor get poorer with ability grouping.

Heathers suggested that the gains for higher ability students were due to adaptation of instruction to meet their needs. He suggested several possible explanations for the losses experienced by lower aptitude students. First, in ability-grouped classes, lower aptitude students lacked the stimulation of higher aptitude learners. Second, students placed in groups labeled as "slow" might expect less from themselves and behave accordingly. And third, their teachers, expecting less from slow students, might teach them less.

During the 1970s and 1980s, reviewers continued to be concerned with equity issues in grouping, but their emphasis shifted away from ability grouping per se and toward the related issue of curricular tracking. In curricular tracking, secondary school students are placed into different classes on the basis of their own choice to prepare for college or for a vocation. In grouping programs, on the other hand, the preferences of students and their parents are usually not a factor in group placement; decisions about group placement are usually made by teachers on the basis of test scores and observation of performance. Rosenbaum (1980) and other sociologically-oriented researchers have pointed out that the two practices share important similarities. With both approaches, according to Rosenbaum, students who are thought to be similar are formed into separate groups, and group membership is based on socially valued criteria so that one's group defines one's position in a status hierarchy.

A shift in methodological preference also occurred during the 1970s and 1980s. Researchers interested in grouping and tracking began to show less interest in controlled studies and more interest in ethnographic studies and national surveys of educational achievement. Coleman's (1966) massive Equal Educational Opportunity Survey undoubtedly fueled interest in the survey approach. Coleman's survey not only provided a methodological model for researchers, but it also raised important new substantive questions about grouping and tracking. A key finding in the Coleman report, for example, was that student achievement varies more within schools than between them. Some researchers wondered whether curricular tracking might have produced a good deal of the variation within schools, and they suspected that survey analyses modeled on Coleman's could document the effects of such tracking. British case studies of streamed schools (e.g., Ball, 1981; Lacey, 1970) were a further influence on American researchers. By the 1980s, American ethnographers were carrying out their own case studies of tracked schools. Some of the studies focused on a single school; others focused on samples of schools. Observations were made in a variety of ways, but usually centered on the quality of instruction offered to students in upper and lower curricular tracks.

A study by Jencks and his colleagues (1972) illustrates the survey approach. To investigate the effects of high school tracking, the researchers looked at 91 predominantly white comprehensive high schools throughout the United States that had tested their students for Project Talent in the ninth grade and had retested them in the twelfth grade. They found that students who reported that they were in the college preparatory curriculum averaged 1 point higher on Grade 12 tests than did students of comparable aptitude in other tracks.

Perhaps the best known example of ethnographic research on grouping is that reported by Oakes (1985) in her book *Keeping Track*. The observations that Oakes presents were originally collected for a project that John Goodlad described in his 1984 book *A Place Called School*. The observations came from 299 English and math classes (75 high track, 85 average track, 64 low track, and 75 heterogeneous classes) in a national sample of 25 junior and senior high schools. The observations covered course content, quality of instruction, classroom climate, and student attitudes in each of the classes.

Results of the observations followed a pattern. Instruction usually seemed to be better in the higher tracks. For example, in English classes, the percentage of time spent on instruction was 81 for the high track and 75 for the low track; in math classes, percentage of time spent on instruction was 81 for the high track and 78 for the low track. In English classes, percentage of time off-task was 2 for the high track and 4 for the low; in math classes, it was 1 for the high track and 4 for the low. In all, more time was spent on instruction and less time was spent off-task in the high tracks.

Gamoran and Berends (1987) reviewed research in both the survey and ethnographic traditions and reported that the results are unclear. They noted that some survey analyses show track effects on student achievement, but other survey analyses show no significant effects. Even when track effects are found, they are usually quite small, and they are especially small in studies with stringent statistical controls. In addition, Gamoran and Berends reported that ethnographic studies usually find only small differences between upper and lower tracks. The differences of 2 or 3 percent in time on instruction or time off-task reported by Oakes (1985), for example, are not large differences. In most respects, track levels appear to be much more alike than they are different.

Slavin (1990a) questioned the nature of the contribution made by the survey and ethnographic approaches on different grounds. Slavin believes that surveys of student

achievement have produced inconclusive findings because they have not examined results in *untracked* control schools. Slavin points out that in surveys of student achievement those in the high and low tracks often differ by two standard deviations on pretests. Covariates cannot adequately control for such group differences. In Slavin's words, no statistician on earth would expect that analysis of covariance or regression could handle such a situation.

> In addition, the logic of such comparisons is simply difficult to accept. Do students at Harvard learn more than those at East Overshoe State, controlling for SAT scores and high school grades? Are the San Francisco Forty-Niners better than the Palo Alto High School football team, controlling for height, weight, speed, and age? Such questions fall into the realm of the unknowable. Comparing the achievement gains of students in existing high versus low tracks is not so different. (p. 506)

Slavin (1990b) also concluded that ethnographic comparisons of instructional quality in high versus low tracks have produced only ambiguous results. He points out that it is hard to tell what conclusions can be drawn from such comparisons:

> For example, teachers typically cover less material in low-track classes. . . . Is this an indication of poor quality of instruction or an appropriate pace of instruction? Students in low-track classes are more off-task than those in high-track classes. . . . Is this due to the poor behavioral models and low expectations in the low-track classes, or would low achievers be more off-task than high achievers in any grouping arrangement? (1990b, p. 474)

I agree with Slavin that there are formidable conceptual difficulties in using the track comparisons favored by survey analysts and ethnographers to assess the value of grouping programs. Even if the results of such comparisons were clear and consistent, interpretation would still be guesswork. For example, there are many different situations in which an investigator could find that upper tracks stimulated students more than the lower tracks did. The following are only two:

1. Instruction in both upper and lower tracks is *more* stimulating than instruction in single-track schools.
2. Instruction in both upper and lower tracks is *less* stimulating than instruction in single-track schools.

The only way to distinguish between such situations and to determine whether students are gaining or losing from tracking is by directly comparing effects in tracked and single-track schools.

Such considerations lead me to conclude that survey researchers and ethnographers are making a serious error in ignoring data from schools and classes that are not tracked. Control data cannot be left out of tracking equations; without such data the equations cannot be solved. Nothing in the literature on tracking, therefore, seems more perplexing and unfortunate to me than Oakes's treatment of observations made in mixed-ability classes for John Goodlad's 1984 national study of schooling in America. Oakes's (1985) *Keeping Track* is a book-length report on the project's observations related to grouping and tracking. She reports in her book that observations were made in a sample of 75 mixed-ability classes as well as 224 tracked classes. Although the book deals at length with observations made in upper and lower tracks, it does not describe any of the results from mixed-ability classes. In brushing aside observations of mixed-ability classes, Oakes appears to have brushed aside the possibility of meaningful answers.

Overall, Eash, Findley and Bryan, Oakes, and other recent reviewers have had a powerful effect on thinking about grouping. Their reviews have raised unsettling questions about the practice, and these questions have resonated in people's minds and hearts. Unfortunately, however, their reviews do not provide definitive answers to the questions they pose. The reviews are too casual in their approach to the evidence. They ignore important research results, misreport key findings, and draw conclusions that do not follow from the cases they cite. The reviews are more notable for passionate conclusions than for dispassionate analysis. If judged on the effect that they have had on current thinking, these reviews would have to be considered successful. The reviews would not get high marks, however, for their treatment of experimental evidence.

Overall, reviews on ability grouping written during the past 60 years seem to reflect their times. In an age of ability testing, reviewers concluded that ability grouping could benefit students. When progressive education was in style and ability testing fell from fashion, reviewers concluded that grouping was more likely to harm than help students. In an era that stressed excellence as an educational goal, reviewers concluded that there were great benefits in the special grouping of higher aptitude students. In an era of educational equity, they concluded that grouping was harmful for disadvantaged and minority students. The times sensitized the reviewers to certain truths about grouping, but they also blinded reviewers to other truths. None of the reviews that we have so far examined can, therefore, be considered definitive. Each presents only a part of the picture.

Examining all of the reviews together does not clear up matters. Taken together, the reviews contain correct conclusions about grouping, but they also contain incorrect ones. The problem is to separate what is right from what is wrong in the reviews. The reviewers themselves do not provide much help. No matter how partisan and inadequate a review, reviewers present their conclusions with confidence. Passow (1962) was right, therefore, when he compared research on ability grouping to a maze. Research findings are confusing and contradictory, and we cannot trust the reviews to guide us through the labyrinth.

Even if the conclusions in these reviews agreed, the value of the reviews would still be limited because the conclusions are so imprecise. The clearest conclusions in the reviews are statements like these:

> The balance of the research is in favor of grouping.
> Grouping leads to a loss in student self-acceptance.

The statements are not helpful because they are so vague. How much of an improvement in student performance can we expect from a procedure that "on balance supports grouping"? It is a poor science that does not tell us the size of the gain to be expected from a treatment or the probability that the treatment will produce a result of a certain size.

To be fair, however, we should note that reviews on ability grouping are neither more nor less flawed than other reviews of their time. In 1980, Jackson examined the scientific adequacy of a random sample of articles from leading journals in education, psychology, and sociology. He concluded that virtually every review was seriously flawed. Reviewers seldom examined critically the evidence in earlier reviews; they often discussed and analyzed only a nonrepresentative sample of studies; they seldom represented study findings exactly; they usually failed to assess systematically possible relations between study characteristics and study findings; and they sometimes failed to

recognize that random sampling error produces variation in study results. Reviews of findings on ability grouping suffer from all these flaws.

We have thus examined six decades of reviews of experimental studies on ability grouping without coming to any clear and definite conclusion about the value of the practice. This is not to say that the reviews are without value. They express their times. They show us the hopes and doubts that grouping raised during the last 60 years. They thus give us a good idea of what questions to ask about grouping. They are a good source of hypotheses, but to test the hypotheses that they suggest, we need to apply more objective and scientific methods to the research literature. In the 1980s scientific methods of research reviewing were developed, and I describe them in the next section of this paper.

Meta-analytic Methods

Glass's 1976 presidential address to the American Educational Research Association was a landmark event in the history of research reviews in the social sciences. In his address, Glass argued that researchers should abandon informal and subjective review methods and commit themselves instead to formal quantitative methods of research reviewing. He coined the term *meta-analysis* to refer to such an approach. Reviewers who carry out a meta-analysis first locate studies of an issue by clearly specified procedures. They then characterize the outcomes and features of these studies in quantitative or quasi-quantitative terms. Finally, meta-analysts use multivariate techniques to describe findings and relate characteristics of the studies to outcomes.

One of the key features in meta-analytic reviews is the use of effect-size statistics to describe study findings. Cohen (1977) has described a number of different effect-size statistics, but the one used most frequently in meta-analytic reviews is the standardized difference between treatment and control means on an outcome measure. Sometimes called *Glass's effect size*, this index gives the number of standard-deviation units that separate outcome scores of experimental and control groups. It is calculated by subtracting the average outcome score for the control group from the average score for the experimental group and then dividing this difference by the standard deviation of the measure. For example, if an experimental group obtains an average score of 600 on a criterion test with a standard deviation of 100 and a control group obtains an average score of 550 on the same test, then the effect size for the experimental treatment is (600 - 550)/100, or 0.5. The effect size indicates that the average score in the treatment group is 0.5 standard-deviation units higher than the average score in the control group.

On the basis of a survey of articles in the social sciences, Cohen (1977) proposed the following rough guidelines for interpreting effect sizes. According to Cohen, effect sizes around 0.2 are small, around 0.5 are medium in size, and around 0.8 are large. Effect sizes can also be interpreted in terms of percentile scores. Under the assumption that treatment effects are normally distributed, an effect size of 0.2 would raise student performance from the 50th percentile to the 58th percentile; an effect of 0.5 would raise performance to the 69th percentile; and an effect of 0.8 would raise performance to the 79th percentile. Glass et al. (1981) have also pointed out a useful relationship between effect sizes and grade-equivalent scores. Empirically, the effect of one year of schooling turns out to be an increase in performance on most standardized tests of 1.0 standard deviation. Thus, effect sizes can also be interpreted in terms of grade-equivalent scores.

An effect size of 0.2 would raise scores by 2 months on a grade-equivalent scale; an effect of 0.5 would raise scores by 5 months; and an effect of 0.8 would raise scores by 8 months.

Reviewers began applying meta-analytic methods to findings on ability grouping in the early 1980s, and they have continued to work with these findings to the current day. With Chen-Lin Kulik, I first used meta-analytic methods in 1982 to integrate research findings on ability grouping in secondary schools (C. Kulik & Kulik, 1982). We later extended our reviews to cover grouping in elementary schools (C. Kulik & Kulik, 1984) and programs of accelerated instruction (J. Kulik & Kulik, 1984). Our most recent reports have provided an overview of this earlier work (J. Kulik & Kulik, 1987, 1991). Slavin (1987, 1990b) has also applied his own version of meta-analysis, called *best-evidence synthesis*, to both elementary and secondary school findings on grouping.

Our meta-analyses have repeatedly shown that higher aptitude and gifted students benefit academically from programs that provide separate instruction for them. Academic benefits for higher aptitude students are positive but usually small when the grouping is done as a part of a broader program for students of all abilities (average effects approximately 0.1 for XYZ classes, 0.4 for cross-grade programs, and 0.3 for within-class programs). Benefits are positive and moderate in size (average effect of 0.4) in enriched classes for gifted students, and academic benefits are striking and large (average effect of 0.9) in accelerated classes. We have also reported that grouping programs have smaller effects on middle and lower aptitude learners. XYZ classes, for example, have virtually no effect on the achievement of such students. Programs of cross-grade and within-class grouping, however, raise achievement scores of middle and lower aptitude students by between 0.20 and 0.35 standard deviations.

Slavin's (1987) meta-analysis of findings from elementary schools covered grouping programs of four kinds: comprehensive XYZ classes, single-subject XYZ classes, cross-grade programs, and within-class programs. Slavin found neither positive nor negative effects for all-day XYZ grouping, and he reported that results were unclear in single-subject XYZ programs. He reported clearly favorable results, however, for cross-grade ability grouping in reading (average effect of 0.45) and for within-class grouping in arithmetic (average effect of 0.3). From the difference in results for comprehensive XYZ classes and other programs, Slavin concluded that grouping is most effective when done (a) for only one or two subjects, (b) with students remaining in heterogeneous classes most of the day, (c) with great reductions in student heterogeneity in specific skills, (d) with frequent reassessment of group assignments, and (e) with teachers varying the level and pace of instruction according to student needs.

Slavin's (1990b) meta-analysis of findings from secondary schools covered only XYZ classes. He reported that such programs had neither positive nor negative effects on student performance. Average effect sizes were approximately zero for both single-subject and comprehensive programs, indicating that students learned the same amount in such programs as they did in mixed-ability classes. Effect sizes were also approximately zero in virtually all subjects and for students in high, middle, and low ability groups.

Thus, many of the findings in our meta-analyses and Slavin's were very similar. Both sets of meta-analyses, for example, found that XYZ programs had negligible effects on most students. Slavin, in fact, reported that effects of such programs were negligible on students of high, middle, and low aptitude. Chen-Lin Kulik and I found that effects on middle and lower aptitude students were negligible but that effects on higher aptitude students were positive but small. Both sets of analyses found that cross-grade and within-class grouping had small-to-moderate positive effects on student achievement.

Neither set of meta-analysis was able to document any consistent negative effects for any type of grouping program. The two sets of meta-analyses differed strikingly in one respect, however, and that was in their treatment of enriched and accelerated classes for the gifted and talented. We included studies of such programs in our meta-analyses and found moderate and strong positive effects from them. Slavin did not include studies of enriched and accelerated classes in his analyses.

The different treatment of special grouping programs for the gifted and talented may have contributed to the different tones in the conclusions from the two sets of meta-analyses. In our conclusions we stressed that grouping programs could help students, especially high-aptitude and talented youngsters:

> The strongest and clearest effects of grouping were in programs designed especially for talented students. The talented students in these programs gained more academically than they would have if they had been taught in heterogeneous classes. . . . Separating talented students into homogeneous groups apparently enabled teachers to provide learning opportunities for the students that were unavailable in more heterogeneous groups. Programs that were designed for all students in a grade - not solely for the benefit of talented learners - had significantly lower effects. (J. Kulik & Kulik, 1987, p. 28)

Slavin, who identifies himself as an opponent of grouping, has emphasized the negligible effect that XYZ grouping has on most students. He drew this conclusion, for example, from his review of findings on ability grouping in secondary schools:

> For practitioners, the findings summarized above mean that decisions about whether or not to ability group must be made on bases other than likely effects on achievement. Given the antidemocratic, antiegalitarian nature of ability grouping, the burden of proof should be on those who would group rather than those who favor heterogeneous grouping. . . . Yet schools and districts moving toward heterogeneous grouping have little basis for expecting that abolishing ability grouping will in itself significantly accelerate student achievement unless they also undertake changes in curriculum or instruction likely to improve actual teaching. (Slavin, p. 494)

In the remaining sections of this report, I attempt to present the meta-analytic findings on grouping as accurately and fully as I possibly can. The results that I will present come from a recent updated statistical analysis that takes into account earlier meta-analytic work by both myself and Slavin. The pool of studies used in the analysis is very similar but not identical to the combined pool of studies used in the two earlier sets of meta-analyses. Based on our rereading of all the studies used in earlier analyses and on our understanding of Slavin's critique of various studies, Chen-Lin Kulik and I eliminated from this new analysis a few studies included in earlier analyses. We also reviewed coding of all the studies, and we revised our earlier coding when it seemed appropriate. My goal was to base conclusions in this report on the best interpretation of the best and most complete set of studies that we could assemble.

Meta-analytic Findings

The analysis covers studies of five major types of ability grouping used in elementary and secondary schools: XYZ grouping, within-class grouping, cross-grade grouping, accelerated classes for the gifted and talented, and enriched classes for the

gifted and talented. The analysis does not cover programs of non-graded instruction, special education, or individualized instruction. Such programs are rarely reviewed in articles on grouping, and studies of such programs were included neither in our earlier meta-analyses nor in Slavin's best-evidence syntheses.

My decision to cover enriched and accelerated classes in this analysis was made only after careful deliberation and deserves some comment. In making the decision, I took into account Slavin's (1987) critique of studies of special classes for the gifted and talented. In brief, Slavin believes that studies of such classes are generally of low quality. He has also argued that such studies should not be included in analyses of grouping research because the special curricula, sizes, resources, and goals of accelerated and enriched classes make them fundamentally different from other grouped classes.

Chen-Lin Kulik and I included studies of enriched and accelerated classes in our earlier reviews for several reasons. First, we took a common-language approach to ability grouping and used the term in the way it is commonly used. Since the 1930s experts on classroom organization have treated special classes for high-aptitude students as a form of ability grouping (e.g., Miller & Otto, 1930; Shane, 1960; Yates, 1966). Second, like many other reviewers, we believe that the methodological weaknesses in studies of enriched and accelerated classes are not great enough to warrant their wholesale dismissal (e.g., Borg, 1964; Ekstrom, 1961; Passow, 1962). Third, examining these studies is important for conceptual reasons. Many reviewers have concluded that grouping works only when a curriculum is adapted to the ability level of those who are grouped. In enriched and accelerated classes, the adjustment of curriculum to student aptitude is especially clear. From studies of such classes, therefore, we can begin to estimate the effects that grouping has when it is done for the purpose of providing instruction adapted to student ability level.

For this report, Chen-Lin Kulik and I reread all of the studies used in our earlier meta-analyses (J. Kulik & Kulik, 1991) and in Slavin's (1987, 1990b) reviews. Out of the total of 143 studies used in these analyses, 127 seemed suitable for this review and 16 seemed unsuitable. Of the 16 excluded studies, 9 had been used only in our earlier reviews, 6 had been used only in Slavin's syntheses; and 1 had been used by both Slavin and ourselves. For the most part, the excluded studies seemed to represent idiosyncratic reviewer choices that we now questioned or judged to be indefensible.

Once the final group of 127 studies was assembled, we coded effect sizes of all studies. The coding was not blind. We consulted our earlier estimates of effect size and also took into account Slavin's estimates of effect sizes for the same studies. The effect sizes that we used in these analyses are, therefore, very similar, but not identical, to those used in our earlier meta-analyses. Effect sizes calculated for this analysis correlated 0.89 with those reported in Slavin's syntheses and 0.97 with those reported in our earlier meta-analyses. We were not surprised that these correlations were less than perfect. Like Glass, McGaw, and Smith (1981, p. 77), we have found that effect-size calculation often requires complex judgments about sources of variation and sometimes also requires decisions about simplifying assumptions. Even experts may disagree on such judgement-calls.

We also coded study features of four types for each study. These features included (a) program characteristics, such as flexibility of group assignments, adjustment of curriculum for group ability level, and duration of the grouping; (b) methodological characteristics, such as use of a control for teacher effects or historical effects; (c) other study characteristics, such as grade level of students, subject matter taught, and type of

tests used; and (d) publication features, such as date and type of publication in which a study was reported.

XYZ Classes

XYZ grouping has been used in American schools for nearly 100 years. The earliest multiple-tier plan described in the literature (Otto, 1941) was a three-track plan put into place in Santa Barbara, California, around the turn of the century. The plan divided the pupils of each grade into three groups: A, B, and C sections. All pupils covered the basic C-level content, but the B pupils did more extensive work than the C groups, and the A groups did still more than the B pupils. The plan was apparently used in the Santa Barbara school for only a few years, and it did not leave a lasting impression on American education.

In 1919, Detroit became the first large city to introduce a formal XYZ plan (Courtis, 1925), and the Detroit plan eventually came to be widely known. For some writers in the 1930s (e.g., Keliher, 1931), in fact, the Detroit plan was synonymous with ability grouping. The plan called for intelligence testing of all children at the start of Grade 1 and then placement of children into X, Y, and Z groups on the basis of the test results. The top 20 per cent went to the X classes, the middle 60 per cent to Y classes, and the bottom 20 per cent to Z classes. Children could be moved from one classification to the other based on teacher judgment, and about 40 per cent of the children, in fact, changed classifications during the first five years of school. Standard Detroit materials and methods were used with all sections (Rankin, Andersen, & Bergman, 1936). No real adjustment of curriculum and methods was made for the ability groups.

Although many school systems followed the Detroit model and instituted three-tier grouping in subsequent years, their plans often differed from the Detroit plan in significant ways. Few schools relied so exclusively on intelligence tests for initial placement into groups, and few separated students at such an early age. In addition, in many programs, especially those in high schools, the separation was not for a full day, but was restricted instead to a single subject. Like the Detroit plan, however, most programs were set up simply to make things easier for teachers by reducing pupil variation in their classes. Few programs used XYZ grouping as a way of providing differentiated curricula to the ability groups.

Student Achievement

The statistical analysis of achievement-test results was designed to shed light on most of the questions that reviewers have raised about student achievement in XYZ classes. The questions fell into three areas. First, what are the overall effects of grouping? Are they negative or positive, and how large and how consistent are such effects? Second, are different students affected differently by ability grouping? Third, are effects different in different types of studies? For example, are effects different in earlier and more recent studies? In studies with good versus poor experimental designs?

Study characteristics. Earlier meta-analyses (J. Kulik & Kulik, 1991; Slavin, 1987, 1990b) covered a total of 56 studies of achievement effects of XYZ grouping. We found all but 5 of these studies to be suitable for the present analyses. One of these 5 studies compared results from programs with different amounts of grouping rather than grouped versus ungrouped classes; one study involved nongraded instruction; one was concerned only with low achieving students; one duplicated results found in another study; and one reported results in too vague a fashion for coding. Excluding these 5

studies from the study pool left us a total of 51 studies on which to base our conclusions (Table 1).

Students assigned to the high, middle, and low tracks in the 51 studies were clearly different in aptitude level. Average IQ was approximately 120 in the high group, 105 in the middle group, and 95 in the low group. Differences among groups were equally clear on standardized achievement tests. On a standard scale with a mean of 500 and standard deviation of 100, average pretest achievement scores were 600 for the high group, 500 for the middle group, and 400 for the low group. This means that in a sixth grade class, the average grade-equivalent score would be 7.0 for children assigned to the high group, 6.0 for children assigned to the middle group, and 5.0 for children assigned to the low group.

The 51 studies covered grouping programs at all grade levels. In 25 of the 51 studies, grouping began during Grades 1 through 6; in 21 studies, in Grades 7 through 9; and in 5 studies, in Grades 10 through 12. A total of 29 of the studies examined comprehensive, full-day programs, but 22 studies examined single-subject programs. Only 9 out of the 51 reports indicated that curriculum and methods were adapted to ability in the XYZ classes, and even fewer reports (2 out of 51) indicated that movement between tracks was flexible.

Methodologically, the 51 studies were adequate, but far from perfect. Only 9 of the studies, for example, involved random assignment of students to experimental and control groups; the remainder employed equivalent control groups. In addition, interpretation of results of some studies was complicated by the possibility of historical, teacher, school, and school-district effects. In 6 studies, for example, XYZ and mixed-ability programs were not offered concurrently; in 35 studies, the programs being compared were offered by different teachers; in 25 studies, the programs were offered in different schools; and in 3 studies, they were offered in different school districts. Almost all of the studies (46 out of 51) measured effects on standardized tests, but 3 studies used only locally designed tests, and 2 studies used a combination of locally designed and standardized tests. Most studies (28 out of 51) were one year in length, but 16 studies lasted for more than one year and 7 studies lasted for less than a year.

The studies were done over an extended time period: 4 were conducted during the 1920s, 6 during the 1930s, 2 during the 1950s, 31 during the 1960s, 7 during the 1970s, and 1 during the 1980s. The studies were reported in several different sources: 26 in journal articles or books, 21 in dissertations, and 4 in technical reports in the ERIC system.

Overall effects. Scores on criterion tests were higher in the multilevel classes in 30 of the 51 studies, but achievement scores were higher in the mixed-ability classes in the remaining 21 studies. Although the number of studies with evidence in favor of XYZ programs is greater than the number of studies with unfavorable evidence, the split is too nearly even to be decisive. The box-score count provides insufficient evidence to reject the null hypothesis of no effect of XYZ plans on overall student achievement.

Table 1.

Major Features and Achievement Effect Sizes in 51 Studies of XYZ Classes

Study	Starting Grade	Course Content	Duration of Instruction	Effect Sizes Overall	High	Middle	Low
Adamson, 1972	7,8	M	2 years	0.20	0.04	0.44	0.15
Bailey, 1968	10	M	1 year	0.03	0.27		Ð0.23
Balow & Ruddell, 1963	6	C	1 year	0.12			
Barker Lunn, 1970	2	C	3 years	Ð0.01	Ð0.01	0.02	Ð0.03
Barthelmess & Boyer, 1932	4	C	1 year	0.23	0.21	0.26	0.18
Barton, 1964	9	R	1 year	0.07	0.20		0.01
Berkun, Swanson, & Sawyer, 1966	3Ð5	R	1 year	0.40	0.30		0.20
Bicak, 1963	8	Sc	21 weeks	Ð0.04	Ð0.08		0.00
Billet, 1928	9	R	30 weeks	0.10	Ð0.04	0.02	0.33
Borg, 1964	4, 6Ð9	C	4 years	0.11	0.17	0.18	Ð0.01
Breidenstine, 1937	2Ð9	C	3 years	Ð0.07	Ð0.05		Ð0.14
Bremer, 1958	1	R	1 year	Ð0.10	Ð0.26	0.01	Ð0.06
Chiotti, 1961	9	M	1 year	0.10	0.12	0.05	0.30
Cochran, 1968	8	C	1 year	0.12			
Daniels, 1961	1	C	4 years	Ð0.27			
Davis & Tracy, 1963	4Ð6	M	1 year	Ð0.15			
Drews, 1963	9	R	1 year	Ð0.07	Ð0.20	0.05	Ð0.05
Fick, 1963	7	C	1 year	0.15	0.37	Ð0.01	0.08
Flair, 1964	1	C	1 year	0.04	0.57	Ð0.16	0.01
Fogelman, Essen, & Tibbenham,1978	7	C	5 years	0.02			
Fowkles, 1931	7	C	1 semester	Ð0.14	Ð0.25	Ð0.21	0.04
Goldberg, Passow, & Justman, 1966	5	C	2 years	Ð0.13	Ð0.02	Ð0.20	Ð0.22
Hartill, 1936	5, 6	C	20 weeks	0.02	Ð0.05	0.07	0.23
Holy & Sutton, 1930	9	M	1 semester	0.29			
Johnston, 1973	1	C	1 year	Ð0.03			
Kerckhoff, 1986	5	C	5 years	0.02			
Kline, 1964	9	C	4 years	Ð0.14	0.07	Ð0.14	Ð0.50
Koontz, 1961	4	C	1 year	Ð0.31	Ð0.13	Ð0.33	Ð0.44
Loomer, 1962	4Ð6	C	1 year	Ð0.08	0.02		Ð0.13
Lovell, 1960	10	R	1 year	0.18			
Marascuilo & McSweeney, 1972	8	So	2 years	Ð0.12	0.09	Ð0.24	Ð0.20
Martin, 1959	6Ð8	C	2 years	Ð0.07	0.19	Ð0.25	Ð0.16
Martin, 1927	--	C	--	0.10			
Morgenstern, 1963	4	C	3 years	0.30			
Moses, 1966	4Ð6	R	1 semester	0.07	0.16	0.05	Ð0.01
Newbold, 1977	7Ð9	C	1 year	0.08	0.13	Ð0.02	0.05
Nichols, 1969	1	R	2 years	Ð0.95			
Peterson, 1967	7, 8	C	1 year	Ð0.12	0.12	Ð0.36	Ð0.12
Platz, 1965	9	Sc	1 semester	0.14	0.14	Ð0.04	0.31
Provus, 1960	4Ð6	M	1 semester	0.27	0.63	0.12	Ð0.08
Purdom, 1929	9	C	18 weeks	0.01	Ð0.02	Ð0.08	0.07
Rankin, Anderson, & Bergman, 1936	2Ð5	C	2 years	0.07	0.22	0.08	0.06
Stoakes, 1965	7	C	2 years	Ð0.10			
Svensson, 1962	5, 8	C	48 weeks	0.04	0.00	Ð0.11	Ð0.24
Thompson, 1974	11	So	1 year	Ð0.34	Ð0.32	Ð0.29	Ð0.35
Tobin, 1966	2Ð6	R	3 1/2 years	0.46	0.52	0.46	0.22
Vakos, 1969	11	So	12 weeks	Ð0.04	0.05	Ð0.01	Ð0.10
Wardrop et al., 1967	3	M	1 semester	0.22			
Willcutt, 1967	7	M	1 year	Ð0.09			
Worlton, 1928	4Ð7	C	3 years	0.21	0.25	0.18	0.21
Zweibelson, 1965	9	So	1 year	0.13			

Note. C = Combined; M = Mathematics; R = Reading; Sc = Science; So = Social Science.

Effect sizes make things clearer. The average effect of grouping in all 51 studies is to raise student performance by 0.03 standard deviations. The median effect size is 0.04. This effect is not a statistically significant one. Following Cohen (1977), we can describe effects whose absolute value is between 0.1 and 0.35 as small, those between 0.35 and 0.7 as moderate in size, and those above 0.7 as large. By these standards, the average effect of XYZ grouping is negligible. It is equivalent to a gain on a grade-equivalent scale of about one-third of a month or a gain in percentile rank from the 50th to the 51st percentile.

Not only was the average effect in the studies small, but the variation in treatment effects across studies was also limited. In 48 of the 51 studies, effects of grouping were trivial or small. Two studies (Thompson, 1974; Tobin, 1966) found moderate positive effects of XYZ grouping. Neither of these studies differed in obvious ways from studies that reported smaller effects. One study (Nichols, 1969) found a large negative effect of XYZ grouping. The study involved only two classrooms, and so teacher differences rather than grouping could have accounted for the unique outcome. The study also involved young children. Difficulties in testing young children might also have contributed to the anomalous results.

Effects by ability level. The average effect of XYZ grouping would obviously be zero if grouping had a negligible effect on all types of students, but it could also be zero if grouping had positive effects on one type of student and negative effects on another type. It is important to determine whether the former or latter situation produced the near-zero average effect that we observed for XYZ programs. Does XYZ grouping have the same effect on all types of students, or does it affect higher and lower aptitude students differently?

A total of 36 studies reported results separately by ability level. The average effect size is 0.10 for higher aptitude, Ð0.02 for middle aptitude, and Ð0.01 for lower aptitude students. The median effect size is 0.13 for higher aptitude, Ð0.02 for middle aptitude, and Ð0.01 for lower aptitude students. The effects for middle and lower aptitude students are not significantly different from each other, nor are effects on these students significantly different from zero. The effect on higher ability students is significantly greater than zero, however, and it is also significantly different from the effects on middle and lower ability students.

XYZ grouping, therefore, affects different students differently. It gives higher aptitude children a boost and helps them move slightly ahead of their peers in mixed-ability classrooms. XYZ programs have virtually no effect, however, on the achievement of middle and lower aptitude children. It seems possible that teachers introduce more challenging materials and methods into higher aptitude classes than they would use in mixed-ability situations. Teachers in middle and lower tracks, on the other hand, may teach in much the same way that they do in mixed-ability settings.

Effects by study features. In his review of grouping in elementary schools, Slavin (1987) listed four features that seem to contribute to the effectiveness of grouping programs:

1. **Curricular differentiation.** Programs in which curricular materials and methods are adjusted to the ability groups seem to produce larger effects than do programs without curricular adjustment.
2. **Flexibility of grouping.** Programs in which group placement is flexible seem to be more effective than inflexible programs.

3. **Method of group assignment.** Programs in which students are assigned to groups on the basis of a specific skill (e.g., a test in reading or arithmetic) seem to be more effective than programs in which assignment is based on an intelligence test or test of overall achievement.
4. **Extent of grouping.** Single-subject grouping seems to be more effective than comprehensive grouping.

Our analysis of XYZ studies failed to support Slavin's speculation. None of the 4 features mentioned by Slavin was significantly related to results in the 51 studies of XYZ grouping. Effect sizes are nearly identical in studies with and without these features. For example, in the 22 studies in which grouping was restricted to a single subject, average effect size is 0.02; in 29 studies of comprehensive, full-day grouping, it is 0.04. In 18 studies in which students were assigned to groups on the basis of a specific skill, average effect size is 0.07, whereas it is 0.01 in studies in which group assignment was based on overall intelligence or achievement level. Average effect size is 0.02 in 41 studies without curricular adjustment, whereas, it is 0.08 in 9 studies with curricular adjustment. Average effect size is 0.03 in 48 studies where grouping was not flexible, and it is 0.02 in 2 studies with flexible grouping.

Reviewers of the literature on grouping have suggested that two other factors may influence study findings. Slavin (1987) has speculated that study quality is a key factor. He expects higher quality studies to produce effects that are consistently around zero in magnitude; lower quality studies are expected to produce higher effect sizes. In addition, several reviewers have suggested that results of grouping have been less favorable in studies carried out after 1960 (e.g., Findley & Bryan, 1971; Heathers, 1969). The damaging effects of grouping have become greater in recent years, the argument goes, because the discrimination inherent in grouping has become more evident.

In fact, neither of these speculations is supported by evidence. Effect sizes in 9 true experiments average 0.03; in 42 quasi-experiments, they also average 0.03 standard deviations. The correlation between year of study and effect size ($r = .17$) is small and nonsignificant. Average effect size in 12 studies published before 1960 is 0.05; average effect sizes in those published since 1960 is 0.03.

Meta-analysts sometimes report finding relationships between methodological features of studies and study outcomes in educational research. For example, Chen-Lin Kulik and I have found that four study features are often related to effect size in research literatures (J. Kulik & Kulik, 1989):

1. **Source of publication.** Journal articles and technical reports often report larger effects from educational treatments than do dissertation studies.
2. **Study length.** Short studies, in which an educational treatment is given to learners for only three or four weeks, often report stronger results than do longer studies.
3. **Control for teacher effects.** Studies without a control for teacher effects (i.e., those in which different teachers instruct experimental and control classes) often report stronger results than do studies in which the same teachers are in charge of experimental and control classes.
4. **Test authorship.** Studies that use locally developed instruments as criterion tests sometimes report stronger results than do studies using standardized tests.

None of these factors is significantly related to the size of effect in the 51 studies of XYZ programs. Effects are similar, for example, in studies published in dissertations,

journals, and technical reports. Average effect size is 0.02 in 26 studies published in journals, 0.05 for 21 dissertation studies, and 0.01 in 4 studies released only as technical reports. Effect sizes average 0.08 in 7 studies lasting one term or less, 0.03 in 28 studies lasting approximately 1 year, and 0.01 in 16 studies lasting longer than 1 year. Average effect size is 0.06 in 15 studies with a control for teacher effects, whereas it is 0.01 in 35 studies without such a control. Effect sizes average 0.03 in 45 studies using standardized tests, -0.01 in 3 studies with locally developed tests, and 0.02 in 2 studies using a combination of local and standardized tests.

Student Self-esteem

Our statistical analysis was also designed to examine the effects of XYZ classes on student self-esteem. A popular hypothesis is that self-esteem of low-aptitude children drops in such classes. Advocates of this hypothesis usually consider labeling, or stigmatizing, to be the basic problem. They point out that labels become attached to ability groups, and they believe that these labels come to function as self-fulfilling prophecies that cause low-aptitude children to lose academic pride and motivation. An alternative hypothesis has also been proposed, however. Advocates of this hypothesis believe that the self-esteem of lower aptitude children may actually go up in ability grouped classes because lower aptitude children have more of an opportunity to participate, to compete, and even to shine in such classes. In mixed-ability classes, lower aptitude children are often overshadowed by quicker classmates. This hypothesis stresses the social comparisons that children make (Hoge & Renzulli, 1992).

It is important to note that self-esteem and academic aptitude usually covary in mixed ability classes. When taught in mixed-ability classrooms, high-aptitude students get higher scores on self-esteem measures than low-aptitude students do. In Goldberg, Passow, and Justman's (1966) study, for example, the difference between the higher and lower aptitude groups in self-esteem (as measured by the "I Am" scale on the researchers' inventory *How I Feel About Myself*) is approximately 0.25 standard deviations. In Drews's (1963) study, the difference in self-esteem between the higher and lower aptitude groups (as measured on her scale of *Concept of Self as Learner*) is approximately 2 standard deviations. The size of the difference varies from scale to scale, but higher self-esteem scores for higher aptitude students is the general finding in mixed-ability classes.

The fundamental question that we must ask, therefore, is whether XYZ grouping accentuates the difference in self-esteem of good and poor students, or whether it reduces it. The labeling hypothesis suggests that the difference will become more pronounced with XYZ grouping. The social comparison hypothesis, however, predicts that the difference will decline with such grouping.

Study characteristics. Only 13 studies were available in which to examine self-esteem effects of XYZ grouping (Table 2). The pool of 13 studies included all but two of the studies of self-esteem used in our earlier analyses (J. Kulik & Kulik, 1991). One of the two studies was eliminated because its self-esteem data came from a single item rather than a total scale. Another study was eliminated because the study compared results from schools with varying degrees of grouping rather than comparing grouped versus ungrouped classes.

Table 2.

Average Effect Size of XYZ Classes on Self-esteem

Study	Effect Size			
	Overall	High	Middle	Low
Adkison, 1964	0.06	Ð0.39		0.52
Barker Lunn, 1970	Ð0.05	Ð0.07	Ð0.18	0.12
Borg, 1964	Ð0.17	Ð0.10	Ð0.27	Ð0.16
Davis & Tracy, 1963	0.09			
Drew, 1963	0.28	Ð0.07	0.18	0.73
Dyson, 1967	0.13	Ð0.02		0.21
Erickson, 1973	Ð0.60	Ð0.40		Ð0.81
Fick, 1963	Ð0.04	0.00	Ð0.04	Ð0.08
Goldberg, Passow, & Justman, 1966	Ð0.14	Ð0.31	Ð0.16	0.24
Marascuilo & McSweeney, 1972	0.06	Ð0.30	Ð0.40	0.88
Morgenstern, 1963	Ð0.22			
Sarthery, 1968	0.02	Ð0.01		0.04
Tauber, 1963	0.20	0.01	0.08	0.48

Researchers used a variety of scales with a variety of names to measure self-esteem in the 13 studies. The two measures used most often in the 13 studies were Bills' Index of Adjustment and Values (Bills, 1951) and the Self-Acceptance scale of the California Test of Personality (Thorpe, Clark, & Tiegs, 1953). Bills' Index of Adjustment and Values consists of a list of adjectives (e.g., *agreeable, cooperative, happy, obedient, smart, understanding*). Children first check the adjectives that seem to describe them best, and they then indicate whether they are satisfied with the characteristics denoted by the adjectives. On the Self-Acceptance scale of the California Test of Personality, children respond *Yes* or *No* to questions like these:

Do your friends generally think that your ideas are good?
Do your folks seem to think that you are doing well?
Can you do most of the things you try?

Of the 13 studies, 7 measured self-esteem on a standardized scale, whereas 6 studies used locally developed scales. Eight studies used measures of general self-concept; 2 measured academic self-concept; and 3 studies measured a combination of the two. Whatever the names of the instruments and whatever their paternity, however, they were all designed to measure the degree to which children held positive or negative views of themselves.

A total of 7 of the studies were conducted in elementary schools, whereas 6 were done in junior high schools. The 13 studies were all of fairly recent vintage: 10 from the 1960s and 3 from the 1970s. They came from standard sources: 5 were found in journal articles, 6 in dissertations, and 2 in technical reports.

All of the studies examined effects after at least one year of XYZ grouping, and most of the studies (9 out of 13) examined effects of comprehensive, full-day programs. In other respects, the 13 studies were similar to our larger pool of studies of XYZ grouping. Material was not adjusted to group ability level in any of the programs; none involved flexible assignment to groups. The 13 studies were also similar to studies in the

larger pool in methodological features. Few of the studies involved random assignment of subjects to conditions, and most lacked a control for teacher or school effects.

Study findings. Does XYZ grouping raise or lower self-esteem level? In 7 of the 13 studies, self-concepts were more favorable overall with XYZ grouping; in the remaining 6 studies, self-concepts were more favorable in mixed-ability classes. The average overall effect of grouping in the 13 studies is a drop in self-esteem scores of 0.03 standard deviations. This effect is both very small and statistically nonsignificant.

Only 1 of the 13 studies, however, found an average effect that was not trivial or small in size. The study (Erickson, 1973) found negative effects of XYZ grouping on children of high, middle, and low aptitude. The study compared self-esteem scores of children in two separate school districts with different grouping policies. The differences in self-esteem for children in the two districts could have been produced by factors other than grouping policy.

Eleven of the 13 studies reported results separately by ability level. The average effect size is -0.15 on high-aptitude students, -0.09 on middle-aptitude students, and 0.19 on low-aptitude students. The median effect size is -0.07 for higher aptitude, -0.16 for middle aptitude, and 0.21 for lower aptitude groups. The effects on the higher and lower aptitude children are significantly different, and the effect on the higher aptitude children is significantly lower than zero.

The effect of XYZ programs on student self-esteem thus appears to be a leveling effect. In mixed-ability classes, higher and lower aptitude children are clearly different in self-esteem. In XYZ programs, they become more similar in self-esteem levels. Brighter children lose some of their self-assurance when they are put into classes with equally talented children. Slower children gain in confidence when they are taught in classes with other slow learners. They may feel less overwhelmed and less overshadowed in such classes.

Conclusions

The important point to note about XYZ grouping is that its overall effect on achievement is trivial. It raises achievement in the total population by an average of about 0.03 standard deviations. This gain is only slightly less than the one (0.06) found in my earlier meta-analyses (e.g., J. Kulik & Kulik, 1991), and it is consistent with the effect of zero found by Slavin for what he calls *ability-grouped class assignments*. The gain of 0.03 standard deviations is not large enough to be considered statistically different from zero.

It is impossible to say with statistical certainty why certain XYZ programs produce positive effects and others produce negative ones. Slavin, for example, has speculated that grouping has maximum positive effects on student achievement (a) when it is done for only one or two subjects; (b) when students remain in mixed ability classes most of the day; (c) when grouping greatly reduces heterogeneity in a specific skill; (d) when group assignments are frequently reassessed; and (e) when teachers vary the level and pace of instruction according to student need. We examined results of programs that had these features and compared them to results of programs that lacked these features. We found no direct evidence that these, or any other study features we examined, were significantly related to results of XYZ studies.

The effects of XYZ programs were not uniform, however, on all types of children. Instead, effects varied as a function of student aptitude level. XYZ programs have small but positive effects on the achievement of students of higher aptitude, but they have no consistent effect on middle and lower aptitude students. Teachers and researchers have long suspected that the effects of grouping depend on the amount of curricular adjustment that is associated with it, and it is possible that curricular adjustment can explain this pattern of effects. Teachers may introduce more challenging material into classrooms composed largely of higher aptitude students, whereas they may teach middle and lower aptitude groups in much the same way that they teach mixed-ability ones.

XYZ programs do not have devastating effects on student self-esteem. The net gain in self-esteem from XYZ grouping is virtually zero, but effects may be slightly positive for lower ability students and slightly negative for higher aptitude ones. Talented students may become slightly less satisfied with themselves when taught with their intellectual peers; slower students may gain slightly in self-confidence when they are taught with other slower learners. XYZ programs thus appear to have a leveling effect on student self-esteem scores. In mixed-ability classes, higher and lower aptitude students sometimes differ markedly in self-esteem. In XYZ classes, higher and lower aptitude students tend to become more similar in self-esteem.

Cross-grade Grouping

On October 26, 1957, the *Saturday Evening Post* published an article entitled "Johnny *Can* Read in Joplin" that described a remarkable reading program that had been established in the elementary schools of Joplin, Missouri. According to the article, the program was revitalizing elementary education in the Missouri town. School children who had once been indifferent to school were now reading dozens of library books each year. Parents and teachers were jubilant about the changes that they saw in their children. Other school districts in Missouri and neighboring states were looking toward Joplin for ideas on remodeling their reading programs, and word of what was happening in Joplin had begun to spread throughout the country.

The Joplin plan was devised by Joplin Assistant Superintendent of Schools Cecil Floyd in 1953, and it involved cross-grade grouping of fourth, fifth, and sixth graders for reading instruction. During the hour reserved for reading, children in these grades would break up into groups that went to reading classes on anything from the second- to the ninth-grade level. In these classes, the children would work with other fourth, fifth, and sixth graders who were reading at the same level. After this period was over, the children returned to their age-graded homerooms for a twenty-five-minute period of reading for enjoyment. A variety of books and magazines were available in each homeroom, and the children were free to read anything they liked.

Floyd did not carry out a formal evaluation of the outcomes of the plan, but he noted that the program seemed to have a strong effect on student performance in reading. He reported that the top 500 pupils graduating from the Joplin elementary schools in 1957, children who had been exposed to the reading plan for three years, scored at approximately the ninth grade level on entry into junior high school. Earlier tests, made in 1950, showed that the top 500 students at that time averaged only slightly above the beginning-seventh-grade level. In other words, the test-score gain attributable to the program appeared to be about 1.5 to 2.0 standard deviations. By any standards, this is a remarkably large effect from an educational program.

The Joplin plan of cross-grade grouping is like XYZ plans in that students of different ability levels are taught in separate classrooms. But in cross-grade plans, there

are typically more levels. In a typical Joplin program, for example, a fifth grader might be assigned to any one of nine different reading groups. In addition, cross-grade grouping is single-subject grouping, and so group placement is usually tied closely to a specific skill. Perhaps the most important difference between cross-grade and XYZ programs, however, is in the amount of curricular adjustment in the two approaches. Materials and methods are completely adapted to group level in cross-grade programs. Pupils in different ability groups work with different materials and different methods. In most XYZ programs, little or no effort is made to adjust curriculum to group ability level.

Although the results described in the *Saturday Evening Post* were dramatic, the account was anecdotal. Educators wanted to know whether careful studies would support the claims for the method. Schools around the country began experiments with the plan, and educational researchers began careful evaluation studies of the plan. By the late 1960s enough studies were available for some conclusions to be reached. Formal syntheses of results, however, did not appear in print until the 1980s. In 1987, Slavin reviewed results from 14 studies of the Joplin plan and reported that the gains from the Joplin plan were 0.45 standard deviations higher than gains from mixed-ability classes. In 1987, we reviewed 16 studies of cross-grade grouping and found that gains from the approach were 0.3 standard deviations higher than those from mixed-ability classes.

Study characteristics. Our meta-analysis (J. Kulik & Kulik, 1991) and Slavin's (1987) together covered a total of 17 studies of cross-grade grouping. All but 3 of these studies seemed to be suitable for use in this review. Two of the unsuitable studies evaluated nongraded classrooms rather than cross-grade grouping, and the other study evaluated a program for low-achieving pupils rather than a program for a representative population. We, therefore, excluded these 3 studies from the study pool, leaving a total of 14 studies from which to draw conclusions (Table 3).

Table 3.

Major Features and Achievement Effect Sizes in 14 Studies of Cross Grade Grouping

Study	Starting Grade	Course Content	Duration of Instruction	Effect Size			
				Overall	High	Middle	Low
Anastasiow, 1968	5-6	R	1 year	0.17			
Carson & Thompson, 1964	4-6	R	1 year	0.00			
Chismar, 1972	4-8	R	1 year	0.14			
DeGrow, 1964	4-6	R	1 year	0.09			
Green & Riley, 1963	4-6	R	1 year	0.46			
Halliwell, 1963	1-3	R	1year	0.62			
Hart, 1959	4-5	R	1 year	0.98			
Ingram, 1960	1	R	3 years	0.81			
Jones et al., 1967	1	R	3 years	0.33			
Kierstead, 1963	3-8	R	1 year	Ð0.01	Ð0.04	Ð0.01	0.08
Moorhouse, 1964	4	R	5 terms	0.26			
Morgan & Strucker, 1960	5-6	R	1 year	0.35	0.28		0.50
Rothbock, 1961	4-5	R	1 year	0.49			
Russell, 1948	4-6	R	2 years	Ð0.03			

Note. R= Reading.

The majority of the studies (11 of 14) were published during the 1960s. One study, however, dated back to the 1940s; one came from the 1950s; and one was done in the 1970s. Reports on the studies were found in two different types of publications: 2 reports were found in dissertations, and 12 in journal articles. All the studies were carried out in elementary schools, and all examined effects after at least one year of cross-grade grouping.

In several major respects these cross-grade programs were different from the XYZ programs that we studied earlier. All of the studies, for example, examined single-subject grouping for reading. Grouping was based on a specific aptitude (reading level) in each of the studies, and material was adjusted to group ability in all of the studies. Finally, in 5 of the 14 studies, placement into groups was flexible; pupils were moved from level to level as appropriate during a year.

In methodological characteristics, however, the 14 studies were very similar to studies of XYZ programs. All of the studies used standardized tests as criterion measures of student achievement. The major threats to study validity were a failure to control experimentally for teacher and school effects and the use of equivalent control groups rather than random assignment of subjects to groups.

Achievement findings. Eleven of the 14 studies found that students achieve more when taught under cross-grade plans; two studies found that students achieve less; the effect size was zero in one study. Although suggestive, these box-score results do not provide conclusive evidence that cross-grade plans have positive effects on student achievement.

The average effect of cross-grade grouping in the 14 studies was to raise performance on criterion tests by 0.33 standard deviations. This effect is significantly different from zero, and it is significantly different from the average effect found in programs of XYZ grouping. The median effect in the 14 studies is 0.30. By conventional standards, the effect of cross-grade grouping on student achievement is moderate in size. It is equal to a grade-equivalent gain of 3 months, or a rise in percentile scores from 50 to 62.

Effects of cross-grade grouping varied from a slightly unfavorable effect of -0.03 in one study to a highly positive effect of 0.98 in another. In four of the studies, effects of cross-grade grouping were trivial; in four studies, they were positive but small; in four studies, they were positive and moderate in size; and in two studies, they were positive and large. Because of the small number of studies of cross-grade grouping, however, it was not possible to carry out analyses of relations between study features and study outcomes.

Only two studies reported effects separately for students of high, middle, and low aptitude. Average effects were 0.12 for higher aptitude learners, -0.01 for middle aptitude, and 0.29 for lower aptitude learners. In the absence of evidence to the contrary, it seems safe to assume that cross-grade grouping has beneficial effects on both good and poor students.

Conclusions. Unlike XYZ grouping, cross-grade grouping clearly works. Cross-grade grouping produces small to moderate positive effects on student achievement scores in most studies, and its average effect is to raise achievement test-scores by approximately 0.30 standard deviations. The positive effects of cross-grade grouping are not restricted to a single type of student. Cross-grade programs instead appear to work for all students, with both good and poor students profiting from it.

Why are cross-grade programs more effective than XYZ programs? The two types of programs differ in several ways. Cross-grade programs are more likely to be single-subject programs, for example, and they are also more likely to involve placement based on a specific skill. It seems unlikely, however, that the restriction of grouping to a single subject or the placement of students based on a specific skill could be important matters. The literature contains many studies of XYZ programs that have these characteristics, and the studies show that such programs are neither more nor less effective than programs that lack these features.

A more important factor may be the large amount of curricular adaptation in cross-grade programs. By definition, cross-grade programs involve adjustment of curriculum to ability level. In cross-grade programs, students are taught material at their level in separate classrooms. For high-aptitude students, cross-grade grouping is accelerated instruction. For lower aptitude students, it is remedial instruction. For students at both extremes, the curriculum differs from that followed by middle-aptitude students. The close fit between curriculum and aptitude may be the key factor that makes cross-grade grouping so successful.

Within-class Grouping

Elementary school teachers often group the children within a class into subgroups for specific activities and purposes. They use such subgroups especially often for reading and arithmetic lessons, and they sometimes form subgroups for science and social science projects as well. The teacher usually presents a lesson to one of the subgroups while the remaining groups engage in other activities.

Although such within-class grouping may have as long a past as XYZ grouping, it has a much shorter history. In 1953 Petty wrote an important volume that reviewed the professional literature on the topic. She concluded that the literature established a philosophy for within-class grouping, but she also pointed out that it did not provide concrete information about techniques that could be used in teaching subgroups. She also reported that she was unable to find studies that compared the effects of teaching with and without such grouping.

Within a few years, the situation changed. Ability grouping became an active area for research and development during the late 1950s and 1960s, and within a few years of Petty's report, educators and researchers were publishing articles describing programs of within-class grouping that had been established for teaching arithmetic and reading. Social priorities of the times focused attention especially on the programs of arithmetic instruction. They became a target not only of development efforts but also of research and evaluation studies.

Dewar (1963) described a typical within-class program of arithmetic instruction in a school in Johnson County, Kansas. Each 6th grade class in the school was divided into three groups on the basis of achievement test results, teacher opinion, and school records. Membership in the groups was to remain constant for the entire term, and teachers were to use textbook materials from the 4th through the 8th grades, as well as especially prepared curriculum guides in teaching the groups. Group I used 6th, 7th, and 8th grade texts; Group II used 5th, 6th, and 7th grade texts; and Group III used 4th, 5th, and 6th grade texts. Each teacher spent 55 minutes per day in arithmetic instruction. The teacher presented material to a group for approximately 15 minutes before moving on to another group. While the teacher was presenting material to one group, the other groups worked on their arithmetic assignments.

Two facts about within-class grouping plans make them especially interesting. First, most within-class grouping plans call for differentiated instruction for the groups. For the practice of within-class grouping to make sense, the teacher must present different material to each group. It would be inefficient for a teacher to divide a class into thirds on the basis of ability and then to make the same presentation separately to each of the three groups. Thus, within-class programs are like cross-grade programs in that they involve differentiated curriculum. Second, within-class programs do not involve assignment of groups to separate classrooms. Within-class programs differ from both XYZ and cross-grade programs in this respect.

By the 1970s enough evaluation studies of within-class programs had accumulated for separate review of these studies. Begle (1975) reviewed 8 such studies, and he concluded that the studies showed that within-class grouping was a worthwhile practice in mathematics education. Three of the 8 studies found positive results; 5 found no significant effects; but no study reported negative or mixed results. Meta-analyses by Slavin (1987) and ourselves (J. Kulik & Kulik, 1991) have also reached favorable conclusions about the effects of within-class ability grouping.

Study characteristics. Our earlier meta-analysis and Slavin's synthesis covered a total of 16 studies of within-class grouping. Five of the 16 studies seemed unsuitable for use in this review. In 2 of the studies, the treatment was individualized instruction rather than within-class grouping; in 1 study, subgroupings were not formed for instruction; in 1 study, the subgroups were not formed on the basis of ability; and in 1 study, the within-class program was used with a low-achieving rather than representative population. The exclusion of the 5 studies from the study pool left 11 studies from which conclusions could be drawn (Table 4).

Table 4.

Major Features and Achievement Effect Sizes in 11 Studies of Within Class Grouping

Study	Starting Grade	Course Content	Duration of Instruction	Effect Size			
				Overall	High	Middle	Low
Campbell, 1964	7	M	1 year	Ð0.18	0.26	Ð0.41	Ð0.36
Cignetti, 1974	7,8	O	9 weeks	0.09	0.27	0.22	Ð0.41
Dewar, 1963	6	M	23 weeks	0.48	0.47	0.50	0.56
Eddleman, 1971	5	M	9 weeks	Ð0.09			
Jones, 1948	4	C	1 year	0.21	0.19	0.23	0.40
Putbrese, 1971	4	M	1 year	0.16			
Shields, 1927	7	R	6 weeks	0.82			
Slavin & Karweit, 1984	3-6	M	1 semester	0.43	0.41	0.38	0.50
Smith, 1960	2-5	M	1 semester	0.22	0.18	0.15	0.30
Spence, 1958	4-6	M	30 weeks	0.60			
Wallen & Vowles, 1960	6	M	1 semester	0.06			

Note. C = Combined; M = Mathematics; O = Other.

The 11 studies included in this analysis were published over a period of several decades. One came from the 1920s; 1 came from the 1940s; 1 from the 1950s; 4 from the 1960s; 3 from the 1970s; and 1 from the 1980s. The studies came from different sources: 4 from journal articles, 6 from dissertations, and 1 from a technical report in the ERIC

system. Eight of the studies were carried out in elementary schools and 3 in junior high schools. Five of the studies examined effects after one year of within-class grouping; 6 of the studies lasted between 6 weeks to one semester.

In several important characteristics, the studies were like studies of cross-grade grouping. Most of the studies, for example, examined single-subject grouping. In eight of the studies, within-class grouping was used for arithmetic only; in one study, it was used for reading only; and in one study, it was used for typewriting only. In the remaining study, within-class grouping was used in all subjects. In addition, material was adjusted to group ability in all studies, and grouping was based on a specific aptitude in all but one of the studies. Group assignments were described as flexible in 2 of the 11 studies.

In methodological characteristics, the 11 studies were similar to other studies in our larger pool. All but one of the 11 studies used standardized tests as criterion measures of student achievement. The major threats to validity of the studies were the use of intact groups for experimental and control comparisons and the lack of controls for instructor and school effects.

Achievement findings. Nine of the 11 studies found higher overall achievement for students grouped within classes, and 2 studies reported higher overall achievement in classes taught without such grouping. In 5 of the 11 studies, differences in overall achievement in the two types of classes were large enough to be statistically significant. In each of these 5 studies, the performance of students from the grouped class was higher. These box-score differences tend to favor within-class grouping over whole-class instruction, but the box-score count is not lopsided enough to be absolutely conclusive.

Analysis of size of effects makes the situation clearer. The average overall effect size in the 11 studies is 0.25 standard deviations. This average effect is significantly different from zero. It is also significantly greater than the effect of XYZ programs, but it is not significantly different from the effect of cross-grade programs.

Six of the studies of within-class grouping reported results separately by ability group. Effects were similar for students of high, middle, and low ability taught in grouped and ungrouped classrooms. The average effect size was 0.30 for higher ability students; 0.18 for middle ability students; and 0.16 for lower ability students. The differences in effect size are not statistically significant.

Effects also varied in size in the 11 studies. The largest positive overall effect was 0.82 standard deviations; the largest negative effect was -0.18 standard deviations. Because of the small number of studies available, further analyses were not carried out to determine whether such differences in overall effects were related to study features.

Conclusions. Like cross-grade programs, within-class programs have a good record of effectiveness in the evaluation literature. Both types of programs raise student achievement on criterion tests by about 0.2 to 0.3 standard deviations. Like cross-grade programs, within-class programs seem to work for all sorts of students. They help the lower aptitude learner, the learner of middle aptitude, and the higher aptitude learner. Like cross-grade programs, within-class programs have a better record of increasing student achievement than do XYZ programs.

Within-class and cross-grade programs differ from XYZ programs in several ways. Within-class and cross-grade programs are almost invariably single-subject programs in which group assignment is based on a specific skill. We do not believe,

however, that these characteristics are critical ones in explaining the effectiveness of within-class and cross-grade programs. XYZ programs may be single-subject or comprehensive, and they may be based on a specific or a general skill. In whatever form they come, XYZ programs seem to produce very small effects on the typical student.

Another factor that is common to cross-grade and within-class grouping programs seems to us to be far more important in explaining their effectiveness. The factor is adaptation of curriculum to student level. Like cross-grade programs, within-class programs usually involve such adaptation. Indeed, without such adaptation of material to group level, it would be pointless for a teacher to use within-class grouping. Repeating the same lesson several times for different student groups would be a waste of teacher time, and it would thus be a waste of student time. Instead of providing such repetitive lessons, teachers using within-class groups adapt their lessons to their audience. This adaptation may be the key to their success.

Special Accelerated Classes

American education has a long tradition of providing special classes for children whose educational needs differ from those of the majority. Special classes have been formed, for example, of children who are physically handicapped, emotionally or socially maladjusted, lacking in proficiency in English, and so on. Children may be placed in special classes for the duration of their schooling, for a transitional period, or for a short time.

One of the longest of these traditions is providing special classes for gifted and talented children. The first classes devised especially for such children were accelerated ones. The Cambridge Double Track Plan of 1891, for example, put bright children into special classes that covered the work of six years in four, and the special-progress classes of New York City, originally established in 1900, allowed pupils to complete the work of three years in two. Other school systems introduced other forms of acceleration in the next decades, and by the 1920s accelerated instruction seemed to be established as the basic method for dealing with gifted school children.

The basic idea of educational acceleration is to modify a school program so that students complete it at an earlier age or in less time than is usual. Such acceleration can be achieved in a variety of ways. Students may accelerate their progress through elementary and high school by entering kindergarten or first grade early; by grade skipping, or double promotion; by participating in programs that compress instruction (e.g., four years in three); or by taking extra courses or summer sessions to shorten total time in school. Students may accelerate their college education by entering college as a full-time student without completing high school; by entering with sophomore standing based on advanced placement credits; by accumulating college credits rapidly through examinations; by taking heavier-than-average course loads; or by attending college year round.

Programs of accelerated instruction have been in and out of favor among educators during the past century. Although interest in accelerated programs declined during the 1930s, 1940s, and 1950s, it revived after the launching of Sputnik by Russia in 1957. That event signalled the start of a technological competition between the U.S. and the U.S.S.R. and brought about widespread reassessment of programs for nurturing technological and scientific talent in U.S. schools. Another stimulus to renewed interest in accelerated instruction was the publication of results from several major studies of educational accelerates. Terman and Oden (1947), for example, presented compelling evidence that exceptionally able students who had been accelerated in school were more

successful academically and vocationally than equally talented students who had not been accelerated. Pressey and his colleagues (e.g., Flesher & Pressey, 1955) reported similarly impressive academic and life outcomes for accelerated college students, and the evaluation of the Ford Foundation's program of early entrance to college produced additional confirming evidence (Fund for the Advancement of Education, 1957).

Interest in accelerated instruction has also been stimulated in recent years by work on radical acceleration carried out by Julian Stanley and his colleagues at the Johns Hopkins University (e.g., Stanley, 1980). For more than a decade this group has developed programs for identifying, describing, and nurturing the talents of mathematically and verbally precocious youth. They have disseminated information about their model programs widely and carried out a number of evaluations of accelerated instruction.

Most reviewers of studies of acceleration have come to favorable conclusions about its effects. In her 1958 review, for example, Goldberg pointed out that it was hard to find a single research study showing acceleration to be harmful and that many studies proved acceleration to be a satisfactory method of challenging able students. Begle (1976), Ekstrom (1961), and Passow (1958) drew equally positive conclusions about programs of accelerated instruction. More recently, a meta-analyses by Rogers (1991; 1992) have reported favorable results from accelerated programs of a variety of types.

My meta-analysis with Chen-Lin Kulik (J. Kulik & Kulik, 1984) covered studies where the achievement of students in accelerated classes was compared to achievement of comparable students in nonaccelerated control classes. It focused on programs of moderate acceleration of a whole class of students rather than on programs of individual or radical acceleration. The review was thus narrower in scope than Rogers' review, but it also covered a more homogeneous group of studies. It nonetheless produced strong evidence for the effectiveness of accelerated classes. The analysis showed that examination performance of students who were accelerated by one year surpassed by nearly one grade level the performance of nonaccelerates of equivalent age and intelligence. The studies produced no evidence that acceleration had negative effects on nonintellective outcomes.

Study characteristics. All but one of the reports on acceleration that we used in our earlier meta-analysis (J. Kulik & Kulik, 1984) were appropriate for use in this analysis. The excluded study examined a program of individual grade-skipping and thus differed from the studies of accelerated classes that were our focus here. We also discovered that one study used in our analysis of enriched classes also contained data on an accelerated class, so we added the study to our pool of studies of accelerated classes. The present analysis is therefore based on 23 studies (Table 5). The 23 studies examined modest forms of rapid advancement. Eighteen of the studies examined programs of grade compression (e.g., 4 years in 3). The remaining 5 studies examined programs that extended the calendar to speed up the progress of gifted and talented students (e.g., completing the work of 4 years in 3 school years with five summer sessions). The effects of 21 programs were evaluated after one or more years of accelerated instruction; the effects of the remaining 2 programs were evaluated after only one semester of acceleration. Nine of the programs involved subject-matter acceleration in mathematics, and 14 studies involved comprehensive programs of acceleration. In 6 of the programs, the accelerated classes began in the elementary school years; in the remaining 17, acceleration took place in the junior high grades.

Table 5.

Major Features and Achievement Effect Sizes in 23 Studies of Accelerated Classes

Study	Starting Grade	Duration of Study	Effect Size	Method of Acceleration
Studies with same-age control groups				
Arends & Ford, 1964	7	2 years	1.14	Acceleration in math in Grades 7,8
Enzmann, 1961	9	4 years	0.30	Acceleration in math in Grades 9-12
Fox, 1974	7	1/2 year	0.46	Summer algebra program for Grade 7 girls
Justman, 1953	7	2 years	0.54	Completion of 3 years of school in 2 years
Klausmeier & Ripple, 1963	2	3/4 year	0.80	Placement of bright older pupils from Grade 2 in Grade 4 after 1 summer session
Klausmeier & Wiersma, 1964	9	2 years	1.48	Completion of 6 semesters of math in Grades 9, 10
Ludeman, 1969	7	6 year	0.85	Completion of Grade 7 and 8 math in 1 year
Montgomery,1968	8	5 years	0.84	Accelerated program in Grade 8-12 math
Passow, Goldberg, & Link, 1961	7	3 years	1.34	Acceleration in Grade 7 and 8 math
Rusch, & Clark, 1963	5	3 years	0.80	Completion of Grades 5-8 in 3 years with 5 summer sessions
Simpson & Martison, 1961	1, 9, 12	1 to 3 years	1.04	Completion of Grades 1 and 2 in 1 year; completion of Grades 7-9 in 2 years with 3 summer sessions; enrollment in college courses during Grade 12
Studies with older control groups				
Adler, Pass, & Wright, 1963	9	4 years	0.11	Completion of 5 years program in 4 years
Culbertson, 1963	7	2 years	Ð0.08	Completion of Grades 7-9 in 2 years
Fredstrom, 1964	7	2 1/2 years	Ð0.30	Completion of Grade 7 and 8 math in 1 year
Herr, 1937	7	2 years	0.12	Completion of Grades 7-9 in 2 years
Justman, 1954	7	4 1/2 years	Ð0.04	Completion of Grades 7-9 in 2 years
Klausmeier, 1963	2	3 years	0.76	Placement of bright older pupils from Grade 2 into Grade 4 after 1 summer session
Klausmeier & Wiersma, 1964	7, 9	2 to 3 years	0.20	Completion of Grade 7-9 math and science in 2 years; completion of 6 semesters of math in Grades 9 and 10
Matlin,1965	4	2 years	Ð0.01	Completion of Grades 4-6 in 2 years
Mikkelson, 1962	8	1 year	Ð0.84	Completion of Grade 9 math in Grade 8
Morrison, 1970	5	7 years	Ð0.07	Completion of Grades 5 and 6 in 1 year
Rusch & Clark, 1963	5	3 years	0.00	Completion of Grades 5-8 in 3 years with 5 summer sessions
Unzicker, 1932	7	2 years	Ð0.03	Completion of Grades 7 and 8 in 1 year

The earliest of the studies was published in 1932, and the most recent in 1974. In all, 2 of the studies were from the 1930s, 2 from the 1950s, 17 from the 1960s, and 2 from the 1970s. Of the 23 studies, 13 were found in journal articles; 7 were found in dissertations; and 3 were found in technical reports in the ERIC system.

In most respects, the research methodologies used in the studies of acceleration were similar to the methodologies used in other studies of grouping. Standardized tests were used as the criterion measure of student achievement in all the studies. All but two of the studies were quasi-experiments with equivalent experimental and control groups rather than true experiments in which subjects were randomly assigned to treatment conditions. The major threats to validity in the studies were the lack of controls for teacher and school effects.

In one design characteristic, however, studies of accelerated classes were different from other grouping studies. The 23 studies used two fundamentally different experimental designs that reflected two fundamentally different purposes. In one group of 11 studies, the researcher's purpose was to determine whether accelerated students learned more than initially comparable students who were not accelerated. In these studies, students in the groups being compared were initially equivalent in age and aptitude, but because one group was accelerated and the other was not, the two groups differed in grade level when educational outcomes were measured. A second group of 12 studies had a different purpose. Talented accelerated students often end up in the same classrooms with talented nonaccelerates who are a year or more older. The purpose of the second group of studies was to determine whether the younger accelerates performed as well on tests as did the older nonaccelerates. In studies of this type, the groups being compared were equivalent in grade level and intelligence quotient when outcomes were measured, but the groups differed in both chronological and mental age.

Findings. The distribution of effect sizes in the 23 studies was bimodal in shape. One of the modes of the distribution was equal to 0.0 standard deviations; the other was equal to 0.75 standard deviations. This bimodality confirmed our initial impression that we were dealing with two distinct groups of studies. Studies where the control students were one year older than the accelerated students clustered around the mode of 0.0. Other studies were spread more loosely around the mode of 0.75.

In each of the 11 studies with same-age control groups, the achievement was higher for students in the accelerated classes. The average effect size in these studies was 0.87; the median effect was 0.84. This means that on a grade-equivalent scale the scores of the accelerated students would be approximately one grade higher than the scores of bright, nonaccelerated students of the same age.

In all but 2 of the 12 studies with older control groups, effect sizes were small. In one of the 2 exceptional studies, the effect size was large and positive; in the other study, it was large and negative. The average effect size in the 12 studies, however, was -0.02; the median effect size was 0.0. In the typical study, therefore, the difference in examination performance of accelerates and older nonaccelerates was trivial in size.

Only a small number of studies investigated other outcomes of acceleration, and findings were not entirely consistent from study to study. On the average, however, acceleration appeared to have little or no effect on attitude toward school or school subjects. Acceleration had a strong effect on vocational plans in two studies but trivial effects on student plans in three other studies. The effect on vocational plans apparently varied as a function of program type. There was no evidence of consistent positive or

negative effects from acceleration on student participation in school activities, popularity, or adjustment.

Conclusions. This meta-analysis showed that gifted students are able to handle the academic challenge that accelerated programs provide. Two major findings supported this conclusion. First, talented youngsters who were accelerated into higher grades performed as well as the talented, older nonaccelerates already in those grades. Second, in the subjects in which they had been moved ahead, talented accelerates showed almost a year's advancement over talented same-age nonaccelerates.

The results from studies comparing accelerates with older pupils seemed especially impressive to us because the accelerates were at a clear disadvantage in these studies. In most of the studies, the accelerates were at least one year younger in chronological age. Because performance on standardized tests in subjects such as mathematics and English is strongly influenced by mental age, the accelerates could hardly be expected to equal the older nonaccelerates in test performance. Nonetheless, the accelerates did very well in most of the comparisons. Overall, their performance was indistinguishable from that of older talented, nonaccelerates.

The results of the same-age comparisons were almost as remarkable. It is unusual for groups that are equivalent in general intelligence and age to differ by almost one grade level in performance on achievement tests. Nonetheless, that is the size of the difference between scores of accelerates and nonaccelerates in the average study. In a review of approximately 100 different meta-analysis of findings of educational research, Chen-Lin Kulik and I were not able to find any educational treatment that consistently yielded a higher effect size than this one (J. Kulik & Kulik, 1989).

Perhaps we should not have been surprised to find such strong effects from programs of acceleration. Reviewers have been noting for many years that programs of acceleration almost always produce good results. "Perhaps what is needed," Gallagher suggested in 1969, "is some social psychologist to explore why this procedure is generally ignored in the face of such overwhelmingly favorable results" (p. 541). Getzels and Dillon in 1973 also lamented the lack of interest in acceleration and offered a social psychological explanation:

> Apparently the cultural values favoring a standard period of dependency and formal education are stronger than the social or individual need for achievement and independence. This is an instance of the more general case one remarks throughout education: When research findings clash with cultural values, the values are more likely to prevail. (p. 717)

Special Enriched Classes

Accelerated programs were the first accommodation made for gifted and talented students in age-graded schools, and in the early years of this century, they were the major method by which schools met the special needs of their high-aptitude students. But by the 1920s some educators began to question the wisdom of acceleration. Their main concern was that accelerated programs might not be meeting the emotional and social needs of gifted youngsters. Two alternative approaches came to be emphasized: enrichment in special classes and enrichment in the regular class.

Several factors played a role in the shift of emphasis toward enrichment and away from acceleration. One was the increasing interest in progressive education during the

1920s. From the vantage point of progressive education, accelerated programs did not deal adequately with human individuality. They put too much emphasis on subject-matter learning and too little on the social needs and interests of children. Psychological studies of the time reinforced the notion that children differed from one another in innumerable important ways. Critics of acceleration reasoned, therefore, that few intellectually talented children were ahead to the same degree in all areas, and they warned that many were not physically and socially mature enough to handle advanced classes. No one could say for certain, they noted, which children should be moved ahead in school and which should be left behind.

Some critics of acceleration came to view as an attractive alternative the enriched classes that Leta Hollingworth began setting up in the city schools of New York City in 1916. In these classes, children did not simply follow a telescoped regular curriculum. Instead, they spent about half of their school hours working on the prescribed curriculum, and about half pursuing enriching activities. In classes that Hollingworth set up for seven- to nine-year-olds, for example, enrichment activities included conversational French; the study of biography; study of the history of civilization, with reference to food, water, clothing, shelter and sanitation; and a good deal of extra work in science, mathematics, English composition and music (Gray & Hollingworth, 1931).

These special enriched classes for the gifted did not address all the criticisms raised by Dewey's followers and specialists in child study. For one thing, children were chosen for the classes on the basis of I.Q. scores, and critics believed that I.Q. tests did not do justice to the complex achievements of children. Pupils with the same I.Q. scores could differ greatly in learning strengths and weaknesses, and their differences were even more marked when nonintellective traits were taken into account. Moreover, segregating talented children seemed socially unfair. The critics felt that the experience of both talented children and their age-mates was diminished by the restriction in social contacts that was a necessary result of separate classes.

Some critics of accelerated and enriched classes argued, therefore, that gifted and talented children should be kept in regular classrooms. They proposed that the special needs of such children could be met by providing them with enrichment activities there, and in the 1930s the arguments against special classes prevailed. Schools increasingly tried to meet the needs of gifted and talented youngsters by providing enrichment activities for them in their regular classrooms. Tannenbaum (1958), surveying developments in gifted and talented education during the first half of the 20th century, described the first two decades of the century as a time when the special needs of gifted and talented were met through acceleration, the 1920s as years when enrichment in special classes was the preferred method, and the 1930s and war years as the time of enrichment in regular classrooms.

Study characteristics. An earlier meta-analysis (J. Kulik & Kulik, 1991) covered a total of 26 studies of enrichment in special or regular classes. All but one of the studies cited in the earlier report were suitable for use in this analysis (Table 6). The unsuitable study used a correlational rather than experimental design, and it, therefore, seemed significantly lower in quality than other studies in the pool.

Table 6.

Major Features and Effect Sizes in 25 Studies of Enriched Classes

Study	Starting Grade	Course Content	Duration of Instruction	Achievement Effect Size	Self-esteem Effect Size
Alam, 1968	3	C	7 years	0.43	0.10
Atkinson & O'Connor, 1963	6-7	C	2 years	0.46	
Bell, 1959	5	C	1 year	0.93	0.01
Bent, 1969	4	C	4 years	0.22	
Cluff, 1964	4	C	2 years	0.13	
Doolin, 1956	11	So	26 weeks	0.48	0.18
Enzmann, 1963	9	C	4 years	0.28	
Evans & Marken, 1982	6-8	C	2 years	Ð0.04	
Gray & Hollingworth, 1931	2-5	C	3 years	0.17	
Hinze, 1957	4-8	C	1 semester	Ð0.01	
Howell, 1962	9	C	1 year	1.25	
Ivey, 1965	4	M	28 weeks	0.57	
Karnes et al., 1963	2-5	C	2 1/2 years	0.52	
Kellogg, 1960	4	C	3 years	0.60	
Koukeyan, 1976	4-6	M	26 weeks	0.11	
Long, 1957	11	M	25 weeks	0.40	
Luttrell, 1959	6	C	28 weeks	0.56	Ð0.09
Mahler, 1962	7-8	C	1 year	0.30	
McCall, 1928	2-6	C	2 years	0.60	
McCown, 1960	10	C	3 years	0.36	
Mikkelsen, 1963	7	M	1 year	Ð0.06	
Schwartz, 1943	1-8	C	1 semester	0.34	
Simpson & Martison, 1961	1,5,6,8,11, 12	C	1 year	0.38	0.29
Tremaine, 1979	12	C	1 year	0.55	
Ziehl, 1962	2,3	C	4 years	0.62	

Note. C = Combined; M = Mathematics; So = Social Science.

The 25 studies were carried out over a period of more than a half-century. The earliest of the studies was published in 1928; 1 came from the 1930s; 1 from the 1940s; 5 from the 1950s; 14 from the 1960s; 2 from the 1970s. The most recent study was published in 1982. Eight of the studies were described in journal articles, 14 in dissertations, and 3 in technical reports in the ERIC system.

The studies were diverse in several important respects. Twenty-two of the studies examined effects of enrichment in separate classes, but 3 examined enrichment in regular classrooms. In 20 studies, the curriculum was adjusted to group ability, but in 3 studies, children followed the same basic curriculum as in the regular classroom. In 2 of the studies, placement in the enriched program was based on a specific ability; in the remaining 23 studies, placement was based on a measure of general ability. Sixteen of the studies were carried out in Grades 1 through 6; 5 were carried out in Grades 7 through 9; and 4 were carried out in Grades 10 through 12. Five of the studies involved enrichment in a single subject, and 20 studies involved enrichment in several subjects. All of the studies examined effects after one year or more of enrichment.

In methodological characteristics, the 23 studies of enrichment were similar to other grouping studies that we have examined. All of the studies used standardized tests as criterion measures of student achievement. The major threats to study validity were the assignment of intact groups to treatments (rather than random assignment of subjects to treatments) and the lack of controls for instructor and school effects.

Student achievement. Twenty-two of the 25 studies found that talented students achieved more when they were taught in enriched classrooms. In the remaining 3 studies, performance of talented students was better when they were taught in mixed-ability classes. In 13 of the 25 studies, the achievement differences in enriched versus mixed-ability classes were great enough to be considered statistically significant. Each of these 13 studies favored enriched classes for talented students.

The average effect size in the 25 studies was 0.41. The median effect size was 0.40. This effect size is significantly greater than zero. It is also significantly greater than the average effect of 0.03 standard deviations for XYZ programs, but it is not significantly greater than the average effect sizes for programs of cross-grade grouping (0.33) and within-class grouping (0.25). An effect size of 0.41 also means that in the typical study, approximately 66% of the talented students in the special classes outperform the typical talented student in a mixed-ability class.

Although the effect of enriched classes for the gifted was modest in the typical study in this group, effects varied in size from a low of -0.06 to a high of 1.25 standard deviations. The variation was great enough to lead us to suspect that factors other than grouping played a role in determining study outcome. We were unable to establish through further analyses, however, that study features were significantly related to achievement outcomes. The small number of studies available for analysis might account in part for this failure to find significant relationships.

Student self-esteem. Five of the 25 studies of separate classes for the gifted examined effects on self concept. In 4 of the 5 studies, self-concepts were more favorable when the talented students were taught in separate classes. In the remaining study, self-concepts were more positive when talented students were taught in heterogeneous classes. The size of the effect was small or trivial, however, in all the studies. The average effect size in all 6 studies was 0.10.

This effect is quite different, however, from the effect of XYZ grouping on high-aptitude students. The self-esteem of higher aptitude students goes down slightly when they are placed in the top groups in XYZ programs. Presumably, higher aptitude students lose some of their self-assurance when they are placed in classes whose members are all intellectual peers. In contrast, the self-esteem of gifted students may go up slightly in enriched programs. Teachers in enrichment programs may be better prepared to help students deal with emotional and social pressures of giftedness.

Conclusions. The main effects of enriched classes are clear. These classes contribute to the intellectual progress of higher aptitude students. Gifted and talented students gain more academically from such classes than they do in regular mixed-ability classes. The students in enriched classes also maintain their sense of self-confidence. There is no evidence available to suggest that students lose self-esteem in programs of enrichment.

These effects seem to us to be remarkable ones, given the goals of most enrichment programs and the criterion tests used to measure their effects. Most enrichment programs are meant to give students varied experiences that would not be

available in regular classrooms. Teachers of enriched classes do not ordinarily try to provide more work on the basic skills. In fact, many teachers cut back on instruction in the basic skills on the assumption that gifted and talented learners can learn the basics in less than the ordinary amount of time. In Gray and Hollingworth's (1931) study, for example, seven- to nine-year-olds spent half of their school time on music, art, foreign languages, and cultural pursuits. Gray and Hollingworth estimated that children in their enrichment programs spent about half as much time on basic skills as did children in control programs.

It is important to note, therefore, that the standardized achievement tests used to evaluate the effects of most enrichment programs do not measure esthetic appreciation, attainment in music and art, achievement in foreign languages, and the like. They focus more on what is taught in conventional classes than they do on what is taught in enriched classes. The tests measure reasonably well the common objectives of enriched and conventional classes, but they do not measure the unique objectives of enriched classes. When children from enriched and regular classes are compared in performance on standardized tests, the two groups are competing on an uneven field, and the test bias favors the children from the regular classrooms. Despite the bias, youngsters from enriched classrooms not only equal the test performance of these from regular classes, they exceed it.

Summary and Conclusions

Researchers have been reviewing the literature on ability grouping for more than 60 years. The older reviews on the topic can serve as a good source of hypotheses about grouping effects, but they are not very helpful as guides to practice. For one thing, the old reviews present only imprecise, nonquantitative conclusions. They may report that grouping programs help students learn, for example, but they seldom report the size of the learning gains that can be expected from grouping. Even worse, the old reviews do not agree about the fundamental research findings. For every review that reports that grouping helps children, there is another that reports that grouping harms them. Conclusions in the older reviews too often reflect prevailing educational philosophies and too seldom reflect the actual research results.

Reviewers have, therefore, begun using objective, scientific methods to summarize and interpret findings on ability grouping. With Chen-Lin Kulik, I carried out one series of meta-analyses (e.g., J. Kulik & Kulik, 1991), and Slavin carried out another set (Slavin, 1987, 1990b). Findings from the two sets of meta-analyses agree quite well, but conclusions in the meta-analyses differ. The disagreement seems to stem from differences in the scope of the analyses. Chen-Lin Kulik and I included studies of special classes for the gifted and talented in our meta-analyses, and in our conclusions we paid special attention to the results from such classes. Slavin excluded all studies of special classes for the gifted and talented from his analyses. His conclusions were, therefore, based largely on findings from XYZ grouping programs.

A careful re-analysis of findings from all the studies included in the two sets of meta-analyses showed once again that higher aptitude students benefit academically from ability grouping. The academic benefits are positive but usually small when the grouping is done as a part of a broader program for students of all abilities. For example, XYZ classes, in which little or no effort is made to adjust curriculum to group ability level, raise the test scores of higher ability students by about 0.1 standard deviations. Within-class and cross-grade programs, which entail curricular adjustment, boost test scores of higher aptitude students by about 0.2 to 0.3 standard deviations.

Benefits are positive and often large in special classes for the gifted and talented. The larger gains are usually found in classes that are accelerated. Classes in which talented children cover four grades in three years, for example, usually boost achievement levels a good deal. Test scores of children accelerated in this fashion are about one year higher on a grade-equivalent scale than they would be if the children were not accelerated. Enriched classes, in which students have a varied educational experience, boost student achievement by more moderate amounts. The average gain on a grade-equivalent scale is 4 months in a typical program. A gain of this size is still impressive, given that some enriched classes spend as much as half their time on cultural material (e.g., foreign languages and music) not covered on standard achievement tests.

Grouping programs usually have smaller effects on middle and lower aptitude learners. XYZ classes, for example, have virtually no effect on the achievement of such students. Test scores of middle and lower aptitude students learning in XYZ classes are indistinguishable from those of similar students in mixed-ability classes. Cross-grade and within-class programs, however, usually raise test scores of middle and lower aptitude pupils by between 0.2 and 0.3 standard deviations. The clear adjustment of curriculum to pupil ability in within-class and cross-grade programs may be the key to their effectiveness.

Evidence on the noncognitive outcomes of grouping is less clear. Despite their importance, noncognitive instructional outcomes are not often studied by educational researchers, and only tentative conclusions can be drawn. One of these conclusions is that grouping programs usually have only small effects on student self-esteem. The programs certainly do not lead talented students to become self-satisfied and smug, nor do they cause a precipitous drop in the self-esteem of lower aptitude students. If anything, XYZ classes seem to have effects in the opposite direction. XYZ programs may cause quick learners to lose a little of their self-assurance, and they may cause slower learners to gain some badly needed self-confidence. The available literature also suggests that grouping programs may have some program-specific effects in noncognitive areas. For example, a few programs of accelerated instruction clearly have an effect on the vocational plans of youngsters; other programs of acceleration have no consistent effect. Design of specific programs undoubtedly plays a role.

These conclusions are obviously quite different from the well-known conclusions about grouping reached by Oakes (1985) in her book *Keeping Track*. According to Oakes, students in the top tracks gain nothing from grouping and other students suffer clear and consistent disadvantages, including loss of academic ground, self-esteem, and ambition. Oakes also believes that tracking is unfair to students because it denies them their right to a common curriculum. She therefore calls for the *de-tracking* of American schools. De-tracked schools would provide the same curriculum to all, and they would not provide special educational opportunities to any on the basis of ability, achievement, or interests.

Oakes's conclusions, however, are based on her own selective and idiosyncratic review of older summaries of the literature and on her uncontrolled classroom observations. Objective analysis of findings from controlled studies provides no support for her speculations. Whereas Oakes believes that grouping programs are unnecessary, ineffective, and unfair, I conclude that the opposite is true. American education would be harmed by the elimination of programs that provide instruction adapted to the aptitude, achievement, and interests of groups with special educational needs.

The effects of de-tracking would vary according to the type of grouping program that was eliminated. If typical XYZ classes were eliminated from all schools, the

achievement level of the country's brightest students would fall slightly, but the effects would not be noticeable on most other students. If the grouping programs that were eliminated were ones that actually adjusted methods and materials to student aptitude, the damage to student achievement would be greater, and the effects would be felt more broadly. Both higher and lower aptitude students would suffer academically from such de-tracking. But the damage would be truly profound if, in the name of de-tracking, schools eliminated enriched and accelerated classes for their brightest learners. The achievement level of such students would fall dramatically if they were required to move at the common pace. No one can be certain that there would be a way to repair the harm that would be done.

Guidelines From Meta-analytic Studies of Ability Grouping

Guideline 1: Although some school programs that group children by ability have only small effects, other grouping programs help children a great deal. Schools should therefore resist calls for the wholesale elimination of ability grouping.

Research support: The effect of a grouping program depends on its features. It is important to distinguish among programs that (a) make curricular and other adjustments for the special needs of highly talented learners, (b) make curricular adjustments for several ability groups at a grade level, and (c) provide the same curriculum for all ability groups in a grade.

Guideline 2: Highly talented youngsters profit greatly from work in accelerated classes. Schools should therefore try to maintain programs of accelerated work.

Research support: Talented students from accelerated classes outperform nonaccelerates of the same age and IQ by almost one full year on the grade-equivalent scales of standardized achievement tests.

Guideline 3: Highly talented youngsters also profit greatly from an enriched curriculum designed to broaden and deepen their learning. Schools should therefore try to maintain programs of enrichment.

Research support: Talented students from enriched classes outperform control students from conventional classes by 4 to 5 months on grade-equivalent scales.

Guideline 4: Bright, average, and slow youngsters profit from grouping programs that adjust the curriculum to the aptitude levels of the groups. Schools should try to use ability grouping in this way.

Research support: Cross-grade and within-class programs are examples of programs that provide both grouping and curricular adjustment. Children from such grouping programs outperform control children from mixed classes by 2 to 3 months on grade-equivalent scales.

Guideline 5: Benefits are slight from programs that group children by ability but prescribe common curricular experiences for all ability groups. Schools should not expect student achievement to change dramatically with either establishment or elimination of such programs.

Research support: In XYZ grouping, all ability groups follow the same course of study. Middle and lower ability students learn the same amount in schools with and without XYZ classes. Higher ability students in schools with XYZ classes outperform equivalent students from mixed classes by about one month on a grade-equivalent scale.

References

Adamson, D. P. (1972). Differentiated multi-track grouping vs. uni-track educational grouping in mathematics. *Dissertation Abstracts International, 32,* 3771A. (University Microfilms No. 72-2564)

Adkison, M. R. (1968). A comparative study of pupil attitudes under conditions of ability and heterogeneous grouping. *Dissertation Abstracts, 28,* 3869A. (University Microfilms No. 66-3322)

Adler, M. J., Pass, L. E., & Wright, E. N. (1963). A study of the effects of an acceleration programme in Toronto secondary schools. *Ontario Journal Education Research, 6*(Autumn), 1.

Alam, S. J. (1969). A comparative study of gifted students enrolled in separate and regular curriculums. *Dissertation Abstracts, 29,* 3354A. (University Microfilms No. 69-6057)

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

Arends, R. H. & Ford, P. M. (1964). *Acceleration and enrichment in the junior high school. A follow-up study*. Olympia, WA.: Washington Office of the State Superintendent of Public Instruction. (ERIC Document Reproduction Service No. ED 001 220)

Atkinson, J. W., & O'Connor, P. (1963). *Effects of ability grouping in schools related to individual differences in achievement-related motivation, Final Report*. Ann Arbor, MI: Michigan University. (ERIC Document Reproduction Service Number ED 003 249)

Ayres, L. P. (1909). *Laggards in our schools*. New York: Charities Publications Committee.

Bailey, H. P. (1968). A study of the effectiveness of ability grouping on success in first year algebra. *Dissertation Abstracts, 28,* 3061A. (University Microfilms No. 68-1249)

Ball, S. J. (1981). *Beachside Comprehensive: A case-study of secondary schooling*. Cambridge, England: Cambridge University Press.

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

Barker Lunn, J. C. (1970). *Streaming in the primary school*. Hove, Sussex, England: King, Thorne, & Stace.

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

Barton, D. P. (1964). An evaluation of ability grouping in ninth grade English. *Dissertation Abstracts, 25,* 1731. (University Microfilms No. 64-9939)

Begle, E. G. (1975). *Ability grouping for mathematics instruction: A review of the empirical literature*. Palo Alto, CA: Stanford Mathematics Education Study Group, Stanford University. (ERIC Document Reproduction Service No. Ed 116 938)

Begle, E. G. (1976). *Acceleration and enrichment in the junior high school. A follow-up study*. Olympia, WA: Washington Office of the State Superintendent of Public Instruction. (ERIC Document reproduction Service No. Ed 001 220)

Bell, M. E. (1959). A comparative study of mentally gifted children heterogeneously and homogeneously grouped. *Dissertation Abstracts, 19,* 2509. (University Microfilms No. 00-22,982)

Bent, L. G., McDonald, R., Rothney, J., & Sowards, W. (1969). *Grouping of the gifted: An experimental approach*. Peoria, IL: Bradley University. (ERIC Document Reproduction Service No. ED 040 519)

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.

Bicak, L. (1963). Achievement in eighth grade science by heterogeneous and homogeneous classes. *Dissertation Abstracts, 23*, 2367. (University Microfilms No. 63-1192)

Billet, R. O. (1928). A controlled experiment to determine the advantages of homogeneous grouping. *Educational Research Bulletin, 7,* 190-196.

Bills, R. E., Vance, E. L., & McLean, O. S. (1951). An index of adjustment and values. *Journal of Consulting Psychology,* 257-261.

Borg, W. R. (1964). *An evaluation of ability grouping*. Logan, UT: Utah State University. (ERIC Document Reproduction Service Number ED 001 177)

Breidenstine, A. G. (1937). 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.

Burr, M. Y. (1931). *A study of homogeneous grouping*. New York: Bureau of Publications, Teachers College, Columbia University.

Burtt, H. E., Chassel, L. M., & Hatch, E. M. (1923). Efficiency of instruction in unselected and selection sections in elementary psychology. *Journal of Educational Psychology, 14,* 154-161.

Campbell, A. L. (1965). A comparison of the effectiveness of two methods of class organization for the teaching of arithmetics in junior high school. *Dissertation Abstracts, 26,* 813-814. (University Microfilms Order No. 65-06726)

Carson, R. M, & Thompson, J. M. (1964). The Joplin plan and traditional reading groups. *Elementary School Journal, 65,* 38-43.

Chiotti, J. F. (1961). *A progress comparison of ninth grade students in mathematics from three school districts in the state of Washington with varied methods of grouping*. Unpublished doctoral dissertation, University of Northern Colorado.

Chismar, M. H. (1972). A study of the effectiveness of cross-level grouping of middle school under-achievers for reading instruction. *Dissertation Abstracts International, 32,* 5101. (University Microfilms No. 72-9249)

Cignetti, M. J. (1974). A study of intraclass grouping and traditional grouping on students' terminal achievements during the last nine weeks in first semester typewriting. *Dissertation Abstracts International, 35,* 5765A. (University Microfilms No. 75-05121)

Cluff, J. E. (1964). The effect of experimentation and class reorganization on the scholastic achievement of selected gifted sixth grade pupils in Wichita, Kansas. *Dissertation Abstracts, 25,* 1676-1677. (University Microfilms No. 64-10,059)

Cochran, J. R. (1968). Grouping students in junior high school. *Educational Leadership, 18,* 414-419.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Coleman, J. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Cooper, H. M. (1984). The integrative research review: A systematic approach. *Applied social research methods series, vol. 2*. Beverly Hills, CA: Sage.

Cornell, E. L. (1936). Effects of ability grouping determinable from published studies. In G. M. Whipple (Ed.), *The ability grouping of pupils,* 35th Yearbook of the National Society for the Study of Education (Part I, pp. 289-304). Bloomington, IL: Public School Publishing.

Courtis. S. A. (1925). Ability-grouping in Detroit schools. In G. M. Whipple (Ed.), *The ability grouping of pupils,* 35th Yearbook of the National Society for the Study of Education (Part I, pp. 44-47). Bloomington, IL: Public School Publishing.

Culbertson, W. P. (1963). An evaluation of an accelerated program in the junior high school. *Dissertation Abstracts, 24,* 1460. (University Microfilms No. 63-6392)

Daniels, J. C. (1961). The effects of streaming in the primary school: I. What teachers believe. II. Comparison of streamed and unstreamed schools. *British Journal of Educational Psychology, 31,* 69-78, 119-126.

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

DeGrow, G. S. (1964). A study of the effects of the use of vertical reading ability groupings for reading classes as compared with heterogeneous groupings in grades four, five, and six in the Port Huron Area Schools of Michigan over a three-year period. *Dissertation Abstracts, 24,* 3166-3167. (University Microfilms No. 64-804)

Dewar, J. A. (1963). Grouping for arithmetic instruction in sixth grade. *Elementary School Journal, 63,* 266-269.

Dewey, J. (1929). *Characters and events.* New York: Henry Holt.

Doolin, R. B. (1956). An experiment with moderately gifted children in the public high schools of Cedar Rapids, Iowa. *Dissertation Abstracts, 16,* 111. (University Microfilms No. 56-2170).

Drews, E. M. (1963). *Student abilities grouping patterns and classroom interactions.* East Lansing, MI: Office of Research and Publications, Michigan State University.

Dvorak, A., & Rae, J. J. (1929). Comparison of achievement of superior children in segregated and unsegregated first-grade classes. *Elementary School Journal, 24,* 380-387.

Dyson, E. (1967). A study of ability grouping and the self-concept. *Journal of Educational Research, 60,* 403-405.

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)

Ekstrom, R. B. (1961). Experimental studies of homogeneous grouping: A critical review. *School Review, 69,* 216-226.

Enzmann, A. M. (1961). An evaluation of the science and arts curriculum for selected students of high ability at Cass technical high school, Detroit, Michigan. *Dissertation Abstracts International, 22,* 3484. (University Microfilms No. AAD62-00907)

Enzmann, A. M. (1963). A comparison of academic achievement of gifted students enrolled in regular and in separate curriculums. *Gifted Child Quarterly, 7,* 176-179.

Erickson, G. R. (1973). A study of the self-esteem and academic self-concepts of ability- and randomly-grouped ninth graders. *Dissertation Abstracts International, 33,* 5550A. (University Microfilms No. 73-10548)

Evans, E. D., & Marken, D. (1982). Multiple outcome assessment of special class placement for gifted students: A comparative study. *Gifted Child Quarterly, 26*, 126-132.

Fick, W. W. (1963). The effectiveness of ability grouping in seventh grade core classes. *Dissertation Abstracts, 23,* 2753. (University Microfilms No. 63-794)

Findley, W. G., & Bryan, M. (1971). *Ability grouping: 1970 Status, impact, and alternatives.* Athens: 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)

Flesher, M. A., & Pressey, S. L. (1955). War-time accelerates ten years after. *Journal of Educational Psychology, 46,* 228-238.

Fogelman, K., Essen, J., & Tibbenham, A. (1978). Ability grouping in secondary schools and attainment. *Educational Studies, 4,* 201-212.

Fowlkes, J. G. (1931). Homogeneous or heterogeneous groupÑwhich? *The Nation's Schools, 8,* 74-78.

Fox, L. H. (1974). Facilitating the development of mathematical talent in young women. *Dissertation Abstracts International, 35*, 3553. (University Microfilms No. AAD74- 29027)

Fredstrom, P. N. (1964). An evaluation of the accelerated mathematics program in the Lincoln, Nebraska, Public Schools. *Dissertation Abstracts International, 25,* 5628. (University Microfilms No. AAD65-02772)

Fund for the Advancement of Education of the Ford Foundation. (1957). *They went to college early*. New York: Research Division of the Fund.

Gallagher, J. J. (1969). Gifted children. In R. L. Ebel (Ed.), *Encyclopedia of educational research* (4th ed., pp. 537-544). New York: Macmillan.

Gamoran, A., & Berends, M. (1961). The effect of stratification in secondary schools: Synthesis of survey and ethnographic research. *Review of Educational Research, 57,* 415-435.

Getzels, J. W., & Dillon, J. T. (1973). The nature of giftedness and the education of the gifted. In R. M. W. Travers (Ed.), *Second handbook of research on teaching* (pp. 689-731). Chicago: Rand McNally.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3-8.

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

Goldberg, M. L. (1958). Recent research on the talented. *Teachers College Record, 60,* 150-163.

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

Good, C. V. (1973). *Dictionary of education* (3rd ed). New York: McGraw-Hill.

Goodlad, J. I. (1984). *A place called school*. New York: McGraw-Hill.

Gray, H. A., & Hollingworth, L. S. (1931). The achievement of gifted children enrolled and not enrolled in special opportunity classes. *Journal of Educational Research, 24,* 255-261.

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

Halliwell, J. W. (1963). A comparison of pupil achievement in graded 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.

Hartill, R. W. (1936). *Homogeneous grouping*. New York: Bureau of Publications, Teacher's College, Columbia University.

Heathers, G. (1969). Grouping. In R. Ebel (Ed.), *Encyclopedia of educational research* (4th ed., pp. 559-570). New York: Macmillan.

Herr, W. A. (1937). Junior high school accelerants and their peers in senior high school. I. Scholastic achievement. *The School Review, 45,* 186-195.

Hinze, R. H. (1957). Achievement of fast learners in a partially segregated elementary school program, with special reference to science instruction. *Dissertation Abstracts, 18,* 496. (University Microfilms No. 58-4203)

Hoge, R. D., & Renzulli, J. S. (1992). *Self-concept and the gifted child* (RBDM9104). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Holy, T. C., & Sutton, D. H. (1930). Ability grouping in the ninth grade. *Educational Research Bulletin, 9,* 419-422.

Howell, W. J. (1962). Grouping of talented students leads to better academic achievement in secondary school. *Bulletin of NASSP, 46*, 67-73.

Husen, T., & Svensson, N.-E. (1960). Pedagogic milieu and development of intellectual skills. *School Review, 68,* 36-51.

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

Ivey, J. D. (1965). Computation skills: Results of acceleration. *The Arithmetic Teacher, 12*, 39-42.

Jackson, G. (1980). Methods for integrative reviews. *Review of Educational Research, 50,* 438-460.

Jencks, C. (1972). *Inequality*. New York: Basic Books.

Johnston, H. J. (1973). The effect of grouping patterns on first-grade children's academic achievement and personal and social development. *Dissertation Abstracts International, 34,* 2461A. (University Microfilms No. 73-25893)

Jones, D. M. (1948). An experiment in adaptation to individual differences. *Journal of Educational Psychology, 39*, 247-272.

Jones, J. C., Moore, J. W., & Van Devender, F. (1967). A comparison of pupil achievement after one and one-half and three years in a nongraded program. *Journal of Educational Research, 61,* 75-77. (University Microfilms No. 60-5240)

Justman, J. (1953). A comparison of the functioning of intellectually gifted children enrolled in special progress and normal progress classes in junior high school. *Dissertation Abstracts, 14,* 65. (University Microfilms No. AAD00-06641)

Justman, J. (1954). Academic achievement of intellectually gifted accelerants and non-accelerants in senior high school. *School Review, 62,* 142-150.

Karnes, M. B., McCoy, G., Zehrbach, R. R., Wollersheim, J. P., & Clarizio, H. F. (1963). The efficacy of two organizational plans for underachieving intellectually gifted children. *Exceptional Children, 29,* 438-446.

Keliher, A. C. (1931). *A critical study of homogeneous grouping*. New York: Bureau of Publications, Teachers College, Columbia University.

Kellogg, R. M. (1960). An analysis of the achievement of segregated and non-segregated gifted pupils. *Dissertation Abstracts, 21,* 2630.

Kerckhoff, A. C. (1986). Effects of ability grouping in British secondary schools. *American Sociological Review, 51,* 842-858.

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

Klausmeier, H. J. (1963). Effects of accelerating bright older elementary pupils: A follow up. *Journal of Educational Psychology, 54,* 165-171.

Klausmeier, H. J., & Ripple, R. E. (1963). Effects of accelerating bright older pupils from second to fourth grade. *Journal of Educational Psychology, 53*, 93-100.

Klausmeier, H. J., & Wiersma, W. (1964). Effects of condensing content in mathematics and science in the junior and senior high school. *School Science and Mathematics, 64*, 4-11.

Kline, R. E. (1964). A longitudinal study of the effectiveness of the track plan in the secondary schools of a metropolitan community. *Dissertation Abstracts, 25*, 324. (University Microfilms No. 64- 4257)

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

Koukeyan, B. B. (1976). Evaluation of a vertical- horizontal enrichment program for the math-gifted students in fourth, fifth and sixth grades. Dissertation Abstracts International, 37, 5587A. (University Microfilms No. 77- 04835)

Kulik, C.-L. C., & Kulik, J. A. (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. C., & 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. (ERIC Document Reproduction Service No. ED 255 329)

Kulik, J. A., & Kulik, C.-L. C. (1984). Effects of accelerated instruction on students. *Review of Educational Research, 54,* 409-426.

Kulik, J. A., & Kulik, C.-L. C. (1987). Effects of ability grouping on student achievement, *Equity and Excellence, 23,* 22-30.

Kulik, J. A., & Kulik, C.-L. C. (1989). Meta-analysis in educational research [monograph]. *International Journal of Educational Research, 13,* 221-340.

Kulik, J. A., & Kulik, C.-L. C. (1991). Ability grouping and gifted students. In N. Colangelo and G. A. Davis (Eds.), *Handbook of gifted education* (pp. 178-196). Boston, MA: Allyn & Bacon.

Kulik, J. A., & Kulik, C.-L. C. (in press). Meta-analytic findings on grouping programs. *Gifted Child Quarterly.*

Lacey, C. (1970). *Hightown Grammar.* Manchester, England: Manchester University Press.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Long, R. G. (1957). A comparative study of the effects of an enriched program for the talented in advanced algebra classes. *Dissertation Abstracts, 18,* 529. (University Microfilms No. 00-24831)

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

Lovell, J. T. (1960). Bay High School experiment. *Educational Leadership, 17*, 383-387.

Ludeman, C. J. (1969). A comparison of achievement in an accelerated program and a standard program of high school mathematics in Lincoln, Nebraska, Schools. *Dissertation Abstracts International, 31,* 299. (University Microfilms No. AAD70-12271)

Luttrell, J. (1959). A comparative investigation of academic achievement and personality development of gifted sixth grade pupils in a special class and in regular classrooms in the public school of Greensboro, North Carolina. *Dissertation Abstracts, 19,* 2536. (University Microfilms No. 59-52)

Mahler, F. L. (1962). A study of achievement differences in selected junior high school gifted students heterogeneously or homogeneously grouped. *Dissertation Abstracts, 22,* 2267. (University Microfilms No. 61-5675)

Marascuilo, L. A., & McSweeney, M. (1972). Teaching and minority student attitudes and performance. *Urban Education, 6,* 303-319.

Martin, W. B. (1959). Effects of ability grouping on junior high school achievement. *Dissertation Abstracts, 19,* 2810. (University Microfilms No. 59-1108).

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

Matlin, J. P. (1965). Some effects of a planned program of acceleration upon elementary school children. *Dissertation Abstracts, 26,* 827. (University Microfilms No. 65-8293)

McCall, W. A. (1928). Comparison of the educational progress of bright pupils in accelerated and in regular classes. *Twenty-seventh yearbook of the National Society for the Study of Education, Part II.* Bloomington, IL: Public School Publications.

McCown, G. W. (1961). A critical evaluation of the four track curriculum program of the District of Columbia Senior High School with recommendations for improvements. *Dissertation Abstracts, 21,* 2558. (University Microfilms No. 60-4928)

Mikkelson, J. (1963). An experimental study of selective grouping and acceleration in junior high school mathematics. *Dissertation Abstracts, 23,* 4226. (University Microfilms No. 63-2323)

Miles, C. C. (1954). Gifted children. In L. Carmichael (Ed.), *Manual of child psychology* (pp. 984-1063). New York: John Wiley & Sons.

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

Montgomery, W. G. (1968). An analysis and appraisal of the Sioux City, Iowa, secondary school accelerated mathematics program. *Dissertation Abstracts, 29,* 2489. (University Microfilms No. AAD69-03125)

Moorehouse, W. F. (1964). Interclass grouping for reading instruction. *Elementary School Journal, 64*, 280-286.

Morgan, E. F., Jr., & 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 and homogeneous (ability) grouping on the academic achievement and personal-social adjustment of sixth-grade children. *Dissertation Abstracts, 24,* 1054. (University Microfilms No. 63-6560)

Morrison, W. A. (1970). A comparative study of secondary school academic achievement and social adjustment of selected accelerated and non-accelerated elementary pupils. *Dissertation Abstracts International, 31,* 2015-A. (University Microfilms No. 70-21,166)

Mort, P. R. (1928). *The individual pupil in the management of class and school.* New York: American Book.

Moses, P. J. (1966). A study of the effects of inter-class grouping on achievement in reading. *Dissertation Abstracts, 26,* 4342. (University Microfilms No. 66-741)

Newbold, D. (1977). *The Banbury group enquiry*. Oxford, England: NEER Publishing.

Nichols, N. (1969). Interclass grouping for reading instruction. *Educational Leadership Research Supplement, 26*, 588-592.

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

Oakes, J., & Lipton, M. (1990). *Making the best of schools*. New Haven, CT: Yale University Press.

Otto, H. J. (1941). Elementary education - II. Organization and administration. In W. S. Monroe (Ed.), *Encyclopedia of Educational Research* (1st ed., pp. 428-446). New York: Macmillan.

Otto, H. J. (1950). Elementary education - III. Organization and administration. In W. S. Monroe (Ed.), *Encyclopedia of Educational Research* (rev. ed., pp. 376-388). New York: Macmillan.

Passow, A. H. (1958). Enrichment of education for the gifted. In N. Henry (Ed.), *Education for the gifted*. 57th Yearbook of the National Society for the Study of Education (Part II, pp. 193-221). Chicago: University of Chicago Press.

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

Passow, A. H., Goldberg, M. L., & Link, F. R. (1961). Enriched mathematics for gifted junior high school students. *Educational Leadership, 18,* 442-452.

Peterson, R. L. (1967). An experimental study of the effects of ability grouping in grades seven and eight. *Dissertation Abstracts, 28*, 130A. (University Microfilms No. 67-7768)

Petty, M. C. (1953). *Intraclass grouping in the elementary schools*. Austin, TX: The University of Texas Press.

Pevec, A. E. (1965). Some problems of academically accelerated senior boys in selected high schools of the Catholic diocese of Cleveland. *Dissertation Abstracts, 25*, 6350. (University Microfilms No. 65-2331)

Platz, E. F. (1965). The effectiveness of ability grouping in general science classes. *Dissertation Abstracts, 26*, 1459-1460A. (University Microfilms No. 65-6914)

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

Purdom, T. L. (1929). *Value of homogeneous grouping*. Baltimore: Warwick & York.

Putbrese, L. M. (1971). An investigation into the effect of selected patterns of grouping upon arithmetic achievement. *Dissertation Abstracts International , 32,* 5113A. (University Microfilms No. 72-08388)

Rankin, P. T., Anderson, C. T., & Bergman, W. G. (1936). Ability grouping in the Detroit individualization experiment. In G. M. Whipple (Ed.), *The grouping of pupils,* 35th Yearbook of the National Society for the Study of Education (Part I, pp. 277-288). Bloomington, IL: Public School Publishing.

Rogers, K. B. (1991, May). *A best-evidence synthesis of research on acceleration options for gifted students*. Paper presented at the Henry B. and Jocelyn Wallace National Research Symposium on Talent Development. Iowa City: University of Iowa.

Rogers, K. B. (1992). *The relationship of grouping practices to the education of the gifted and talented learner* (RBDM9102). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Rosenbaum, J. E. (1980). Social implications of educational grouping. In D. C. Berliner (Ed.), *Review of Research in Education* (Vol. 8, pp. 361-401). Washington, DC: American Educational Research Association.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

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

Rusch, R. R., & Clark, R. M. (1963). Four years in three: An evaluation. *Elementary School Journal, 63*, 281-285.

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

Sarthory, J. A. (1968). The effects of ability grouping in multi-cultural school situations. *Dissertation Abstracts, 29,* 457A. (University Microfilms No. 68-11664)

Schwartz, W. P. (1942). *Effects of homogeneous classification on the scholastic achievement and personality development of gifted pupils in the elementary and junior high school*. Unpublished doctoral dissertation, New York University, New York City.

Shane, H. G. (1960). Grouping in the elementary school. *Phi Delta Kappan, 41,* 314-317.

Shields, J. M. (1927). Teaching reading through ability-grouping. *Journal of Educational Methods, 7*, 7-9.

Simpson, R. E., & Martison, R. A. (1961). *Educational programs for gifted pupils: A report to the California legislature prepared pursuant to Section 2 of Chapter 2385, Statutes of 1957*. Sacramento, CA: California State Department of Education. (ERIC Document Reproduction Service Number ED 100 072)

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research, 57,* 293-336.

Slavin, R. E. (1990a). Ability grouping in secondary schools: A response to Hallinan. *Review of Educational Research, 60,* 505-507.

Slavin, R. E. (1990b). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60,* 471-499.

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

Smith, W. M. (1960). The effect of intra-class ability grouping on arithmetic achievement in grades two through five. *Dissertation Abstracts, 21,* 563. (University Microfilms No. 60-02984)

Spence, E. S. (1959). Intra-class grouping of pupils for instruction in arithmetic in the intermediate grades of the elementary school. *Dissertation Abstracts, 19,* 1682. (University Microfilms No. 58-5635)

Stanley, J. C. (1980). On educating the gifted. *Educational Researcher, 9,* 8-12.

Stoakes, D. W. (1964). *An educational experiment with the homogeneous grouping of mentally advanced and slow learning students in the junior high school*. Unpublished doctoral dissertation, University of Colorado.

Svensson, N.-E. (1962). *Ability grouping and scholastic achievement*. Stockholm: Alqvist & Wiksell.

Symonds, P. M. (1931). Homogeneous grouping. *Teachers College Record, 32,* 501-517.

Tannenbaum, A. J. (1958). History of interest in the gifted. In N. Henry (Ed.), *Education for the gifted*. 57th Yearbook of the National Society for the Study of Education (Part II, pp. 21-38). Chicago: University of Chicago Press.

Tauber, M. C. (1963). An experimental study of the relationship between certain selected social and emotional factors and ability grouping of high school students. *Dissertation Abstracts, 23,* 3263. (University Microfilms No. 63-1353)

Terman, L. M., & Oden, M. H. (1947). *The gifted child grows up*. Stanford, CA: Stanford University Press.

Thompson, G. W. (1974). The effects of ability grouping upon achievement in eleventh grade American history. *Journal of Experimental Education, 42,* 76-79.

Thorpe, L. P., Clark, W. W., & Tiegs, E. R. (1953). *California Test of Personality - Elementary Form AA*. Los Angeles: California Test Bureau.

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)

Tremaine, C. D. (1979). Do gifted programs make a difference? *Gifted Child Quarterly. 23*, 500-517.

Tunley, R. (1957, October 26). Johnny *can* read in Joplin. *Saturday Evening Post,* 107-108, 110.

Turney, A. H. (1931). The status of ability grouping. *Educational Administration and Supervision, 17,* 110-127.

Unzicker, S. P. (1932). A study of acceleration in the junior high school. *The School Review, 40,* 346-356.

Vakos, H. N. (1969). The effect of part-time grouping on achievement in social studies. *Dissertation Abstracts International, 30,* 2271A. (University Microfilms No. 69-20066)

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

Wardrop, J. L., Cook, D. M., Quilling, M., Klausmeier, H. J., Espeseth, C., & Grout, C. (1967). *Research and development activities in R/I units of two elementary schools of Manitowoc, Wisconsin, 1966-1967.* Madison, WI: University of Wisconsin. (ERIC Document Reproduction Service No. ED 019 796)

Whipple, G. M. (1919). *Classes for gifted children.* Bloomington, IL: Public School Publishing.

Wilcox, J. (1963). A search for the multiple effects of grouping upon the growth and behavior of junior high school pupils. *Dissertation Abstracts, 24,* 205. (University Microfilms No. 63-4574)

Willcutt, R. E. (1967). Ability grouping by content subject areas in junior high school mathematics. *Dissertation Abstracts, 28,* 2152A. (University Microfilms No. 67-16,440)

Worlton, J. T. The effect of homogeneous classification on the scholastic achievement of bright pupils. *Elementary School Journal, 1928, 28,* 336-345.

Yates, A. (Ed.). (1966). *Grouping in education.* New York: Wiley.

Ziehl, D. C. (1962). An evaluation of an elementary school enriched instructional program. *Dissertation Abstracts, 24*, 2743. (University Microfilms No. 62-04644)

Zweibelson, I., Bahnmuller, M., & Lyman, L. (1965). Team teaching and flexible grouping in the junior high school social studies. *The Journal of Experimental Education, 34,* 20-32.

Research-Based Decision Making Series

The National Research Center on the Gifted and Talented
The University of Connecticut
362 Fairfield Road, U-7
Storrs, CT 06269-2007

Editor
E. Jean Gubbins

Production Assistants
Dawn Guenther
Renay Midler
Jonathan A. Plucker
Del Siegle
Siamak Vahidi

Series Reviewers

Susan Demirsky Allan
Francis X. Archambault
John Borkowski
James Borland
Carolyn M. Callahan
Pamela Clinkenbeard
Nicholas Colangelo
Gary Confessore
Bonnie Cramond
James Cross
Gary Davis
Marcia Delcourt
John Feldhusen
David Fetterman
William Foster
David Irvine
David Kenny
Joe Khatena
Jann Leppien
Wilma Lund
Marian Matthews
Stuart Omdal
A. Harry Passow
Jonathan A. Plucker
Brian D. Reid
Sally M. Reis
Joseph S. Renzulli
Del Siegle
Virginia Simmons
Robert J. Sternberg
Kazuko Tanaka
James Undercofler
Karen L.Westberg

Also of Interest

The Relationship of Grouping Practices to the Education of the Gifted and Talented Learner
by Karen B. Rogers

Cooperative Learning and the Academically Talented Student
by Ann Robinson

Self-Concept and the Gifted Child
by Robert D. Hoge & Joseph S. Renzulli

NRC G/T

The National Research Center on the Gifted and Talented Research Teams

The University of Connecticut

Dr. Francis X. Archambault, Associate Director
The University of Connecticut
School of Education, U-64
Storrs, CT 06269-2007
203-486-4031

Dr. Alexinia Y. Baldwin
Dr. Scott W. Brown
Dr. Deborah E. Burns
Dr. David A. Kenny
Dr. Jonna Kulikowich
Dr. Sally M. Reis
Dr. Karen L. Westberg
Dr. Michael F. Young

The University of Georgia

Dr. Mary M. Frasier, Associate Director
The University of Georgia
Department of Educational Psychology
323 Aderhold Hall
Athens, GA 30602-7146
404-542-5106

Dr. Scott L. Hunsaker

The University of Virginia

Dr. Carolyn M. Callahan, Associate Director
Curry School of Education
The University of Virginia
405 Emmet Street
Charlottesville, VA 22903
804-982-2849

Dr. Michael S. Caldwell
Dr. Robert W. Covert
Dr. Marcia A. B. Delcourt
Dr. Mary Catherine Ellwein
Dr. Bruce Gansneder
Dr. Brenda H. Loyd
Dr. Donald Ball

Yale University

Dr. Robert J. Sternberg, Associate Director
Yale University
Psychology Department
Box 11-A, Yale Station
New Haven, CT 06520-7447
203-432-4633

Dr. Pamela Clinkenbeard