

## Grouping and the Gifted

# Ability-Grouping Research Reviews: What Do They Say about Grouping and the Gifted?

If educators are to make informed decisions based on the findings about ability grouping, they must study the *original* research and be sure that the questions they are asking are the same ones posed by the researchers.

The questions of whether, when, and how to group students according to academic ability represent some of the most difficult and frustrating challenges facing educators today. Seeking to help answer these questions, researchers have applied new techniques of research review to this subject. Two prominent sets of reviews—the meta-analyses of James Kulik and Chen-Lin Kulik of the University of Michigan (1982, 1984b) and the best-evidence syntheses of Robert Slavin of Johns Hopkins University (1986, 1990)—attempt to synthesize this information. These reviews, their techniques, and their findings are important to educators who need to make decisions about grouping that are based on accurate knowledge of its effects. This article provides both a synthesis and a critique of these research reviews of ability grouping with the aim of clarifying for practitioners how these synthetic techniques affect

the results; what research questions are being asked and answered, and what is and isn't established by the research.

## Understanding the Methodology

Both the meta-analytic and best-evidence techniques of research review treat all included studies as equally valid. Although the reviewers set criteria for omitting clearly inadequate studies, they give all other studies the same weight, without regard for their relative quality. The best-evidence synthesis is more selective in its criteria, but then becomes vulnerable to the charge of hand-picking the evidence. (For a description of these two methods of research review and the more traditional narrative review, see the sidebar on p. 63.)

A methodological problem that applies primarily to the gifted (the top 3–7 percent) and to a lesser degree to high-ability students (the top 33 per-

cent) is the use of standardized test scores. On most studies included in the meta-analyses, these are the main measure of achievement. The scores of gifted students usually approach the ceiling on standardized achievement tests, making it very difficult to show significant academic improvement on their part. The ceiling effect of standardized tests is also a factor—although to a lesser degree—in evaluating the improvement of high-ability students. Certainly, at the minimum, the degree of academic improvement in the studies would be much greater if it weren't masked by the ceiling effect of standardized testing.

This problem stemming from the inclusion of high-ability students may affect all the major studies. However, I have had difficulty obtaining exact data on the percentage of studies included in the analyses that use standardized test scores. James Kulik (personal communication) reports that the ma-

majority of studies in his meta-analyses used such data. In his study, Slavin (1986) reported (personal communication) that almost all studies where effect size was computed used standardized data (raw scores, grade equivalents, or standard scores). In both the meta-analyses and the best-evidence synthesis, some forms of grouping were found to improve the academic performance of gifted children, and it is likely that the real benefits were greater than could be shown by the method of measurement.

In a more recent synthesis of grouping in secondary schools, Slavin (1990) raises an additional problem concerning the use of standardized testing as a measurement of the effects of grouping on student achievement. Discussing the lack of positive evidence for grouping in his study, Slavin says, "One possibility is that the standardized tests used in virtually all the studies discussed in this review are too insensitive to pick up effects of grouping." Insensitivity of the tests is indeed one possibility. Another is the criticism commonly raised by teachers, particularly at the secondary level, that the tests don't evaluate what they are teaching. One possible check on this difficulty is to compare student progress in ability-grouped vs. heterogeneous classes using teacher-made tests. These are less commonly used in research because they are not comparable across teachers and subject areas. In fact, in both Slavin's elementary synthesis (1986) and secondary synthesis (1990), one of the criteria for inclusion of a research study was that "teacher-made tests, used in a very small number of studies, were accepted only if there was evidence that they were designed to assess objectives taught in all classes" (Slavin 1990). Clearly, if ability grouping is being used effectively, the objectives should vary among the different classes. Therefore, testing for the same (probably minimal) objectives will not permit any benefits of ability grouping in average- or high-ability classes to be demonstrated. A similar problem, related to differentiating instruction appropriately for the students being

taught, arises again when we examine the research questions being asked.

### **Examining the Research Questions**

The most serious difficulty with Kulik and Kulik's meta-analytic reviews and Slavin's best-evidence syntheses on grouping appears when we delve into the studies that actually make up these syntheses. The research questions actually being asked may prove very surprising to educators who have been reading general accounts of the analyses.

One question *not* asked in the Slavin research was whether programs designed to provide differentiated education for gifted or special education students were effective. Those programs were systematically omitted from Slavin's synthesis on the basis that they "involve many other changes in curriculum, class size, resources, and goals that make them fundamentally different from comprehensive grouping plans" (Slavin 1986). It is ironic that some school systems are using the Slavin best-evidence synthesis to make decisions about gifted and special education programs when

**It is ironic that some school systems are using the Slavin best-evidence synthesis to make decisions about gifted and special education programs when such an application clearly is inappropriate.**

such an application clearly is inappropriate. Slavin (1988) addressed such programs in a later narrative review in which he argued that the research on them was biased and the programs were ineffective. However, this subject was not researched in the systematic fashion of the best-evidence synthesis, and, logically, that synthesis cannot provide guidance on it.

Kulik and Kulik did address the effectiveness of gifted programs in their meta-analyses, including such programs when their other methodological criteria were met. Their results show clear positive gains for students in gifted programs, which they attribute to the specialized curriculum and materials used and to the training afforded teachers in such programs.<sup>1</sup>

The importance of the research question being asked arises again when we examine Slavin's (1986) review of regrouping in the elementary school for reading and/or mathematics. Five of seven studies in the best-evidence synthesis found that students learned more in regrouped than in heterogeneous classes, while two found negative results. However, in at least one of the studies in which students in regrouped classes failed to outperform those in heterogeneous classes (Davis and Tracy 1963), no attempt was made to provide differentiated materials to the regrouped classes. Use of the same materials for all groups also occurred in a different study, included in both Slavin's and Kulik and Kulik's analyses, where students were regrouped for reading (Moses 1966). Despite this inadequacy of educational design, Moses found weak positive evidence for regrouping.

A study by Koontz (1961), the other study with negative results noted in Slavin's synthesis, involved regrouping for three subjects (math, language, and reading) and, therefore, had as much similarity to departmentalization models as to limited regrouping. Students changed classes three to four times a day. Most significantly, in the regrouping, language arts and reading each became separate classes, a very questionable educational practice. In contrast, a study by Provus (1960) in a suburban district showed clear and



**The most destructive aspect of the controversy over ability grouping is the misrepresentations of the findings, particularly those of Slavin's best-evidence synthesis, in the popular media.**

sometimes dramatic gains for students who were both regrouped for mathematics and provided with ability-appropriate materials. There were cases of 4th graders who finished the year working on an 8th grade level. Importantly, however, the gains were not limited to high-ability students. There were also clear, if less spectacular, benefits for both average- and low-ability students.

It is difficult to imagine any rational disagreement that could stem from these results. It is hardly reasonable to suggest that students should be ability grouped without the use of appropriate curriculum and materials. Grouping while using the same materials and curriculum for all groups of students is not supported by any segment of the education profession. But it appears that some researchers are attempting to ask the "pure" research question of whether grouping as a single isolated factor has any effect on student achievement. The answer, not surprisingly, is mixed, although generally positive. However, this is *not* the question that educators and parents are asking. They want to know whether grouping, with appropriately differentiated instruction, has any effect on student achievement. When that question is addressed, the results provide a stronger positive answer in both math and reading for all groups of students.

**Interpreting the Findings**

The most destructive aspect of the controversy over ability grouping is the misrepresentations of the findings, particularly those of Slavin's best-evidence synthesis (Slavin 1986), in the popular media. Headlines such as "Is Your Child Being Tracked for Failure?" (*Better Homes and Gardens*), "The Label That Sticks" (*U.S. News and World Report*), and, the most sensational of all, "Tracked to Fail" (*Psychology Today*) distort the research findings and undermine serious discussion of an important issue. The *Psychology Today* article begins with a ridiculous comparison to the categorization of alphas, betas, and gammas in *Brave New World*! There has been too little reaction from the educational community to bring the discussion back to a substantive level. The publications cited above, as well as some general education publications, fail to take note of Slavin's very important and worthwhile distinction between types of grouping. They also paint his research as having determined that grouping is academically harmful, which is not the case. The meta-analyses of Kulik and Kulik are less frequently misinterpreted by the general media, perhaps because they are rarely cited.

In examining the actual conclusions in these research syntheses, it is essential to examine them according to type of grouping rather than as one amorphous whole. When grouping is separated into within-class, comprehensive, and between-class grouping patterns, the results become more specific and useful.

*Within-class ability grouping* can be accomplished in several ways and can use a variety of educational techniques. After considering programs in which students in a grade level were assigned to different groups within heterogeneous classrooms, Slavin and Karweit (1984) concluded that such grouping clearly benefits students. Kulik and Kulik (1989) separated the within-class grouping studies into those designed for all students and those designed specifically for academically talented students. The programs designed for all students showed a positive, but small effect on student achievement. This effect was

similar for high-, average-, and low-ability groups. The within-class groupings for academically talented students were found to have substantial positive academic effects.

In examining techniques used in within-class differentiation of instruction, both Slavin and Kulik and Kulik have published reviews of mastery testing, and Slavin has reviewed cooperative learning. In the area of mastery testing, Slavin (1987) finds little methodologically adequate research support for it. Kulik and Kulik (1987) find that it generally has positive effects on student learning, although those effects were more pronounced for the less able students. However, it also increased the amount of time needed for instruction. On the average, mastery testing groups require 26 percent more instructional time than conventionally taught groups. Cooperative learning was not included in the Kulik and Kulik research, but Slavin is generally supportive of the practice if groups are rewarded on the basis of the individual learning of all members.

**When grouping is separated into within-class, comprehensive, and between-class grouping patterns, the research results become more specific and useful.**

The practice of *comprehensive full-day grouping* of pupils into different classrooms on the basis of general ability or IQ is not supported by Slavin's best-evidence synthesis. However, it is vital to note that he did not



find evidence of academic harm to students in this form of grouping—only lack of academic gain. This lack of academic gain shown among high-ability students in full-day grouping possibly is attributable to the ceiling effect of standardized testing. It also is useful to recall that gifted and special education programs were omitted from this aspect of the best-evidence synthesis, although Slavin has stated his opposition to them in other contexts (with the exception of acceleration programs, which he states may benefit gifted students). In contrast, Kulik (1985) found that students grouped in classes according to general academic ability slightly outperformed non-grouped students. The strongest positive effect size was for students in high-ability classes (0.12) with weaker effects for students in middle-level classes (0.04) and no effect for those in low-ability classes. In a separate analysis of gifted and talented programs, Kulik and Kulik (1989) found that students performed significantly better than they did in heterogeneous classes.

The practice of departmentalization was not addressed by Kulik and Kulik, and Slavin indicated that the small amount of existing research recommends against departmentalization in upper elementary and middle grades.

The final topic of direct contrast between the two reviews is that of *regrouping for specific subject areas*. This includes Joplin and non-graded plans as well as the more traditional regrouping, usually for math and language arts. Slavin (1986) concludes that such an approach can be instructionally effective, particularly when:

- it is done for only one or two subjects—students remain in heterogeneous classes for most of the day;
- it greatly reduces student heterogeneity in a specific skill;
- group assignments are frequently reassessed;
- teachers vary the level and pace of instruction according to student needs.

Slavin's conclusions raise an interesting point of conflict with Kulik and Kulik's research (1989). While they also found a positive effect on achieve-

### Methods of Reviewing Ability Grouping Research

Three main techniques have been used to review research in the area of ability grouping: narrative review, meta-analysis, and best-evidence synthesis. Narrative review is the "traditional" method in which the reviewer surveys and comments in detail upon individual studies in the literature. While narrative review permits a great deal of evaluative commentary on the studies it includes, reviewers have always struggled with the difficulty of comparing studies with different results and different standards of measurement. Meta-analysis and best-evidence synthesis, the methods used in the two sets of reviews that form the focus of this article, were developed in order to make the results more replicable and quantifiable than the narrative technique permits.

The meta-analytic technique (used by James Kulik and Chen-Lin Kulik) requires the reviewer to locate studies of an issue through objective and replicable searches, code the studies for salient features, and describe study outcomes on a common scale. Kulik and Kulik itemized additional qualifications for the use of a study in their meta-analysis. In order for a study to be included in their review, the results had to be reported in quantitative form; the results had to be available from a conventionally instructed control group as well as from the one receiving the experimental treatment; the control group had to be similar to the experimental group in aptitude; and, very important, the studies had to take place in actual classrooms, not labs.

The best-evidence synthesis technique (used by Robert Slavin) is a combination of meta-analysis and narrative review. It has many characteristics in common with meta-analysis, including the computation of effect size<sup>1</sup> and the clear specification of inclusion criteria. There are, however, several crucial differences. One important difference is that studies were included whose effect size could not be computed. Such studies are characterized in the data analyses as positive, negative, or zero rather than excluded. In addition, individual studies and methodological and substantive issues are disclosed in the detail typical of narrative reviews. Finally, the Slavin review included studies that used calculations that Kulik and Kulik considered mathematically inappropriate for their meta-analytic techniques. □

<sup>1</sup>Effect size is computed as the difference between the mean scores of experimental and control groups, divided by the standard deviation of the control group. It provides a common scale that standardizes the various measurements used in different studies (Kulik and Kulik 1989 and Slavin 1989, 1990).

Kulik, J. A., and C.-L. Kulik. (1989). "Effects of Ability Grouping on Student Achievement." *Equity and Excellence* 23, 1-2: 22-30.

Slavin, R. E. (1989). "Grouping for Instruction." *Equity and Excellence* 23, 1-2: 31-36.

Slavin, R. E. (1990). "Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis." *Review of Educational Research* 60, 3: 471-499.

—Susan Demirsky Allan

ment for such regrouping approaches, they further observed that this effect existed even when the regrouping was not limited to only one or two subjects, did not substantially reduce student heterogeneity, and when group assignments were not frequently reassessed. In other words, Kulik and Kulik (1989) did not find evidence to support Slavin's conclusion that grouping programs are most effective when the specific criteria described above are met.

Finally, unlike Slavin, Kulik and Kulik (1982) and Kulik (1985) address the issues of attitude and self-concept.

Their findings in these areas show that grouping has minor effects and is generally positive. They found that students who were ability grouped for a specific subject had a better attitude toward that subject but that grouping did not change attitudes about school in general.

With regard to student self-esteem, Kulik and Kulik's research requires serious consideration. A major criticism of ability grouping is that it will lower the self-esteem of students in low-ability groups. Kulik and Kulik determined that, in general, effects of grouping on self-esteem were very small and



somewhat dependent upon program type. Programs with high-average-low groups have a small overall effect on self-esteem, but effects tend to be slightly positive for low-ability groups and slightly negative for high and average ones. Limited studies of remedial programs (Kulik 1985) provide evidence that instruction in homogeneous groups has positive effects on the self-esteem of slow learners. Programs designed for gifted students have trivial effects on self-esteem (Kulik 1985). Why are these results counter to the prevailing expectation? Kulik (personal communication) raises an interesting point on the relative importance of the effects of labeling versus the effects of daily classroom experience. He suggests that the labeling (by placement of a student into a low-medium-high group) may have some transitory impact on self-esteem but that impact may be quickly overshadowed by the effect of the comparison that the student makes between himself or herself and others each day in the classroom. Low-ability students may experience feelings of success and competency when in a classroom with others of like ability, and high-ability students may encounter greater competition for the first time. While the data cannot, in themselves, identify the cause of these findings, the results make it clear that we must reexamine the arguments about self-esteem in light of them.

### **Other Issues to Consider**

Kulik and Kulik's meta-analyses and Slavin's best-evidence syntheses address a number of important issues about ability grouping for academic instruction. However, other concerns should be considered in making academic grouping decisions. Issues such as the impact of adult attitudes towards grouping, the role of gifted students as role models for other students, and the impact of grouping on student behavior and teacher expectations are all crucial.

Neither of the two studies discusses the importance of teacher and parent attitudes and approaches to grouping, even though educator experience suggests that a low-key, supportive approach by all adults concerned goes a

## **The thorniest issue concerning grouping and the gifted is whether the gifted are needed in the regular classroom to act as role models for other students.**

long way toward minimizing any emotional effects of grouping.

The thorniest issue concerning grouping and the gifted is whether the gifted are needed in the regular classroom to act as role models for other students and whether this "use" of gifted students is more important than their own educational needs. That students constantly make ability comparisons between themselves and others (Nicholls and Miller 1984) is sometimes used as the rationale for having gifted students serve as motivational models for others. While there is nothing inherently wrong with serving as a positive role model on occasion, it is morally questionable for adults to view any student's primary function as that of role model to others.

Further, the idea that lower ability students will look up to gifted students as role models is highly questionable. Children typically model their behavior after the behavior of other children of similar ability who are coping well with school. Children of low and average ability do not model themselves on fast learners (Schunk 1987). It appears that "watching someone of similar ability succeed at a task raises the observer's feelings of efficiency and motivates them to try the task" (Feldhusen 1989). Students gain most from watching someone of similar ability "cope" (that is, gradually improve their performance after some effort), rather than watching someone who has attained "mastery" (that is, can

demonstrate perfect performance from the outset). These data are compatible with Kulik and Kulik's explanation of their data on self-esteem discussed previously in this article.

A final point not considered in either of the major analyses is that teachers of high-ability classes may spend less time on discipline, spend more time interacting with students (particularly at student initiation), have students who spend more time-on-task, use better teaching techniques, and have higher expectations (Veldman and Sanford 1984). The implication is that the differences in teacher behavior may be a result of teacher bias or expectations, rather than a reaction to the behavior and needs of the students. It is questionable whether the same teacher, with the same expectations, would be able to use the same techniques with a lower ability class. However, the point is well taken that teachers need to examine whether they are "under-expecting" performance from all groups of students and thereby not providing them with the opportunity to rise to their potential.

### **Educators as Critical Consumers**

There is a great deal to be learned from the Slavin and the Kulik and Kulik analyses of ability grouping. The separation of the data into types of grouping (comprehensive, between-class, within-class, separate program, and acceleration) is particularly valuable because it has demonstrated that the effects of grouping vary according to type of plan. However, there also has been a great deal of misrepresentation and misinterpretation of the research. Educators need to be critical consumers. I believe the following statements are supported by research results and may reasonably be applied by educators when making decisions on ability grouping.

1. Gifted and high-ability children show positive academic effects from some forms of homogenous grouping. The strongest positive academic effects of grouping for gifted students result from either acceleration or classes that are specially designed for the gifted and use specially trained teachers and



differentiated curriculum and methods. In fact, all students, whether grouped or not, should be experiencing a differentiated curriculum that provides options geared to their learning styles and ability levels.

2. Average- and low-ability children may benefit academically from certain types of grouping, particularly elementary school regrouping for specific subject areas such as reading and mathematics, as well as from within-class grouping. These benefits may be small. These students show very little benefit from wholesale grouping by general ability.

3. The preponderance of evidence does not support the contention that children are academically harmed by grouping.

4. Students' attitudes toward specific subjects are improved by grouping in those subjects. However, grouping does not have any effect on their attitudes toward school.

5. It is unclear whether grouping has any effect on the self-esteem of students in the general school population. However, effects on self-esteem are small but positive for low-ability children and slightly negative for average- and high-ability children. There is limited evidence that remedial programs have a positive effect on the self-esteem of slow learners.

I support the plea of many in the educational field that educational decisions stand upon a firm research base. The original research, however, must itself be examined rather than relying on distillations or selective, possibly biased reports in the media. Further, the questions the researcher is asking must match the questions being asked by the practitioner. Then, our decisions about ability grouping will stand on a sound research base. □

<sup>1</sup>R. Slavin (personal communication) suggests a distinction between enrichment and acceleration programs for the gifted. This is not always an easy distinction to make. Acceleration is clear when a 7th grader takes Algebra I or French. But is it acceleration or enrichment when a gifted program class introduces more sophisticated literature or science concepts than those used in the regular curriculum? Such material may be characteristic of that usu-

ally offered to older children but does not advance them through the instructional continuum. Many studies evaluate programs that are not clearly identifiable as being either enrichment or acceleration. Although the Kuliks did not make the enrichment/acceleration distinction in their meta-analyses on grouping, a separate meta-analysis on accelerated instruction (Kulik and Kulik 1984a) showed very strong positive benefits for acceleration. The performance of accelerated students surpassed by nearly one grade level the performance of nonaccelerates of equivalent age and intelligence. In their grouping meta-analysis, the Kuliks added an additional 24 studies on gifted children (there is only one overlap with the acceleration meta-analysis), and they obtained the positive results cited above.

#### References

- Davis, O. L., and N. H. Tracy. (1963). "Arithmetic Achievement and Instructional Grouping." *Arithmetic Teacher* 10: 12-17.
- Feldhusen, J. P. (1989). "Synthesis of Research on Gifted Youth." *Educational Leadership* 46, 6: 6-11.
- "Is Your Child Being Tracked for Failure?" (October 1988). *Better Homes and Gardens*: 34-36.
- Koontz, W. F. (1961). "A Study of Achievement as a Function of Homogeneous Grouping." *Journal of Experimental Education* 30: 249-253.
- Kulik, C.-L. (1985). "Effects of Inter-Class Ability Grouping on Achievement and Self-Esteem." Paper presented at the annual convention of the American Psychological Association (93rd), Los Angeles, California.
- Kulik, C.-L., and J. A. Kulik. (1982). "Effects of Ability Grouping on Secondary School Students: A Meta-Analysis of Evaluation Findings." *American Educational Research Journal* 19: 415-428.
- Kulik, J. A., and C.-L. Kulik. (1984a). "Effects of Accelerated Instruction on Students." *Review of Educational Research* 54, 3: 409-425.
- Kulik, C.-L., and J. A. Kulik. (1984b). "Effects of Ability Grouping on Elementary School Pupils: A Meta-Analysis." Paper presented at the annual meeting of the American Psychological Association, Toronto (ERIC No. ED 255 329).
- Kulik, C.-L., and J. A. Kulik. (1987). "Mastery Testing and Student Learning: A Meta-Analysis." *Journal of Educational Technology Systems* 15, 3: 325-345.
- Kulik, J. A., and C.-L. Kulik. (1989). "Effects of Ability Grouping on Student Achievement." *Equity and Excellence* 23, 1-2: 22-30.
- 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).
- Nicholls, J., and A. T. Miller. (1984). "Development and Its Discontents: The Differentiation of the Concept of Ability." In *The Development of Achievement Motivation*, pp. 185-218, edited by J. Nicholls. Greenwich, Conn.: JAI Press.
- Provus, M. M. (1960). "Ability Grouping in Mathematics." *Elementary School Journal* 60: 391-398.
- Rachlin, J. (July 3, 1989). "The Label That Sticks." *U. S. News and World Report* 51-52.
- Schunk, D. H. (1987). "Peer Models and Children's Behavioral Change." *Review of Educational Research* 57, 2: 149-174.
- Slavin, R. E. (1986). *Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis*. (Rep. No. 1). Baltimore, Md.: Johns Hopkins University, Center for Research on Elementary and Middle Schools.
- Slavin, R. E. (1987). "Mastery Learning Reconsidered." *Review of Educational Research* 57, 2: 175-213.
- Slavin, R. E. (1988). "Synthesis of Research on Grouping in Elementary and Secondary Schools." *Educational Leadership* 46, 1: 67-77.
- Slavin, R. E. (1990). "Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis." *Review of Educational Research* 60, 3: 471-499.
- Slavin, R. E., and N. Karweit. (1984). "Within-Class Ability Grouping and Student Achievement." Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Tobias, S. (September 1989). "Tracked to Fail." *Psychology Today*: 54-60.
- Veldman, D. J., and J. P. Sanford. (1984). "The Influence of Class Ability Level on Student Achievement and Classroom Behavior." *American Educational Research Journal* 21, 3: 629-644.

*Author's note:* I conducted this review while employed by Falls Church Public Schools in Virginia and gratefully acknowledge their sponsorship and encouragement of the project.

**Susan Demirsky Allan** is Consultant for Gifted Education/Fine Arts, Dearborn Public Schools, Department of Instructional Services, 18700 Audette, Dearborn, MI 48124.

Copyright © 1991 by the Association for Supervision and Curriculum Development. All rights reserved.