Journal of Evidence-Based Medicine ISSN 1756-5391

REVIEW ARTICLE

Four decades of research on the effects of detracking reform: Where do we stand?—A systematic review of the evidence

Ning Rui

Center for Research and Evaluation in Social Policy, University of Pennsylvania, USA

Correspondence
Ning Rui, Center for Research and Evaluation in Social Policy, University of Pennsylvania, USA
Email: nrui@dolphin.upenn.edu

Received 10 January 2009; accepted for publication 1 July 2009.

doi: 10.1111/j.1756-5391.2009.01032.x

Abstract

Objective To review and synthesize evidence about academic and non-academic effects of detracking reform.

Methods Fifteen studies conducted from 1972 to 2006 were located and reviewed, including 4 experimental studies, 2 quasi-experimental studies, 7 observational studies, and 2 qualitative studies. Meta-analyses using fixed effects and random effects models were conducted for all and subsets of selected studies (by the academic ability of students and research design), followed by extensive discussion of individual studies.

Results Generally speaking, students in detracked groups performed slightly better academically than their equivalent-ability peers in tracked groups ($d = 0.087$, $k = 22$, $N = 15{,}577$, $p < 0.0001$), using a fixed effects model. A random effects model also indicated the overall positive effects of detracking ($d = 0.202$, $k = 22$, $N = 15{,}577$, $p < 0.01$). However, the effect sizes of individual studies are generally heterogeneous with $I^2(21) = 94.033$. Using a random effects model, the study shows that average or high ability students in detracked groups performed no differently than their equivalent-ability peers in tracked groups with a 95% confidence interval of (−0.047, 0.388). For low-achieving students, both the fixed effects model [$d = 0.113$, $k = 8$, $p < 0.0001$, 95% CI (0.056, 0.169)] and random effects model [$d = 0.283$, $k = 8$, $p < 0.005$, 95% CI (0.087, 0.479)] revealed positive effects of detracking on student achievement for the 8 low-ability subgroups in 6 studies. The evidence with respect to the non-academic impact of detracking is mixed.

Conclusion The findings suggest that the detracking reform had appreciable effects on low-ability student achievement and no effects on average and high-ability student achievement. Therefore, detracking should be encouraged, especially in schools where the lower-track classes have been traditionally assigned fewer resources.

Introduction

Definition of policy problem

The practice of school tracking — using their scholastic capabilities to group students in differentiated classrooms or academic programs—fueled a debate spanning nearly the entire twentieth century, which still continues (1). Tracking is practiced in 60 percent of elementary and 80 percent of secondary schools in the United States (2). With the advent of standards-based education reform, interest and concern about tracking has increased in recent years among the lay public, educational researchers, and policymakers. This concern has been expressed through state legislative initiatives to abolish tracking in California and Massachusetts (3); increased federal and foundation derived funding for research and service programs on inclusion, detracking, ability grouping, mainstreaming, and cooperative

learning[1] (4); and an emerging professional literature on related issues (5–8). Some meta-analyses of past research have reported inconsistent effects of tracking on academic achievement in elementary and secondary schools (6, 7, 9). Other meta-analyses found varying psychosocial and institutional effects of tracking and detracking (10, 11). This uncertainty underscores the need to reexamine the academic and non-academic outcomes of detracking versus tracking as reported in recent literature, in order to inform more effective policymaking regarding grouping practice in the K-12 school system.

Opinions on the tracking practice have always been divisive. Advocates of tracking maintain that teachers can gear lessons specifically to the needs of students when they are grouped by alleged abilities, arguing that higher-track students can make greater gains (2, 12, 13). Opponents to tracking argue that lower-track students are often exposed to meager education resources and less effective instruction than higher-track students, which may lead to socioeconomically and racially segregated classrooms (5, 14, 15). Given the perceived inequality associated with tracking, detracking strategies started being implemented during the 1990s in an attempt to close the achievement gap between lower-track (mostly poor and minority students) and upper-track students (16). Detracking is a means for all students to work together regardless of their ability levels. Proponents of detracking maintained that low-achieving students would benefit from this practice (14). Opponents to detracking claimed that low-achieving students would hold back the high-achieving students making instruction more challenging for teachers, as well as wearing down low achievers' self-confidence by confronting them with instructional material beyond their capabilities (3, 17–19).

Implementation of state detracking policies

Since the 1970s, U.S. policymakers and educators have shown great concern about inequitable distribution of educational resources and insisted that all students have access to a rigorous curriculum, which has led to a series of federal initiatives (eg Goals 2000, No Child Left Behind) during the Clinton and Bush administrations calling for states to establish rigorous standards for what all students should learn and be able to do. The publication of Jeannie Oakes's *Keeping Track*: *How Schools Structure Inequality* in 1985 indicated that tracking created social and racial differences in access to learning. Since then, a wide range of national and local educational organizations started recommending detracking in K-12 schools. Organizations such as the National Education Association, National Governors Association, and the National Council of Teachers of English were in the front line of eliminating tracking in the early 1990s (17, 20). In 1987, the officials of California mandated that the state's middle schools eliminate or reduce tracking. In 1993, Massachusetts followed suit for its middle schools. The policy received positive responses from many local districts and schools in these two states. It was found that urban and rural schools, schools with higher minority enrollments, schools with fewer high-achieving students, and schools with low parental influence on policy were more likely to embrace detracking (3).

Philosophical arguments on tracking vs. detracking

The equity-efficiency tradeoff of detracking reform has been discussed intensely among political scientists and philosophers. Findings from two sets of meta-analyses on the effects of tracking (6, 7, 9) show that tracking tends to benefit students with higher aptitude (excluding exceptionally talented individuals), although the benefit is small across all studies reviewed. Other studies show that abolishing tracking would have a large positive impact on achievement of students who used to be placed in the low tracks, but that this gain would come at the expense of students in upper tracks (21). This raises a fundamental dilemma for educational policymakers who are looking for the optimal strategies to improve overall school performance.

Two main guiding but contradictory philosophical arguments in regard to the issue are based on the Pareto principle in neoclassical economics and Rawls's social contract theory (22). The Pareto principle claims that 80% of the effects come from 20% of the causes for many events. No individual can be better off without someone else being made worse off under a Pareto efficient economic system (23). Therefore, it is commonly accepted among economists that Pareto efficiency is an important criterion for evaluating public policies given the assumption that resources in a capitalist society are not equitably distributed. From the Pareto perspective, any egalitarian-based education policy would be an unproductive policy because the high-achieving students' performance might be held back and hence the overall achievement might be affected adversely. On the other hand, John Rawls (1921–2002) developed the Justice as Fairness theory, which consists of two principles. One of the principles states that "social and economic inequalities are to be arranged so that. . . they are to be of the greatest benefit to the least-advantaged members of society" (22). This justifies the coercive use of state power to ensure equitable distribution of educational resources. Guided by this theory, detracking should be widely accepted if the least advantaged and low-performing students tend to realize substantial academic gain

[1] In this article, "tracking" and "homogeneous grouping" are used interchangeably; "detracking" and "heterogeneous grouping" are used interchangeably.

in detracked classes. Policymakers should target the progress of the gain made by the lowest performers when evaluating the relative effectiveness of the policy.

Tracking in a global context

Previous studies show that schools in other nations that surpassed their American counterparts in standardized scores were often tracked. For example, in New Zealand, classes are usually composed of students of mixed ages, and students advance in accordance with their developmental capabilities (12). Singapore is one of few countries where tracking starts in primary schooling. The new Ability-Driven Education paradigm recently adopted by the government proposed subject-based banding, which allows the 5th graders to take classes in Standard or Foundation subjects based on their proficiency in each subject. The education paradigms in Singapore, including the traditional primary school streaming, have rarely been contested by the general public because of a preexisting social covenant between the government and the public that the population is the country's most important resource and education is deemed a strategy of resource development. The objective of the education system is to maximize the development of every individual's aptitude, so that each could make the best contribution to the economy (24).

In both Japan and China, tracking usually starts when students enter high school, following the 9-year compulsory education (including primary and junior middle schooling). To most students –and parents–, the beginning of high school represents a key milestone in life, which will determine or restrict their access to higher education. The tracking in these two countries is characterized by high school stratification. In Japan, those junior middle school graduates who aim at going to colleges will go to standard high schools where teachers will prepare them for the college entrance examination. Those who would like to enter the workplace after high school usually choose to go to vocational schools. This is similar to the wholesale tracking of students into discrete vocational and academic tracks in the U.S. In modern Japan, family background was found to have stronger influence on students' access and choice to high school and postsecondary education than in the U.S. (25). Many high schools in China assign students to a fast-track class (sometimes called "experimental class" or "potential class") or a slow-track class (or "common class") in the first year based on each student's prior achievement or their performance on the entrance exam. During the second year (sometimes even earlier), students will be assigned to art or science cohorts for the purpose of differentiated instruction and preparation for the National College Entrance Examination, which includes different mixes of subjects based on students' interests and choices. While criticism has been raised about the inequity issue and unequal allocation of instruction associated with tracking, some argue that tracking, when implemented with careful control, may meet each individual's learning needs and maximize the development of each student's talent. It is important to differentiate absolute equity (without consideration of efficiency) and relative equity based on the acknowledgement of individual differences (26).

In other countries, tracking is not encouraged. For example, primary school students in Switzerland are not differentiated by ability level at all. Cooperative learning is strongly advocated in classrooms, and usually, a faster learning child is carefully paired with a slower learning one to prevent the development of a substantial achievement gap (27). A teacher will be assigned to the class during the first 3 years and years 4–6 respectively so that the teacher can detect the students with learning difficulties.

Based on the practice of student tracking on the global scene, some argue that it is not tracking itself that affects the American students' achievement, but rather how schools track and the extent to which differentiated instruction allows each student to learn at the optimal depth and pace, as well as providing motivation and excitement about learning (12). At the national level, the degree to which detracking is accepted and legitimized in a country is usually determined by nation-specific cultural expectations (28, 29) and public conceptions about the role of education in national economic development.

Methods

Research questions to be addressed

Because of the institutional, organizational, political, and social factors surrounding the detracking versus tracking debate and effective policy implementation, it is essential that research on the outcomes of detracking be reviewed and summarized. This review examines and synthesizes currently available research findings about the effects of detracking compared to tracking on student academic and non-academic outcomes. Such a review is needed to inform practitioners, policymakers, and researchers about the current state of evidence on this topic and about gaps in the knowledge base in need of further scientific investigation. To achieve the goal, this paper presents a comprehensive review of major studies that examined the effects of detracking or heterogeneous grouping on K-12 student since the early 1970s. The following three questions are addressed:

1. What are the academic outcomes of detracking for average or high-achieving students (the students originally placed in the upper tracks) in elementary and secondary schools?
2. What are the academic outcomes of detracking for low-achieving students (the students originally placed

in the low tracks) in elementary and secondary schools?
3. What are the non-academic outcomes (social, psychological, attitudinal, and behavioral) of detracking for all students in elementary and secondary schools?

In concluding this study, I summarize major findings from the review and provide research-based policy recommendations on student grouping for school administrators, teachers, and policymakers.

Literature search and selection criteria

The studies included in this review were located through an extensive search. Principal sources included university libraries, the Education Resources Information Center (ERIC), PsychInfo, ProQuest digital dissertations databases, and citations in the identified reviews and meta-analyses. Attempts were made to obtain a complete set of published (preferably from peer-reviewed journals) and unpublished studies (doctoral dissertations and conference research papers) that contain such as keywords as detracking, school integration, heterogeneous grouping, tracking, ability grouping, or homogeneous grouping. Any citation starting from the early 1970s that fit into these categories became part of the review pool. Preference was given to experimental and quasi-experimental studies that compared detracked groups (or heterogeneously grouped classes) with tracked groups (or ability-grouped classes), but some observational or pre-experimental studies were also selected so that the findings from experimental studies and non-experimental studies could be compared. In this review, an experimental study is defined as a study where each subject is randomly assigned to an intervention group or a control group before the implementation of the intervention. A quasi-experimental study shares many characteristics of a true experimental study, except for the lack of random assignment. An observational study is defined as a study that is based on observed data and draws inferences about the effect of an intervention using regression, matching of subjects, or propensity score analysis, where the assignment of subjects into a group is outside the control of the investigator.

The studies included in this review have to meet a set of a priori criteria with respect to the content and methodological adequacy. These include:

1. Only studies of comprehensive detracking that incorporated most or all students were included in the review. In other words, studies on programs for gifted children, special education programs, and remedial programs were excluded due to their lack of generalization.
2. All empirical research articles have to be available in English, but no such restrictions were placed on background, theoretical, and commentary literature. For example, one background article and one commentary article in Chinese were used in the literature review of tracking in a global context.
3. Each study must clearly describe a practice that can be identified as detracking or heterogeneous grouping.
4. Each quantitative study must have a sample of at least 30 students.
5. The detracking or tracking has to be in place for at least a semester in each study.

Procedures for research synthesis

Meta-analyses were conducted to combine the results of a set of quantitative studies that address the achievement effect of detracking using experimental or quasi-experimental design or regression-based inference. These meta-analyses provide a summary of previous research, using quantitative methods to compare outcomes across a range of studies (30). Separate meta-analyses were conducted for subsets of selected studies based on the academic ability of students and research design. Effect sizes for each subgroup in each study were calculated using *Cohen's d,* which is computed as the experimental mean minus the control mean with the result divided by a pooled standard deviation (31). A positive effect size indicates that students in detracked classes had better outcomes; a negative effect size indicates that students in tracked or ability-grouped classes had better outcomes. Traditional statistics such as t tests or F ratios are inappropriate for comparisons of various studies because the values of those statistics are partially a function of the sample size. A common impediment to meta-analyses in education research is that there are usually too few experimental studies where effect sizes are reported or can be derived. Additionally, in this specific case, many of the empirical studies on the benefits of detracking have design and methodological limitations. Such problems include different definitions of the intervention, imbalanced design, the lack of comparison groups, and selection bias. Given these limitations, this review applied the best-evidence synthesis (32), in addition to the meta-analyses. The best-evidence synthesis uses a systematic literature search, quantification of outcomes as effect sizes, and extensive discussion of individual studies that meet inclusion criteria. Best-evidence synthesis also requires an extensive description of key studies.

If effect sizes were not reported directly from a study, they were estimated using r, t, F, and p values or other well-established estimation methods (33). For studies that lacked means and standard deviations, reported no significant difference between the experimental and control groups, and did not indicate the direction of the effect, an estimated effect size of zero was used. All data were entered into the beta version of the Comprehensive Meta-Analysis Program (34), to estimate the effect sizes of each study, calculate the overall mean

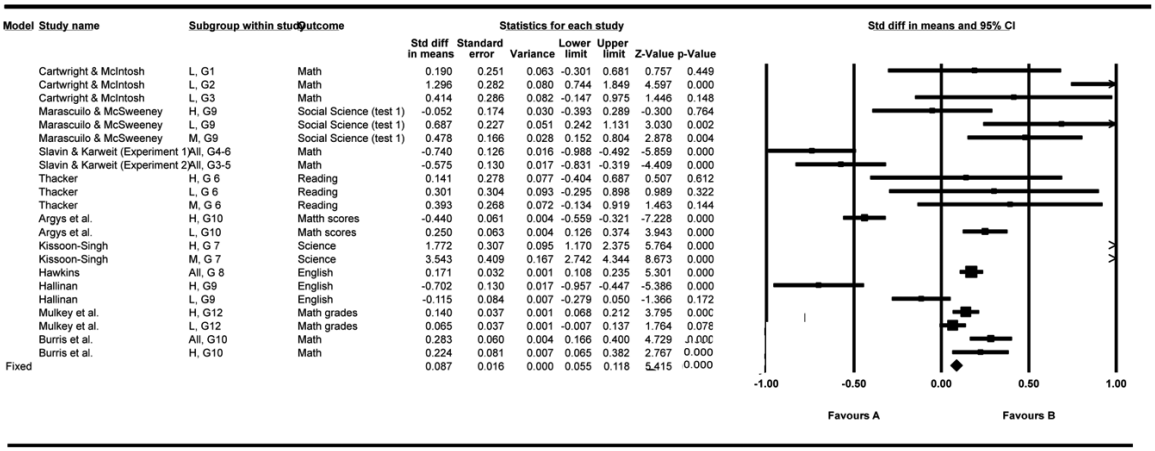

Figure 1 Summary of meta-analysis of studies on achievement effects of detracking (combined).

weighted effect sizes, and test whether the mean weighted effect size was derived from a homogeneous set of studies. The weighting factor was sample size, so that effect sizes from larger samples contributed more to the meta-analysis mean than did those from smaller samples.

In this review, a total of 15 single studies were reviewed, including 4 experimental studies, 2 quasi-experimental studies, 7 observational studies, and 2 qualitative studies. In the detracking literature, the homogeneous group is usually considered the control group; while in the tracking or ability grouping literature, the heterogeneous group is almost always considered the control group. Positive effect sizes are those that favor detracking in the former type of literature, whereas negative effect sizes are in favor of detracking in the latter type of literature.

Research findings

Results of meta-analyses

Of the 15 studies reviewed, 10 reported results on the effects of detracking or heterogeneous grouping on student achievement in at least one educational subject. The meta-analysis of the results for 22 subgroups (by student initial ability and grade level) in these 10 studies, conducted in 9 states or provinces in the United States and Canada, revealed that students in detracked groups performed slightly better academically than their equivalent-ability peers in tracked groups ($d = 0.087$, $k = 22$, $N = 15{,}577$, $p < 0.0001$), using a fixed effects model. A random effects model also indicated the overall positive effects of detracking ($d = 0.202$, $k = 22$, $N = 15{,}577$, $p < 0.01$). Random effects models try to account for the possibility that population parameters (d) vary from study to study, while fixed effects models assume a priori that the same effect size value underlies all studies in the meta-analysis (or standard deviation of $d = 0$). If there is very little variation between studies, then the I^2 for heterogeneity will be low and a fixed effects model would be appropriate (35). Figure 1 summarizes and compares the effect sizes associated with detracking effects for the 22 subgroups using a forest plot. Note that the standard differences in means larger than 0 favor the detracking group, suggesting that the intervention enhanced student achievement. Although the average effect size was significantly larger than zero [$t(21) = 5.42$, $p < .0001$], the effect sizes of individual studies were heterogeneous with $I^2(21) = 94.033\%$. This result suggests that factors other than sampling error influenced the sampling distribution.

Two possible correlates or moderators of the effect size are the study design and the initial abilities of the participating students. To have a better understanding of the differential effects of detracking on various subgroups of students, I did a meta-analysis separately for each of the following sets of studies: (1) all studies on average or high-achieving students, (2) all studies on low-achieving students, (3) experimental studies for average or high-achieving students, and (4) experimental studies for low-achieving students. The results of these four meta-analyses were summarized in Figures 2 through 5. Figure 2 revealed that average or high ability students in detracked groups performed slightly better academically than their equivalent-ability peers in tracked groups ($d = 0.075$, $k = 14$), using a fixed effects model. Although the average effect size was significantly larger than zero [$t(13) = 3.91$, $p < 0.0001$], the effect size was heterogeneous with $I^2(13) = 95.83\%$. Given the large heterogeneity ($I^2 > 95\%$) among the selected studies, the meta-analytic results may be unreliable. In fact, when a random effects model was applied, the overall effect size ($d = 0.170$) was no longer significant ($p = 0.125$), with a 95% confidence limits of (-0.047, 0.388), suggesting that the detracking reform had no appreciable effects on achievement, on average, for student of average and high ability.

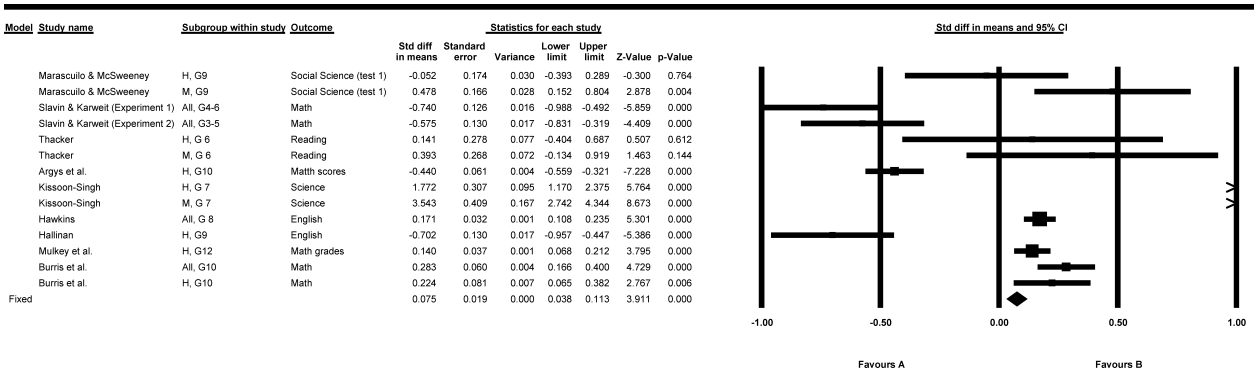

Figure 2 Summary of meta-analysis of studies on achievement effects of detracking for all or high-achieving students.

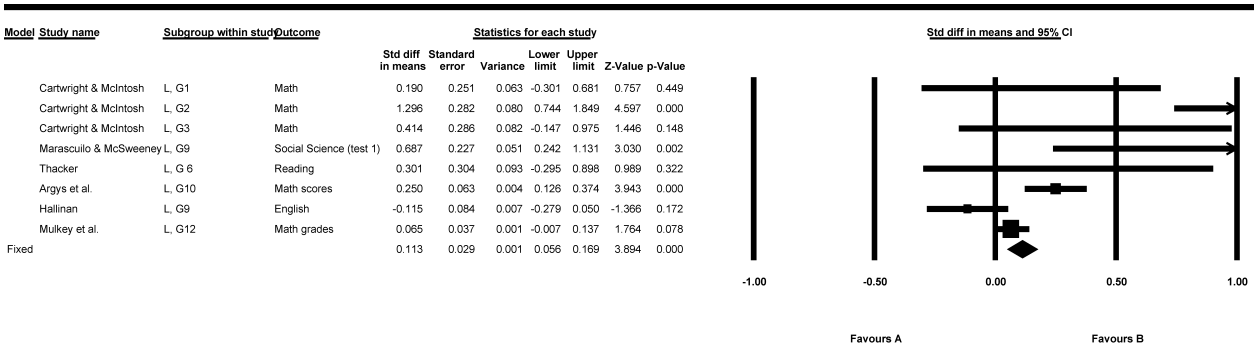

Figure 3 Summary of meta-analysis of studies on achievement effects of detracking for low-achieving students.

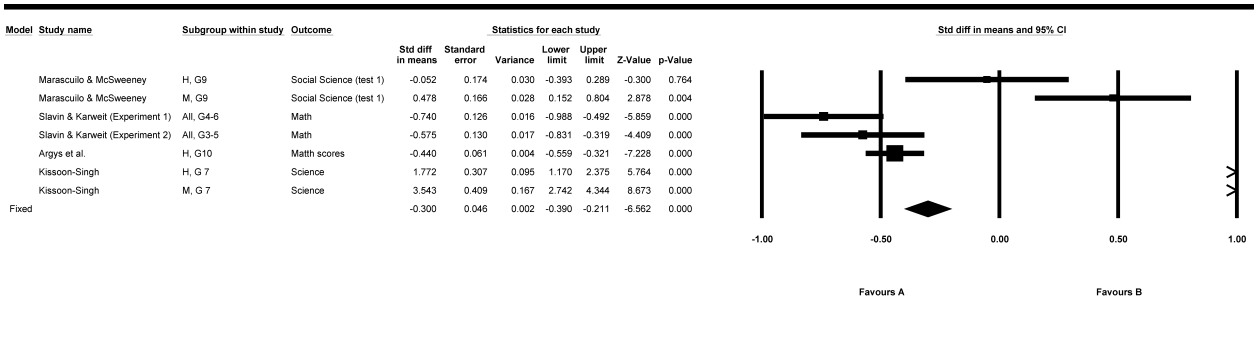

Figure 4 Summary of meta-analysis of experimental studies on achievement effects of detracking for all or high-achieving students.

On the other hand, for low-achieving students, both the fixed effects model [$d = 0.113$, $k = 8$, $p < 0.0001$, 95% CI (0.056, 0.169)] and random effects model [$d = 0.283$, $k = 8$, $p < 0.005$, 95% CI (0.087, 0.479)] revealed positive effects of detracking on student achievement for the 8 low-ability subgroups in 6 studies. As shown in Figure 3, the effect sizes of most studies are to the right of zero, even though there is considerable heterogeneity:

$I^2(7) = 82.20\%$.

Figure 4 summarizes the results of 7 subgroups with high or average ability students in 4 experimental studies, which reveals that average or high ability students in

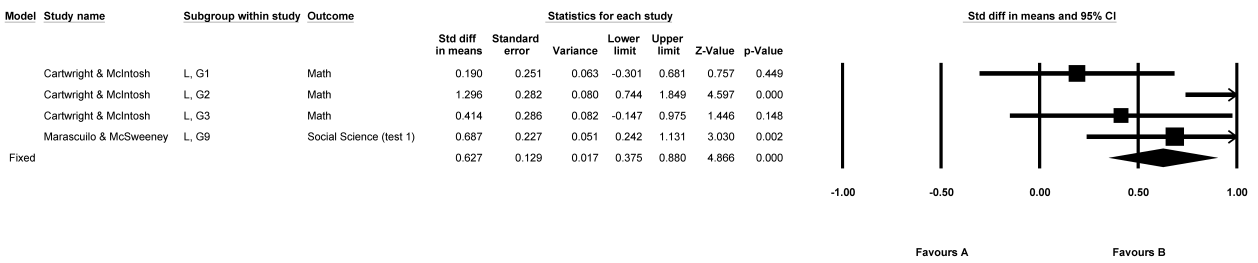

Figure 5 Summary of meta-analysis of experimental studies on achievement effects of detracking for low-achieving students.

detracked groups performed moderately worse than their equivalent-ability peers in tracked groups ($d = -0.300, k = 7, p < 0.0001$), using a fixed effects model. However, visual examination suggests that there is a high degree of heterogeneity of effect sizes reported by these 4 studies. When a random effects model was applied to the data, the average effect size turned positive, but was not statistically significant ($p = 0.125$). This result further confirmed that the detracking reform had no clear effects on average and high-ability students' achievement.

Figure 5 summarizes the results of 4 subgroups with low ability students in the 2 experimental studies, which reveals that low ability students in detracked groups enjoyed substantially higher academic achievement than their low achieving peers in tracked classes, as revealed by a fixed effects model [$d = 0.627, k = 4, p < 0.0001$, 95% CI (0.375, 0.880)] and a random effects model [$d = 0.640, k = 4, p < 0.005$, 95% CI (0.191, 1.088)]. This finding provides further confirmation of a positive academic effect of detracking for students in the lower track.

One of the reasons that there is such a high degree of heterogeneity of effect sizes for studies on high or average ability students is that the study by Kissoon-Singh (36) reported extremely large effect sizes (1.772 for high ability and 3.543 for average ability). The intervention in this study incorporates specially designed instructional methods in a computer-based setting, which may influence its results. To obtain a more objective overall effect size for experimental studies on high or average ability students, I conducted a sensitivity analysis by excluding the Kissoon-Singh study. The results (Fig. 6) indicate that there were no effects of detracking in either direction for high and average ability students [$d = -0.005, p = 0.837$, 95% CI (-0.053, 0.043)]. However, considerable heterogeneity still exists with $I^2(8) = 94.22\%$.

Narrative findings from reviewed studies

Studies on academic outcomes of detracking for average or high-achieving students

Experimental studies. Marascuilo and McSweeney (37) conducted a randomized trial that examined the effects of detracking practice on social science achievement and individual attitudes toward self, school, and classes for high-, average-, and low-ability student groups, respectively. Four experimental (detracked) classes of 28 junior high school students were created by proportional allocation and random

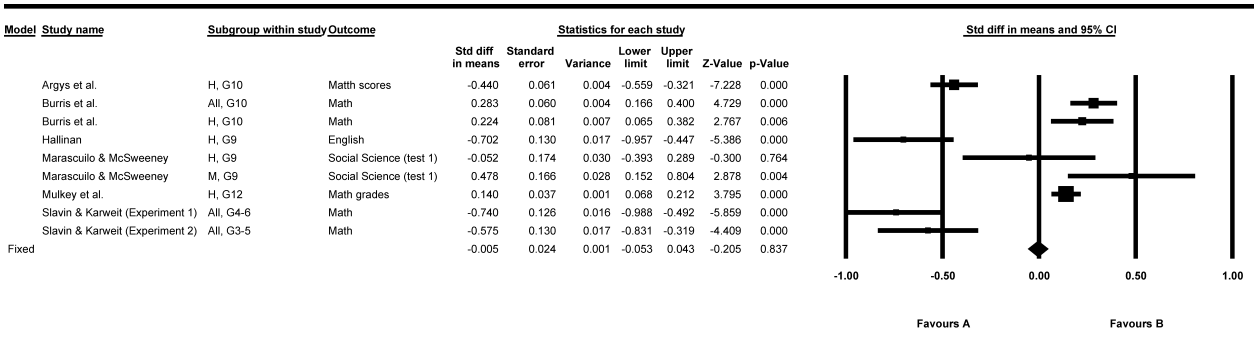

Figure 6 Sensitivity analysis for all or high-achieving students (excluding Kissoon-Singh study).

assignment of students from high-, middle-, and low-tracks so that the class composition would resemble that of all students in the 8th grade. The remaining students were randomly distributed to their regular tracked classes, including 6 high-track classes, 7 medium-track classes, and 3 low-track classes. For the academic outcomes, standardized tests such as the Cooperative Social Studies Test (CSST), a test on the U.S. Constitution, and a teacher-made test were used. For non-academic outcomes, an attitude questionnaire aimed at assessing attitudes toward self, school, and social studies classes were administered during the second semester. Results from regular eighth-grade testing showed no initial differences between the students assigned to detracked classes and those in tracked classes for all three ability groups. At the end of the first year, the authors found no statistically significant differences in the performance of high- and medium-achieving students in detracked classes and their peers in tracked classes on CSST (H: $F = 1.7$, $p > 0.05$; M: $F = 0.9$, $p > 0.05$) and the Constitution test (H: $F = 1.0$, $p > 0.05$; M: $F = 1.5$, $p > 0.05$). At the end of the second year (when the students were in the 9th grade), there were still no significant differences between the performance of detracked and tracked high-achieving students on the CSST and a teacher-made test ($F = 1.1$, $p > 0.05$). These results suggest that detracking did not affect the performance of high-achieving students. However, for medium-achieving students, the students in detracked classes significantly outscored their counterparts in tracked classes at the end of the second year on both CSST ($F = 2.9$, $p < 0.05$) and the teacher-made test ($F = 7.4$, $p < 0.05$).

Slavin and Karweit (38) conducted two randomized field experiments to investigate the achievement effects of three instruction methods: team-assisted individualized instruction, within-class ability grouping, and a whole-class, untracked instruction. It should be noted that the grouping practice referred to in this study took place within classes, which is different from the usual tracking into separate classes. The first experiment was conducted in 15 grade 4–5 classrooms in an urban district where heterogeneous class assignments were mandated as part of a desegregation plan. The second took place in a relatively homogeneous rural school setting. For the within-class tracking classes, teachers were instructed to differentiate pace and materials for a high-ability group and a low-ability group within a class based on initial test scores for the students. Results from the first experiment showed that there were significant cross-group differences in computational skills, [$F(2,12) = 5.11$, $p < 0.05$] but not in concepts and applications achievement. Post-hoc analyses revealed that students in within-class grouping and individualized instruction classes performed significantly better than those in whole-group classes by 74% of a standard deviation, and there was no difference between within-class grouping and individualized instruction in computations achievement. For the second experiment, the overall nested analysis of variance (students nested in classes) was marginally significant for computations [$F(3,18) = 2.71$, $p < 0.076$]. Similar to the findings from the first experiment, within-class grouping and individualized instruction did not differ in effects on computational ability, but both were superior to uniform class instruction ($d = 0.58$, SE $= 0.13$). Kissoon-Singh (36) conducted an experimental study to investigate the effects of studying in heterogeneous-ability groupings compared to homogeneous groupings in computer-based settings on achievement in mathematics and science and the self-efficacy of high and average ability students. The author started the study by classifying 130 seventh-grade students into high and average ability groups based on their Canadian Cognitive Abilities Test (CCAT) scores. The high-ability and average-ability students were randomly assigned to three groups: homogeneous high ability, homogeneous average ability, and heterogeneous ability. The study examined students' scores using four analysis groups: high ability students in homogeneous classes ($n = 44$), high ability students in heterogeneous classes ($n = 21$), average ability students in homogeneous classes ($n = 44$), and average ability students in heterogeneous classes ($n = 21$). The multivariate analysis of variance of pretest scores indicated that neither homogeneous nor heterogeneous groups had a distinct advantage over the other in performance on any of the dependent measures for high- and average-ability students respectively ($F = 0.28$, $p > 0.05$). Both an independent t-test of the posttest scores and analysis of covariance with pretest as covariate showed that the academic outcomes were significantly better in heterogeneous versus homogeneous groups for average ability [$d = 3.54$, 95% CI 2.71,4.30] and high ability students [$d = 1.77$, 95% CI 1.15,2.35]. Although this study found highly positive achievement effects of heterogeneous grouping, the study was conducted in a cooperative computer-based instruction (CBI) setting. Therefore the treatment variable in this study was not only placing students with various abilities in the same classroom, but also use of cooperative learning with computer-based instruction. The relative effects of the placement, instructional, and technological components of the treatment were not disentangled in this study. However, the highly positive effects observed suggest that the instructional and technological aspects of the heterogeneous-ability classes were important conditions for optimizing the use of a heterogeneous learning environment. These conditions should be integrated in a class to serve a facilitative role in amplifying student achievement and the efficacy of detracking.

Quasi-experimental studies. Thacker (39) conducted a Nonequivalent Control Group Design study that examined the effects of administrator implemented homogeneous and heterogeneous instructional grouping techniques on sixth grade reading comprehension achievement in the United

States. Students from two schools in northern Indiana were selected, with one school fully utilizing homogeneously grouping ($n = 113$) and the other relying on heterogeneous grouping practices ($n = 59$). Students who were grouped homogeneously received instruction appropriate to their needs. The reading score mean for high-ability students exposed to heterogeneous instruction (398.17) was about 10 points higher than that for high-ability students exposed to homogeneous instruction (388.37). However, analysis of variance (ANOVA) showed that this difference was not statistically significant. Similarly, no grouping effect was detected for average-ability students, although the reading scores for average-ability students exposed to heterogeneous instruction (357.19) was about 13 points higher than that for average-ability students exposed to homogeneous instruction (344.77).

Burries, Heubert, and Levin (40) conducted a longitudinal quasi-experimental study to examine the effects of offering an accelerated mathematics curriculum in heterogeneously grouped middle school classes. Two outcome variables were defined: whether a student subsequently completed advanced high school mathematics courses, and their mathematics achievements. Burries et al. applied an interrupted time series design in that the first three student cohorts (who entered high school in 1995 to 1997) were traditionally tracked, while the last three cohorts (who entered in 1998 to 2000) were the first three in which all students were accelerated and heterogeneously grouped through grades 6–8. Descriptive statistics showed that the percentage of initially high-, average-, and low-achieving students taking advanced mathematics courses all increased after the detracking reform. The results of multiple logistic regression analysis with initial achievement as covariate showed that members of universal acceleration (detracked) cohorts were still more than 2.6 times as likely as members of the pre-detracked cohorts to take and pass Sequential Math III after controlling for initial achievement and underrepresented status. For the post-treatment achievement, it was found that universal acceleration was associated with increased achievement among all students (1.51) and also high achievers (beta = 0.22) using linear regression with initial achievement as covariate. The main drawback of this study is that the causal inference of detracking effects was based on the cross-sectional data of student cohorts enrolled in different years. The author had to make a strong assumption that the student demographic characteristics and ability levels remained stable across the six years of the study. In addition, the effect of detracking in this study was confounded with the accelerated mathematics curriculum. It is unclear which factor contributed more to the student academic outcomes.

Observational studies. Hawkins (41) conducted a single group pretest-posttest study to examine the change in achievement grades of 970 eighth-grade students, who had experienced ability based grouping in the seventh grade and were then assigned to mixed-ability classrooms in Philadelphia, United States. Student grades in English, reading, social studies, mathematics, and science over three marking periods in the seventh and eighth grades were analyzed using repeated measures analysis of variance. Few significant differences appeared between achievement grades in the seventh (when students were ability-grouped) versus the eighth grade (when students were detracked). The only one statistically significant difference occurred in the third marking period English grades where the students achieved at a higher level in the eighth grade when compared to the seventh grade ($F = 1.74$, $p = 0.002$). This suggested that detracking had little short-term effect on academic grades. However, due to the lack of a comparison group, this study was susceptible to many factors that could threaten the validity of its findings. Furthermore, since the author did not analyze the disaggregated data by the prior achievement level of the students, it was unclear whether high-, average- and low-achieving students experienced differential growth by the detracking practice.

Mulkey, Catsambis, Steelman, and Crain (42) conducted a longitudinal analysis on whether heterogeneous or homogeneous grouping produced better academic and non-academic outcomes for middle school students. The study followed 5,895 students from grade 8 through 12 using data from the National Education Longitudinal Survey (NELS: 88). The main advantage of the longitudinal panel data is that it represents a much broader spectrum of students than individual-site studies. However, the results should be interpreted with caution since NELS: 88 used multi-stage sampling, and students were not randomly assigned to tracked and untracked conditions. To adjust for this limitation, Mulkey et al. used propensity analysis to achieve a balance between intervention and control groups (43). The results showed that middle school 8th grade tracking was positively associated with the 12th grade test scores in mathematics for students ($d = 0.33$) in all tracks. However, the authors also found that the 8th grade tracking was negatively associated with high-achieving students' grades in the 12th grade, as shown by a tracking d of -0.21 for high-achieving girls and -0.07 for high-achieving boys. Middle school tracking experience had no direct relationship with the number of mathematics courses taken in high school ($d = 0$ for all). This study suffers inevitably from a number of threats to internal validity, such as history (other events that happened between the 8th grade and 12th grade may explain the outcomes) and maturation (the outcome may simply be a function of the passage of time). There are also factors jeopardizing the external validity, such as multiple treatment interference (any other grouping experience between the 8th grade and 12th grade may have lingering effects).

Another study that used the NELS: 88 data was conducted by Argys, Rees, and Brewer (21). In this study, the authors

focused on the effect of tenth grade tracking on grades at the end of the year. Like the Mulkey et al. study, the authors used response data to the teacher questionnaire to determine student' class tracks. The study sample consisted of 3,405 students who attended public schools in both 8th and 10th grades. A multinomial logit model was applied to estimate track placement outcome (the upper-track, average-track, low-track, and heterogeneous class, respectively) with such explanatory variables as a student's prior achievement (at the 8th grade), gender, race/ethnicity, socioeconomics background, and region. The results indicated that moving from upper tracks to a heterogeneous class would lead to an 8.4% decrease in mathematics scores and that moving all students to an untracked class would lead to an average decrease in scores by 2%. The computed Cohen's *d* using pooled standard deviation for detracking effects on high- and average-ability students are −0.44 and −0.16 respectively, suggesting a small to medium negative effect for high-achievers and a less than small negative effect for medium achievers. One possible threat to the internal validity of the study is experimental 'mortality', in that there is possible association between school tracking and student attrition and moving to others schools. If two such factors have a significant influence on subsequent achievement at the 10th grade, then selection bias may become an issue. In addition, it is unclear whether the authors used study weights to compensate for unequal probabilities of selection as a result of multistage cluster sampling used in NELS. Since there is no note on adjustments using study weight, the standard errors for OLS parameter estimates were presumably calculated based on the original sample rather than the weighted national population estimates. Since students within the same school or class tend to resemble each other more than their peers from another school or class, the standard errors reported in the study might be substantially underestimated, which would lead to an inflation of effect size. Therefore, the results should be interpreted with caution.

The study by Angrist and Lang (44) used static-group comparison and one-group pretest-posttest design to examine the effect of the Metropolitan Council for Educational Opportunity (Metco), a school integration program that sent low-achieving minority students from schools in Boston, United States to more affluent, higher-achieving suburbs, on Metco-receiving schools (compared to those without Metco) and students (both Metco students and Non-Metco students), respectively. OLS regression models were applied using both school- and student-level data. School-level data analysis showed that schools with Metco students had much higher average scores than those without. However, the study suffers from obvious selection bias in that the metco-receiving districts and schools tended to be more affluent and have higher scores; therefore one should not attribute the higher average scores to the Metco program alone. This bias is also demonstrated by regression analysis after controlling for district effect, which showed a consistent negative relation between the fraction of Metco students and average school scores. As for the student-level data, a parallel analysis of white student scores from Metco-receiving districts and other districts shows that the scores of white students are unaffected by the presence of Metco students (all minorities).

Hallinan (19) conducted an observational study into the effects of ability grouping on student growth in academic achievement, based on a longitudinal survey of students in grade 9 from six public high schools in a city in the Midwest of the United States. Hallinan used ability group assignment at mid-year as the treatment variable, which has four levels for English classes and five levels for mathematics classes, respectively. The results showed that students from schools that did not track students in eighth grade generally performed better than other students (as shown by positive parameter coefficient estimates across all ability cohorts) for English and mathematics, after controlling for pretest score, gender, race/ethnicity, SES, days absent from school, as well as each school as dummy variable. However, only the detracking coefficients for English ($\beta = 3.92$, $p < 0.05$) and mathematics achievement ($\beta = 10.22$, $p < 0.01$) of students in honors classes were statistically significant, suggesting that the honors-class ninth graders who came from detracked classes in eighth grade had scores that were on average 3.92 percentiles higher on the English test and 10.22 percentiles higher on the mathematics test than other honors-class 9th graders. This finding was surprising because few previous studies found appreciable academic effects of detracking for higher-achieving students. The study also found that assigning a student to a higher ability group had a generally positive effect on the achievement of the student, regardless of the student's learning ability. However, the predicted gains for higher ability placement should be interpreted with caution for several reasons. First, hypothetically moving a student from one group to another assumes that the unmeasured characteristics of the students for the two groups have the same distribution. In reality, students in different ability groups usually differ from each other in other ways that could affect achievement, such as motivation towards academic work, academic climate, and overall quality of instruction. None of these was measured in the study. Second, the parameter estimates were based on the particular distribution of students in each ability group in the sample. The distribution of achievement and potential confounding variables might affect the estimation of the coefficients.

Studies on academic outcomes of detracking for low-achieving students

Experimental studies. Cartwright and McIntosh (45) conducted a randomized controlled study that examined the

relative effects of three grouping procedures (heterogeneous, homogeneous, and flexible) on the academic achievement of 260 disadvantaged pupils in grades 1–3. The pupils were randomly assigned to one of the three groups taught by nine teachers. The authors also matched the teachers in each group based on the number of years that they had been teaching. The children assigned to heterogeneous groups were placed in the classroom on the basis of chronological age, while those assigned to homogeneous groups were placed in the classes based on intellectual ability and academic achievement in reading. The study found that first graders in the detracked group had the highest mean score in reading. The first graders in the flexible group had the highest mean score in computation. For the second graders, only the F ratio associated with computation was statistically significant with the detracked group having the highest mean score (39.1 vs. 22.0 for tracked and 18.7 for flexible). The detracked second graders also had the highest mean scores on the other three domains, even though the differences were not statistically significant at the 0.05 level. For the third graders, significant differences among the three groups were again found on computations, with the flexible group third graders scoring the highest.

Marascuilo and McSweeney's (37) study found no statistically significant differences in the performance of low-achieving students in detracked classes and their peers in tracked classes on the standardized Cooperative Social Studies Test (CSST) at the end of the first year ($F = 1.8, p > 0.05$), but low-achieving students who were assigned to detracked classes obtained higher scores on the U.S. Constitution test than did their counterparts in tracked classes ($F = 3.2$, $p < 0.05$). The study also found that the low-achieving students assigned to detracked classes significantly outscored their counterparts in tracked classes at the end of the second year on both the CSST ($F = 3.1$, $p < 0.05$) and a teacher-made test ($F = 6.0$, $p < 0.05$).

Quasi-experimental studies. Thacker (39) conducted a Nonequivalent Control Group Design study that examined the effects of administrator implemented homogeneous and heterogeneous instructional grouping techniques on sixth grade reading comprehension achievement at two schools in northern Indiana, United States. Descriptive statistics showed that the reading score mean for low-ability students exposed to heterogeneous instruction (317.50) was higher than that for low-ability students exposed to homogeneous instruction (308.75). However, ANOVA showed that this difference was not statistically significant ($F = 0.995$).

In the quasi-experimental study by Burries, Heubert, and Levin (40), the authors looked at the effects of universal acceleration with heterogeneous grouping on mathematics course completion and achievement of minority and economically disadvantaged students. One striking finding is that the percentage of minority students who met the mathematics commencement requirement after the universal acceleration tripled, from 25% to 75%. Also, a higher percentage of black, Latino, and low-SES students passed the exam in eighth-grade detracked classes than in eighth-grade tracked classes prior to the universal acceleration.

Saunders (46) conducted a quasi-experimental study that investigated the student achievement scores of 305 students in three schools in Arizona, United States. The study found that the high-track classes progressed at a more accelerated and in-depth rate than the medium and low-track classes. The students placed in ability-grouped mathematics classrooms increased learning at a higher rate than students placed in a heterogeneous group. This finding was consistent across gender and ethnicity. It was found that ability grouped students, regardless of their specific track, increased their previous year's mathematics test scores by 15.9% compared to the detracked students who had an increase by 1.3%. This was not statistically significant based on a paired samples t-test. The author concluded "ability-grouped students significantly outperformed their heterogeneous-grouped counterparts" (46, p. 100). However, descriptive statistics showed that the heterogeneously-grouped students' mean mathematics scores in 2003 (M = 50.40, SD = 16.89) and 2004 (M = 51.07, SD = 17.09) were much higher than those of the ability-grouped students (2003: M = 39.85, SD = 14.39; 2004: M = 46.18, SD = 16.96). This is troubling because the students in the two groups clearly did not start at the same academic level, which may be attributed to the difference in learning growth. The author also disaggregated the ability-grouped students into high-, medium-, and low-track groups. Tracked high-ability students had a substantial increase in overall mathematics score of 9.67 (from 67.37 to 77.04). However, no ANCOVA was conducted for high-ability students because the subsample size was less than 30. A statistically significant increase of 13.84 [$F(11,50) = 2.32$, $p < 0.05$] was detected for ability-grouped medium-group students (45.06 to 58.90). For ability-grouped low-ability students, no statistically significant increase (from 32.68 to 34.71) was found by ANCOVA [$F(21,102) = 2.32$, $p = 0.11$]. Due to the unique outcome variable and overt disparity between the detracked and tracked groups, results from this study were not included in the meta-analysis.

Observational studies. The study by Mulkey, Catsambis, Steelman, and Crain (42) also examined the academic outcomes of tracking on low-achieving students. Since the effect sizes were calculated using the pooled standard deviation of both tracked and untracked groups, one can also interpret the results as effect of detracking or heterogeneous grouping by changing the sign of the effect sizes (all effect sizes were converted to detracking effect in Table 1). In general, the study found that the tracking experience in middle school had a positive effect on all low-achieving students regardless

Table 1 Summary of part of reviewed empirical studies on achievement effects of detracking

Study	Location/Data Source	Grades	Design	Sample Size	Outcome Measures	Effects			Sources of Invalidity
						By ability level	By subject	Total	
Experimental Studies									
Cartwright & McIntosh, 1972	Linaouni School, Honolulu, Hawaii	1–3	RCT. All students were randomly assigned to 1 of 3 experimental groups: detracked, tracked, and flexible.	260 students	Word knowledge, reading and computation skills measured by Metropolitan Achievement Test (MAT), Peabody Picture Vocabulary Test (PPVT)		Rdg g1: +0.76 Rdg g2: +0.45 Rdg g3: −0.55 Math g1: +0.19 Math g2: +1.30 Math g3: +0.41	+0.34 +0.14	Confounding of teachers and treatment variable
Marascuilo & McSweeney, 1972	Berkeley Unified School District, CA	8–9	RCT. Detracked classes were created by proportional allocation and random assignment of students from high, mid, low tracks. Regular tracked classes were used as control group.	603 students 1 school	Teacher-made and standardized tests on social studies achievement	Teacher-made test H: −0.19 M: +0.52 L: +0.61 CSST H: −0.05 M: +0.48 L: +0.69	Social studies Teacher-made test: +0.31 CSST: +0.37	+0.31 +0.37	Treatment group is selected from volunteered students; unbalanced groups
Slavin & Karweit, 1985a	Wilmington, DE	4–6	RCT. Students & teachers were randomly assigned to team-assisted individualized instruction, within-class ability grouping, and a whole-class, group-paced instruction	345 students	Mathematics Computations Subscale of Comprehensive Test of Basic Skills		Math −0.74	−0.74	Within-class ability grouping

Table 1 Continued

Study	Location/Data Source	Grades	Design	Sample Size	Outcome Measures	Effects			Sources of Invalidity
						By ability level	By subject	Total	
Slavin & Karweit, 1985b	Hagerstown, MD	3–5	RCT. Students & teachers were randomly assigned to team-assisted individualized instruction, within-class ability grouping, and a whole-class, group-paced instruction	231 students	Mathematics Computations Subscale of Comprehensive Test of Basic Skills		Math −0.58	−0.58	
Kissoon-Singh, 1996	Canada	7	RCT. Students were randomly assigned to homogeneous or heterogeneous-ability groups	130 students	Canadian Cognitive Abilities Test, Students' Science Achievement Test, Students' Self-Efficacy Scale (SSES)	Science achievement H: +1.77 M: +3.54 Self efficacy H: +0.45 M: +2.09		+2.65	Study was conducted in a computer-based setting. The relative effects of instruction and technology were not disentangled.
Quasi-Experimental Studies									
Thacker, 1987	Indiana	6	Nonequivalent Control Group Design	172 students 2 schools	SRA Reading Comprehension Test	H: +0.14 M: +0.39 L: +0.30		+0.28	Lack of randomization
Burris, Heubert, & Levin, 2006	Nassau County, Long Island	8–12	Interrupted time series	985 students 1 district	ITBS math concepts subtest; advanced placement exams in calculus	All: +0.28 H: +0.22		+0.25	Selection bias, confounding of the accelerated curriculum and detracking

Table 1 Continued

Study	Location/Data Source	Grades	Design	Sample Size	Outcome Measures	Effects			Sources of Invalidity
						By ability level		By subject	Total	
Observational Studies										
Hawkins, 1999	Philadelphia, PA	7–8	One-group pre- and posttest study	970 students 1 school	Grades in English, reading, social studies, mathematics, and science			English: +0.17 Reading +0.05 Social: +0.05 Math: +0.01 Science: −0.001	+0.06	Selection bias, lack of control group
Hallinan, 2000	A Midwestern city	9–11	Observational study: Two-Limit Tobit Model Analysis	2581 students, 6 schools	Scores in English & math on a state test	English H: −0.70 L: −0.12	Math H: +0.27 L: +0.32	English: −0.41 Math: +0.29	−0.11	Comparisons are made based on model predictions that are unobserved
Angrist & Lang, 2004	Boston, Springfield, Brookline, MA	K-12	Static-group comparison, one-group pretest-posttest design, OLS regression (Treatment: Metco voluntary school integration program)	300 Metco students and about 6,000 resident students (Brookline)	Massachusetts Comprehen-sive Assessment System (MCAS) test scores	Metco students (minority) are more likely to graduate from high school than other boston public school students; some negative effect on scores of 3rd grade girls; no effect on scores of white students in receiving districts	A 10% increase in fraction Metco is associated with a 26% increase in fraction proficient in MCAS math	Metco students generally show more improvement between third and seventh grades than do non-Metco students.	Selection bias, statistical regres-sion	
Alvarez & Meban, 2006	Preuss School, University of CA, San Diego, CA	6–12	One-shot case study (Treatment: school-wide detracking)	300 students at middle school, 400 at high school, 100% low SES, 93.7% minorities.	Reading and math scores on California Achievement Test (CAT), graduation rate, SAT scores			70%–83% students at each grade scored ≥ 50th percentile on CAT math; >80% 9–12th Latino & Asian scored ≥ 50th percentile on CAT reading.	80% of the students form the first graduating class attended 4-year colleges; 20% attended community colleges	Selection bias (homogeneous student makeup), rival variables other than detracking, include university-school partnership, rigorous courses, abundant resources, etc.

Table 1 Continued

Study	Location/Data Source	Grades	Design	Sample Size	Outcome Measures	Effects: By ability level	Effects: By subject	Effects: Total	Sources of Invalidity
Argys, Rees, & Brewer, 1996	National, NELS: 88	8–10	Multinomial logit model of track placement	3,405 students	Math standardized scores	M diff / *d* *H*: −5.82 −0.44 *M*: −2.11 −0.16 *L*: +5.01 +0.25		Mean diff = −1.61 ES = −0.08	Experimental mortality (attrition, moving), selection bias
Mulkey, Catsambis, Steelman, & Crain, 2005	National, NELS: 88	8, 10, 12	Track placement propensity score analysis	5,895 students	10th grade:	Effect size for detracking		ES	History, maturation, selection bias, and multiple treatment interference.
						High: M / F; *Low*: M / F			
					Certainty graduation	+0.14 +0.14 −0.18 −0.18		−0.02	
					College plans	+0.02 +0.24 −0.11 −0.11		+0.01	
					Locus of control	0.00 0.00 0.00 0.00		0.00	
					Taking math	+0.87 0.00 +0.53 0.00		+0.35	
					School engagement	+0.10 0.00 −0.19 0.00		−0.02	
					Self-concept	+0.12 +0.12 −0.13 −0.13		0.00	
Mulkey, Catsambis, Steelman, & Crain, 2005	National, NELS: 88	8, 10, 12	Track placement propensity score analysis	5,895 students	12th Grade:	*High*: M / F; *Low*: M / F			
					Math test scores				
					Math credits	−0.33 −0.33 −0.33 −0.33		−.33	
					Math grades	0.00 0.00 0.00 0.00		.00	
						+0.07 +0.21 +0.25 −0.12		+.10	
Saunders, 2005	Glendale, AZ	6–7	Non-equivalent control group design, some students were tracked in grade 7	305 students 3 schools	District Assessment Test			Pre-posttest diff: Tracked: 15.88% Detracked: 1.3%	Lack of comparability between s groups in initial achievement

of gender (tracking ES = 0.33). However, there were differential effects of tracking on student mathematics grades by gender. Eighth grade tracking was positively associated with grades at the 12th grade for low-achieving females, who were slightly advantaged over their untracked counterparts with respect to the grades (tracking ES = 0.12); while a negative association of prior tracking experience and mathematics grades at the 12th grade was found for low-achieving males (tracking ES = −0.25). Again, any other experiences between the 8th and 12th grade may confound these effects.

The study by Argys, Rees, and Brewer (21) also used NELS: 88 data to examine the potential impact of detracking on low-achieving students through the estimation of a standard educational production function. The study indicated that moving from a low-track class to an untracked class would increase a student's mathematics score by 8.6% (from 58.07 to 63.08), but the gains would come at the expense of students in upper tracks. Based on the predicted mean score and standard deviation at each cohort (tracked and untracked), one can estimate the effect size of detracking academic outcomes for high-, medium-, and low-achieving students, respectively. The computed Cohen's *d* using pooled standard deviation for low-achieving students is 0.25, suggesting a small to medium effect.

During the second phase of their study on the Metco program, Angrist and Lang (44) conducted a one-group pretest-posttest comparison of the Metco students (low-achieving) in Brookline, Massachusetts, which enrolls 300 Metco students each year. The study revealed that Metco students generally showed more improvement between third and seventh grades than did non-Metco students. A negative impact for Metco students was found on the reading scores of third-grade black girls, which might be driven by the fact that Metco students were more likely to be female. These Metco students displaced relatively high-scoring black girls in the host district and hence lowered the average score for that subgroup. As for the more pronounced effect on the Metco students, there are several threats to the internal validity of the study. First, Metco students had significantly lower pretest scores than non-Metco students and hence might enjoy great gains due to regression to the mean, rather than a genuine effect of the Metco program. Second, it is possible that the results reflect more favorable sample selection for Metco students. The ideal evaluation strategy would use comparisons with an otherwise similar group of non-Metco students from Boston. Overall, there is little evidence of statistically significant effects of the Metco program on the non-Metco student achievement.

Alvarez and Meban (47) conducted a case study of a completely detracked school, the Preuss School, which is a charter school on the University of California, San Diego, United States campus. All students at the school came from low-income families and were enrolled in a single college-prep track. Minorities account for 93.7% of the student population. Eighty percent of the students from the first graduating class attended 4-year colleges, and 20% attended community colleges. The author wrote "this gives us existence proof that detracking can propel students from low-income households toward college eligibility and enrollment." Unfortunately, the study suffers from multiple threats to internal and external validities. First, there was selection bias in that the Preuss School used a lottery to select low-income students with high potential, so that the entering students were more socially and racially homogeneous than students from other schools. Factors such as student motivation, aspiration, and parental involvement can influence student desire to excel. In addition, multiple extraneous variables other than detracking can explain the students' academic progress, such as academic and social supports provided by the university-school partnership (including tutoring services by UCSD students), rigorous curriculum, ample educational resources provided by the university and private foundations etc. If these factors were not well controlled, the estimated effects of detracking might seriously be distorted. As many studies on tracking from other schools show, the lower-track students tend to be assigned less qualified teachers and received a smaller share of educational resources (21). In order to examine the pure effect of tracking or detracking, it is important to control for teacher characteristics and other educational input and use a control group. In short, the circumstances about the school in Alvarez and Meban's study are so different that there is little evidence that the success can be attributed to detracking alone.

Studies on non-academic outcomes of detracking for all students

Experimental studies. In Marascuilo and McSweeney's study, heterogeneously grouped students of all ability levels reported consistently less satisfaction with assignments and coursework than did their counterparts in regularly tracked classes (37). In addition, it was surprising that the dissatisfaction level of the heterogeneously grouped low-track students was the highest (Cohen's $d = -0.82$) even though they performed significantly better than the homogeneously grouped low-track students ($F = 3.2$, $p < 0.05$). As for the evaluation of self, both high and low track students in detracked classes tended to be more self-critical than their counterparts in tracked classes.

Slavin and Karweit conducted two randomized field experiments to investigate the effects of three classroom grouping methods on student attitude towards mathematics learning (38). For the first experiment, team-assisted individualized instruction and within-class ability grouping did not differ in their effects on student attitudes, but were both superior to whole-class instruction (detracking) [$F(2,12) = 4.06$,

$p < 0.05$]. However, the author didn't report the differences between adjusted means in effect sizes, as they did for the achievement outcome. For the second experiment, overall nested analysis of variance was significant [$F(3,18) = 5.41$, $p < 0.01$], with modified Bonferroni comparisons indicating that students in team-assisted individualized instruction classes exceeded all others in positive attitudes toward mathematics.

Kissoon-Singh conducted an experimental study to investigate the effects of studying in heterogeneous-ability groupings versus homogeneous groupings on the mathematics and science achievement and self-efficacy of 130 high and average ability students (36). The results showed that average ability students acquired significantly higher judgment of self-efficacy in heterogeneous rather than in homogeneous groups [$d = 2.09$, 95% CI 1.43, 2.69], but high-ability students' judgment of self efficacy did not differ significantly from high ability peers in homogeneous groups [$d = .45$, 95% CI −0.08, 0.97]. This suggested that heterogeneous-ability grouping was beneficial to average-ability students in terms of enhancing self-efficacy without being detrimental to the high-ability students.

Observational studies. The longitudinal study by Mulkey, Catsambis, Steelman, and Crain (42) examined whether heterogeneous grouping produces better psycho-social outcomes for 10th grade students. The study found that heterogeneous grouping tends to benefit high ability males the most (mean detracking $d = 0.25$) and low ability females the least (mean detracking ES = −0.11). The students placed in upper tracks experienced losses in mathematics self-concept at grade 10 (tracking $d = -0.12$) that negatively affected their mathematics grades at grade 12 (tracking $d = -0.07$ for males and −0.21 for females) and mathematics course-taking decisions for males (−0.87), suggesting that initial drops in self-concept for tracked students continued throughout high school and were correlated strongly with lower mathematics grades in high school. Both males and females with a propensity for the high track are less sure of high school graduation (tracking $d = -0.14$ for both genders) and less likely to make specific plans for attending college after high schools (tracking ES = −.02 for males and −.24 for females) if they are placed in tracked schools than their counterparts in untracked schools. However, both males and females with low-track propensity remain more certain of their high school graduation (tracking ES = 0.18) in schools where they experience tracking. This finding is contradictory with a few other studies that claim tracking tends to place low-achieving students at a disadvantage (eg 21). The authors explained that high-achieving students are less secure about themselves because they may be comparing their abilities with fellow high-track members (peer grouping effect). Like the limitations mentioned earlier regarding academic outcomes, this study suffers from multiple threats to internal and external validities because it examined the lasting effect of middle school tracking experience on psycho-social outcomes. Possible validity threats include history, maturation, selection bias, and multiple treatment interference. One of the virtues of this study is that the authors analyzed the differential effects of grouping by gender and ability level, which sometimes nullify each other when combined.

Most other studies that focused on the non-academic outcomes of detracking reform were qualitative or descriptive studies. Although these studies were not given preference in the present review, some provided a complement to the reviewed quantitative studies and provide more depth to the findings, especially on non-academic impacts of detracking. The studies reviewed above revealed various degrees of negative psychosocial impacts of detracking for students who were previously placed in low-track. Rubin (8) conducted a year-long case study of detracked ninth-grade English and History classes at Cedar High School in the San Francisco Bay Area, United States. Through collecting field notes from classroom observations and interviews of teachers and students, the author found a strong difference in teaching practices in an intentionally "balanced" class, as well as the reality for the students who underwent detracking. While two minority students attributed their academic success to interaction with their mixed race peer group, most students commented that the teachers' intentional placement of lower-achieving students with higher-achieving students sometimes "inadvertently raised tensions" and "ran the risk of exacerbating the dynamic between two groups" (8, p. 21). A few students felt they were disrespected and alienated in group-work situations. The author concluded that a more effective and sensitive implementation of detracking and other equity-geared reforms ought to go beyond simply placing students of various abilities in the same classroom.

Cooper (48) conducted a study using survey data of 1,090 ninth-grade students involved in a detracking reform at Liberty High School in Bakersfield, California, United States. Each core class of 20 students was intentionally balanced by race and ability level and was exposed to the same rigorous academic curriculum. The survey data showed that the majority of students reported a moderate level of enjoyment of their English and history classes, as evidenced by an overall mean score of 4.64 out of 7 (1 being least favorable and 7 being most favorable) for their enjoyment of the new program. The author found that the majority of students indicated that the program not only challenged them intellectually but also made them more culturally sensitive about issues in the society.

Conclusion

Previous studies indicated the negative consequences of tracking in public schools (49). In comparison with the

extensive research on ability-based grouping, there are considerably less empirical studies on detracking or alternatives to tracking. The systematic review reported here offers new insights into this topic by synthesizing the best available evidence concerning the academic and nonacademic effects of detracking on both low-achieving and high-or average-achieving students in the literature from the 1970s to present. As well as a review of background, theoretical framework, and practices outside of the United States, four types of empirical studies conducted in North America were reviewed and synthesized.

The findings of this review extended previous research on ability groupings, particularly on effects of heterogeneous groupings. Despite the variability in the types of grouping practice and outcomes in the education literature, the 15 reviewed studies, conducted over a 35-year period, provide evidence that detracking practice had moderately positive effects on the academic outcomes of low-achieving students, and no significant effects on the academic outcomes of high- or average-ability students. The evidence with respect to the non-academic impact of detracking is mixed. Students of various ability levels in heterogeneous classes reported outcomes ranging from being less secure, through raised tensions and no difference, to higher self-efficacy, and more positive attitudes. Understanding the academic and non-academic effects of detracking versus tracking in schools is an important education issue and continues to generate interest among educational practitioners and researchers in the 21st century. With the increasing pressure from both the state and federal governments to improve student achievement, schools are required to be more cognizant of which instructional method works best for enhancing student performance.

Implications for practice and future research

The main finding from this review is that detracking, or heterogeneous-ability grouping, was beneficial to low ability students in terms of enhancing their academic achievement without being detrimental to the high- and average-ability students. Therefore, a major implication is that heterogeneous grouping should be encouraged and promoted, especially in schools where the lower-track classes have been traditionally assigned fewer resources and less qualified teachers. This review does not support the competing claims that the performance of higher achieving students would decrease as a result of detracking. At a time when all students are expected to meet high standards, this review of evidence on the effectiveness of detracking provides valuable information for educators and scholars. Although there are methodological problems associated with some of the reviewed studies, the findings were consistent based on either all studies or only experimental studies, and revealed a lack of empirical evidence in the literature suggesting that detracking is detrimental to students of a particular group.

However, school administrators do need consider specific school contexts and characteristics of student populations while making decisions about optimal instructional strategies. Detracking should not be approached as a technical reform that simply changes the course structures. The reviewed studies suggest that the reform should be not only regrouping and restructuring of classrooms to accommodate students with various abilities, but a challenge to the status quo and the basic norms, policies, and practices that have traditionally governed schools. As Grant and Sleeter suggested, restructuring in and of itself offers no guarantee of an improved quality of education (50). Creating a learning environment in which all students feel valued and treated as capable learners is the first step in institutionalizing a commitment to both high academic standards and equal educational opportunities. Some schools in the reviewed studies had been struggling in creating a multicultural learning environment where all students can be successful. A purely technical approach to detracking may deflect attention away from more crucial issues for the poor, low-performing schools, such as better teachers and schooling. Without substantial investments in a school's capacity to meet these fundamental needs, detracking or other alternative strategies to tracking sometimes may aggravate existing inequalities and tensions across racial/ethnic and socio-economic groups.

This review is limited by the fact that there have been few published experimental studies on heterogeneous grouping in recent years. In many reviewed studies, it is impossible to determine the specific impact of grouping based on the data presented. As a result, the relative effects of detracking above and beyond the instruction, curriculum, and technology components cannot be disentangled. Literature on this topic needs to be updated to incorporate the more diverse student population that is present in today's public schools. In addition to examining prior academic ability, more studies are needed to examine differential effects of detracking reform on student subgroups across racial and socio-economic factors. Teacher effectiveness should be examined and compared in detail in both the homogeneous and heterogeneous instructional settings. Furthermore, the effect of peer pressure in grouping situations, parental involvement, and interest in various instructional practices should be explored. In conclusion, educators in today's school system continue to be faced with a classical dilemma: how to achieve equity and efficiency at the same time? Additional research is needed to assist in the development of strategies to organize instruction so that the needs of all ability groups are met and the potential of every student is realized.

References

1. Yonezawa S, Wells AS, Serna I. Choosing tracks: "freedom of choice" in detracking schools. *American Education Research Journal* 2002; 39: 37–68.
2. Ansalone G. Poverty, tracking, and the social construction of failure: international perspectives on tracking. *Journal of Children & Poverty* 2003; 9: 3–20.
3. Loveness T. *The Tracking Wars: State Reform Meets School Policy*. Washington, DC: Brookings Institution Press, 1999.
4. Maddy-Bernstein C, Matias ZB, Cunanan ES, Krall BT, Kantenberger J, Iliff L. *Inclusion/Detracking: A Resource Guide*. Berkeley, CA: National Center for Research in Vocational Education, 1995.
5. Oakes J. *Keeping Track: How Schools Structure Inequality*. New Haven, CT: Yale University Press, 1985.
6. Slavin RE. Ability grouping and student achievement in elementary schools: a best-evidence synthesis. *Review of Educational Research* 1987; 57: 293–336.
7. Slavin RE. Achievement effects of ability grouping in secondary schools: a best-evidence synthesis. *Review of Educational Research* 1990; 60: 471–99.
8. Rubin BC. Unpacking detracking: when progressive pedagogy meets students' social worlds. *American Educational Research Journal* 2003; 40: 539–89.
9. Kulik CLC, Kulik JA. Research synthesis on ability grouping. *Educational Leadership* 1982; 39: 619–21.
10. Mosteller F, Light RJ, Sachs J. Sustained inquiry in education: lessons from skill grouping and class size. *Harvard Educational Review* 1996; 66: 797–842.
11. Noland TK, Taylor BL. The effects of ability grouping: a meta-analysis of research findings. Paper presented at the annual meeting of the 70th American Educational Research Association, San Francisco, CA, 1986.
12. Agne K. "Kill the baby": making all things equal. *Educational Horizons* 1999; 77: 140–7.
13. Hallinan MT. The detracking movement: why children are still grouped by ability. *Education Next* 2004; 4: 72–6.
14. Gamoran A, Weinstein M. Differentiation and opportunity in restructured schools. *American Journal of Education* 1998; 106: 385–431.
15. Mallery JL, Mallery JG. The American legacy of ability grouping: tracking reconsidered. *Multicultural Education* 1999; 7: 13–5.
16. Burris CC, Welner KG. Closing the achievement gap by detracking. *Phi Delta Kappan* 2005; 86: 594–9.
17. Loveness T. Will tracking reform promote social equity? *Educational Leadership* 1999; 56: 28–32.
18. Kariya T, Rosenbaum JE. Bright flight: unintended consequences of detracking policy in Japan. *American Journal of Education* 1999; 107: 210–30.
19. Hallinan MT. Ability group effects on high school learning outcomes. Paper presented at the 95th annual meeting of the American Sociological Association, Washington, DC, 2000.
20. Wheelock, A. *Crossing the Tracks: How "Untracking" Can Save America's Schools*. New York, NY: New Press, 1992.
21. Argys, LM, Rees DI, Brewer DJ. Detracking America's schools: equity at zero cost? *Journal of Policy Analysis and Management* 1996; 15: 623–45.
22. Rawls J. *A Theory of Justice*. Cambridge, MA: Harvard University Press, 1971.
23. Osborne MJ, Rubenstein A. *A Course in Game Theory*. Cambridge, MA: MIT Press, 1994.
24. Tan C. The potential of Singapore's ability driven education to prepare students for a knowledge economy. *International Education Journal* 2005; 6: 446–53.
25. Ishida, H. *Social Mobility in Contemporary Japan*. Stanford, CA: Stanford University Press, 1993.
26. Hou J, Shan X, Bi F. The necessity of implementing school tracking in China. *Shandong Education Journal* 2006; 113: 15–6.
27. Whitburn, J. *Motivating 14–16 Year Olds: How do the Swiss do it?* Paper presented at the ESRC Funded Seminar Series at the London School of Economics and Political Science, 2003. http://cep.lse.ac.uk/events/seminars/motivation/WhitburnPaper.pdf (accessed 2 August 2007).
28. LeTendre GK, Hofer BK, Shimizu H. What is tracking? cultural expectations in the United States, Germany, and Japan. *American Educational Research Journal* 2003; 40: 4–43.
29. Ansalone G. Perceptions of ability and equity in the U.S. and Japan: understanding the pervasiveness of tracking. *Radical Pedagogy* 2006; 8(1). http://radicalpedagogy.icaap.org/content/issue8_1/ansalone.html (accessed 2 August 2007).
30. Glass GV, McGaw B, Smith, ML. *Meta-analysis in Social Research*. Thousand Oaks, CA: Sage, 1981.
31. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.
32. Slavin RE. Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Educational Researcher* 1986; 15: 5–11.
33. Cooper H. *Synthesizing Research*. 3rd edition. Thousand Oaks, CA: Sage, 1998.
34. Borenstein M. *Comprehensive Meta-analysis Software*. Englewood, NJ: BioStat, 2005.
35. Hunter JE, Schmidt FL. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: Sage, 2004.
36. Kissoon-Singh SB. Cooperative groupings and computer based instruction: The effects of grouping by ability. Doctoral dissertation, University of Toronto, Toronto, Ontario, Canada, 1996.
37. Marascuilo LA, McSweeney M. Tracking and minority student attitudes and performance. *Urban Education* 1972; 6: 303–19.
38. Slavin RE, Karweit NL. Effects of whole class, ability grouped, and individualized instruction on mathematics achievement. *American Educational Research Journal* 1985; 22: 351–67.
39. Thacker JL. Effects of administrator implemented homogeneous and heterogeneous grouping on reading achievement of selected

sixth-grade students. Doctoral dissertation, Andrews University, Berrien Springs, MI, 1987.
40. Burris CC, Heubert JP, Levin HM. Accelerating mathematics achievement using heterogeneous grouping. *American Educational Research Journal* 2006; 43: 105–36.
41. Hawkins MA. Student and faculty perceptions of educational change: The move from homogeneous to heterogeneous grouping in eighth grade. Doctoral dissertation, Temple University, Philadelphia, PA, 1999.
42. Mulkey LM, Catsambis S, Steelman LC, Crain RL. The long-term effects of ability grouping in mathematics: a national investigation. *Social Psychology of Education* 2005; 8: 137–77.
43. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997; 127: 757–63.
44. Angrist JD, Lang K. Does school integration generate peer effects? evidence from Boston's Metco Program. *The American Economic Review* 2004; 94: 1613–34.
45. Cartwright GP, McIntosh DK. Three approaches to grouping procedures for the education of disadvantaged primary school children. *Journal of Educational Research* 1972; 65: 425–9.
46. Saunders R. A comparison study of the academic effects of ability grouping versus heterogeneous grouping in mathematics instruction. Doctoral dissertation, Arizona State University, Tempe, AZ, 2005.
47. Alvarez D, Meban H. Whole-school detracking: a strategy for equity and excellence. *Theory into Practice* 2006; 45: 82–9.
48. Cooper R. Urban school reform: student responses to detracking in a racially mixed high school. *Journal of Education for Students Placed at Risk* 1999; 4: 259–75.
49. Ascher C. *Successful Detracking in Middle and Senior High Schools*. New York, NY: ERIC Clearinghouse on Urban Education, 1992.
50. Grant C, Sleeter C. Who determines teacher work? the debate continues. *Teaching & Teacher Education* 1987; 3: 61–4.