Journal of Urban Economics **51,** 497–514 (2002)
doi:10.1006/juec.2001.2255, available online at http://www.idealibrary.com on IDEAL®

School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?[1]

David N. Figlio

Department of Economics, University of Florida, Gainesville, Florida 32611-7140; and National Bureau of Economic Research
E-mail: figlio@ufl.edu

and

Marianne E. Page

Department of Economics, University of California at Davis, Davis, California 95616-8578
E-mail: mepage@ucdavis.edu

Received August 18, 2000; revised August 13, 2001; published online November 7, 2001

Tracking programs have been criticized on the grounds that they harm disadvantaged children. The bulk of empirical research supports this view, but existing studies compare outcomes across students placed in different tracks. Track placement is likely to be endogenous with respect to student outcomes. We use a new strategy for overcoming the endogeneity of track placement and find no evidence that tracking hurts low-ability children. Previous studies have also been based on the assumption that students' enrollment decisions are unrelated to whether or not the school tracks. When we account for the possibility that tracking programs affect school choice, we find evidence that they may help low-ability children. © 2002 Elsevier Science (USA)

I. INTRODUCTION

Tracking students into separate classrooms according to their ability level has been common practice in the United States since the turn of the century. In

[1] We appreciate the data and financial support of the U.S. Department of Education, as well as helpful conversations with Laura Argys, Dominic Brewer, David Card, Dennis Epple, Larry Kenny, Dick Murnane, Dan Rees, Steve Rivkin, Rich Romano, and seminar participants at Berkeley, Columbia, Cornell, Florida, Oregon, Norwegian Institute of Science and Technology, South Florida, and SUNY–Albany, and the Econometric Society, APPAM, Population Association of America, and American Educational Finance Association meetings. Thanks to Jan Brueckner and an anonymous referee for helpful comments. The financial support of the National Science Foundation through Grant SBR-9816629 is gratefully acknowledged. This project was completed while Page was a Visiting Scholar at the Joint Center for Poverty Research.

0094-1190/02 $35.00

recent years, however, tracking has fallen into disfavor because research suggests that it reduces achievement among disadvantaged students (Alexander and McDill [1]; Argys *et al.* [2]; Betts and Shkolnik [3]; Gamoran [10]; Hoffer [12]; Kerckhoff [14]; and Vanfossen *et al.* [23]).[2] This finding is so prevalent in the literature that the National Education Association is currently recommending that tracking programs be abolished. It appears as though schools are heeding the NEA's advice: although there are no published statistics on the number of tracking programs in the United States, we estimate that between 1987 and 1993 the number of schools that maintained programs for gifted children declined by 7%.[3]

Proponents of ability grouping argue that narrowing the range of student abilities within a classroom allows teachers to target instruction at a level more closely aligned with student needs. Teachers of high-ability students can provide them with more challenging material or present standard material at a faster pace than is possible when less-able students' needs also have to be met. At the same time, low-ability students are expected to benefit from the slower pace or alternative teaching methods that become feasible when teachers are not simultaneously responsible for engaging the students, high-ability peers. Critics of tracking programs, however, claim that when disadvantaged children are tracked they lose the opportunity to benefit from positive peer effects that might be gleaned from coming into regular contact with more-able students. They also maintain that schools with tracking programs systematically redistribute resources away from low-ability students toward high-ability students, and that less capable teachers are disproportionately assigned to low-ability tracks. Although these arguments are largely speculative, studies of tracking programs that have been conducted to date indicate that they widen, rather than narrow, the test score distribution.

There are several reasons that economists should care about the effect of tracking on student achievement. First, it is well known that the wage distribution has spread over the last 30 years (for example, see Bound and Johnson [5] and Juhn *et al*. [13]), and there is also evidence that student test scores and future earnings are linked (Murnane *et al.* [15]). If tracking does, indeed, increase test score inequality then it may also lead to further increases in wage inequality. Second, economists have long contributed to the literature on education production functions, which focuses on the impact of school inputs on student outcomes. That literature, which has been largely unsuccessful at identifying factors that affect *between* school variation, also points to the fact that much of

[2]Additional critiques can be found in Braddock and Slavin [7], Gamoran [11], Oakes [16], Rosenbaum [20], and Wheelock [24].

[3]These estimates are based on schools included in the Schools and Staffing Survey: in 1987, 74% of the high schools in our sample maintained a program for gifted students, and 54% maintained both a remedial and a gifted program. In 1993, 69% maintained a gifted program and 46% had both a gifted and a remedial program. Among elementary schools, the fraction with gifted programs fell from 78% to 72% and the fraction with both types of programs fell from 67% to 64%.

the variation in student outcomes is *within* schools. Tracking may partly explain why this within school variation exists. Conversely, if students benefit from being educated among similarly skilled children, then tracking may be a school "input" that can help improve outcomes among America's most disadvantaged citizens.

In this paper we highlight three empirical problems that have impeded researchers' ability to assess the effect of tracking programs on students' test scores. We note that previous studies have not adequately accounted for the possibility that a student's track placement is endogenous or that the existence of a tracking program may affect an individual's school choice. Discussion of nonrandom school selection goes back to Tiebout [22] and is of fundamental concern in nearly every empirical study of school effects but has not been acknowledged in the tracking literature. When we address these issues we find no evidence that low-ability students are harmed by being grouped together and conclude that the trend away from tracking is misguided. In fact, we find that tracking programs may be associated with test score *gains* for students in the bottom third of the initial test score distribution. We conclude that the move to end tracking may harm the very students that it is intended to help.

The rest of this paper is laid out as follows: the next section provides a review of the tracking literature and discusses the difficulties associated with identifying program effects. An overview of our data is presented in Section III. In Section IV we use our data to re-estimate the impact of tracking on low-ability children, paying special attention to the problems discussed in Section II. Section V concludes.

II. PREVIOUS RESEARCH

Tracking studies can be roughly classified into two categories: those that estimate the effect of tracking on *mean* achievement and those that investigate the impact of tracking on the *distribution* of achievement. Studies that compare mean outcomes across schools with and without tracking programs generally find that they have small, and typically insignificant, effects (Slavin [21]). More recent investigations have examined the effect of tracking on the distribution of test scores by using within school variation in individuals' track placement (e.g., Alexander and McDill [1]; Gamoran [10]; Vanfossen *et al.* [23]) or by comparing students in different tracks to those placed in heterogeneous classrooms (e.g., Argys *et al.* [2]; Betts and Shkolnik [3]; Hoffer [12]; Kerckhoff [14]). These studies estimate the type of regression model

$$\Delta A_i = \alpha + \beta_1 H_i + \beta_2 M_i + \beta_3 L_i + \gamma X_i + \varepsilon_i, \tag{1}$$

where ΔA_i represents the change in student i's test score from time $t-1$ to time t, H_i, M_i, and L_i are dummy variables representing placement in a high-ability, medium-ability, or low-ability classroom, and β_1, β_2, and β_3 are interpreted as the degree to which being placed in one of these classrooms affects test score growth relative to being placed in an untracked classroom. The vector X_i contains student-level covariates, usually including a base-year test score. The results produced by these studies almost universally support the hypothesis that tracking programs benefit the upper tail of the ability distribution at the expense of the lower tail. Consequently, a number of policy analysts have called for an end to tracking practices (Oakes [17]).

As has been noted in the past, however, at least two factors complicate the identification of differential tracking effects across student types. The first problem researchers face is that estimates of $\beta_1 - \beta_3$ are likely to suffer from omitted variables bias, even when a prior test score is included in the regression equation, because other factors unobservable to the researcher will affect track placement and some of these factors may be correlated with test score growth. Oakes [16] provides evidence that teachers' evaluations of student abilities, track assignments in previous years, and student motivation all affect track placement. Several innovative studies have recently emerged that use sophisticated econometric techniques to try to address this problem, but none of the approaches has been entirely satisfactory. Betts and Shkolnik [3], for example, compare students in tracked classrooms to students in untracked classrooms in which average test scores are similar. While their approach has the advantage of comparing outcomes across students who are similar, it is unclear why we should expect students in classrooms that are differentiated by their label (tracked vs untracked) but not differentiated by the composition of their classmates to produce different outcomes. Hoffer [12] uses a propensity score method to divide students into quintiles based on similar backgrounds and then compares outcomes for students in high, medium, and low tracks to the average heterogeneously grouped student within each quintile. The estimates produced by this method will still be subject to omitted variables bias, however, unless the quintiles are chosen based on all relevant characteristics. Argys *et al.* [2] account for selection into particular tracks using regional indicators, urbanicity indicators, and student body characteristics as instruments in a two-stage least squares procedure, but the authors do not provide evidence that these variables are exogenous with respect to the dependent variable. When we estimate an expanded version of Eq. (1), we find that some of these instruments are correlated with the change in individuals' test scores. Although all of these researchers have made valiant attempts to address the problem, it remains unclear whether their finding that children placed in low-ability tracks have lower test scores is a causal effect.

We propose a simpler and cleaner strategy for circumventing the selection problem, which is to eliminate track placement from the model altogether and, instead, estimate the effect of attending a tracked school separately for students in the top, middle, and bottom thirds of an initial test score distribution.[4] This

[4]An alternative approach would have been to use the full sample to estimate a model of test score growth as a function of the school's tracking status interacted with the individual's initial

allows us to identify the effect of tracking solely from variation across schools, rather than from variation both across schools and across student types. Provided that school choice is unrelated to a school's tracking status, this strategy will yield unbiased tracking estimates for each ability level.[5]

The second empirical issue is that there is a great deal of ambiguity about what it means to say that a school "tracks." This is particularly true at the high school level, where students may naturally sort themselves into different ability groups even if there is not a formal tracking program. If a secondary school offers its seniors courses in both algebra and calculus, and if high-ability students tend to sort into calculus and low-ability students sort into algebra, does this mean that the school tracks? Or does tracking mean that college bound students take an entirely different *set* of courses from those who are not college bound? Researchers and policymakers agree that tracking involves ability grouping, but rarely do studies or policy discussions clarify specifically which types of programs "count" as tracking programs and which do not. School principals, parents, and students also have a wide range of ideas about what constitutes tracking (Gamoran [10]; Rosenbaum [19]; Rees *et al.* [18])[6] and because survey questions rarely define it, this makes interpretation of estimates difficult. Like other studies, our study is also constrained by ambiguity inherent in survey questions, but we add to the literature by investigating the sensitivity of results to a number of different definitions of tracking.[7]

---

test score. We find, however, that coeffecient estimates on many control variables vary substantially across ability groups. We also find that the estimated tracking effect is not linear with respect to the student's initial test score. We, therefore, think that it is preferable to estimate the effect of tracking separately for each ability group.

[5]Although this strategy will not provide direct estimates of the effect of being placed in a specific track, the estimates produced by this strategy answer the question that is at the heart of the tracking debate, which is whether a policy of tracking is harmful to particular types of students. Furthermore, the student's initial test score is a good predictor of track placement, conditional on a school tracking. In tracked schools, students ranking in the bottom third of the eighth grade test score distribution are six times as likely to be placed in a mathematics class identified by the teacher as "generally of low ability" as to be placed in a mathematics class identified by the teacher as "generally of high ability." On the other hand, students ranking in the top third of the eighth grade test score distribution are ten times as likely to be placed in a mathematics class identified by the teacher as "generally of high ability" as to be placed in a mathematics class identified by the teacher as "generally of low ability." The effect of tracking on low-ability students is therefore likely similar to the average effect on students of being placed in a low-ability track.

[6]Indeed, we find a great deal of within school variation in students' reported 8th grade tracking status, which suggests that using student reports would lead to noisy measures of true tracking status.

[7]Rees *et al.* [18], in a comment on Betts and Shkolnik [3], criticize papers that use principal-reported measures of tracking as unreliable, given heterogeneity in the ways in which school tracking policies could be classified. Their preferred method involves relying on teacher reports of whether a specific *class* is tracked. In this paper, we employ both teacher-reported and principal-reported measures of ability tracking, thereby reducing the validity of this criticism as it pertains to our own work.

Our third contribution is to consider the possibility that school choice is affected by whether or not a school tracks. If particular types of students are attracted to schools with tracking programs then the estimates produced by cross-school comparisons may still be biased. Suppose, for example, that tracking helps high-ability students at the expense of low-ability students. Then high-ability students with high levels of motivation (or better information) will be more likely to select schools with tracking programs, and low-ability students with high levels of motivation will be more likely to avoid schools with tracking programs. This would lead to upward biased estimates for high-ability students and downward biased estimates for low-ability students. Previous studies have not considered this possibility. In order to address it, we conduct an instrumental variables procedure using as instruments state and county level variables that might affect a school's decision to track but are unlikely to be correlated with student outcomes. The estimates produced by the IV procedure suggest that tracking programs may actually benefit low-ability students.

III. DATA

Our analyses of the relationship between tracking and achievement gains are based on data from the National Education Longitudinal Study (NELS). The NELS began in 1988 when a nationally representative sample of schools was surveyed. In that year, a random sample of 8th grade students was interviewed within each school, and these students were then reinterviewed in subsequent years, resulting in a dataset that contains test scores and other individual, family background, and school level variables for the same nationally representative set of students in 8th, 10th, and 12th grade. A comparable dataset covering elementary grades does not exist.

Our dependent variable is the change from 8th to 10th grade in the student's item response theory (IRT) math score.[8] We estimate the effects of tracking on mathematics because there is a well-documented link between math test scores and labor market outcomes (e.g., Bishop [4]; and Murnane *et al.* [15]), and because math test scores are the outcome most frequently studied in the tracking literature. The NELS provides standardized test scores in addition to IRT scores, but we chose the IRT score because it concretely characterizes students' progress from one year to the next, whereas the standardized test score is a further transformed measure and merely reflects individuals' *relative* positions in the test score distribution. In our sample, the mean IRT score in

[8]IRT scores are essentially raw test scores in which responses to more difficult questions are weighted more heavily. The NELS uses item response theory to construct its test scores in order to reduce the potential problems associated with ceiling effects. Raw test scores are not provided.

8th grade is 35.8 points, and its standard deviation is 11.8 points. The 10th grade IRT score for our sample is 42.4 points with a standard deviation of 13.9 points.

As discussed in Section II, our ability to distinguish "tracked" from "untracked" schools is complicated by the fact that tracking means different things to different people. At the high school level, tracking may consist of a series of classes intended for precollege vs vocational students, or it may refer to the grouping of particular classes by ability level. Furthermore, a school may not *formally* group its students by ability level but may offer a set of subject specific classes that are aimed at different ability groups (e.g., calculus, precalculus, algebra) and into which students of different ability levels naturally sort themselves. Like other datasets, survey questions that address tracking in the NELS are vague about what exactly tracking is. For example, school principals are asked "Does your school track in mathematics?" but they are not told what types of programs constitute tracking in mathematics.

We are interested in assessing the effect of being schooled in a classroom with similarly skilled students relative to the effect of being schooled in a classroom that has a larger variance in student abilities. Of the tracking-related questions asked in the NELS, principals' answers to the above question seem to us to provide the cleanest information on whether or not mathematics classes are grouped by ability level, but because the definition of tracking is not explicit, we may misclassify some schools in which ability grouping takes place. In order to explore the sensitivity of our results to the way in which tracking is defined, we have created several alternative definitions of tracking which are based on teacher reports of classroom homogeneity, and on teacher reports of the official "track" of the class that he or she teaches. The more consistent results are across varying definitions of ability tracking, the more confident one can be that an estimated tracking effect is genuine.

Our alternative tracking measures are created from the following information: for each student surveyed in the NELS, two of her teachers were also surveyed. The teachers provided instruction in either math, science, English, or history and were asked to answer questions about the math, science, English, or history class in which the student was enrolled. Each teacher was asked whether he or she "best describes the achievement of the [eighth] tenth graders in this class compared with the average [eighth] tenth grade student in the school" as "higher achievement levels," "average achievement levels," "lower achievement levels," or "widely varying achievement levels." Teachers also reported on whether the track that best described the class was "advanced," "academic," "general," or "vocational." We used the responses to these questions to create additional tracking variables, which we discuss in the next section. The answers to these questions have the advantage of providing information on the composition of a particular classroom but may be even more subject to measurement error than principals' responses.[9]

Our sample includes all public school students with 10th grade school identifiers who have both 8th and 10th grade test scores, information on tracking, and information on the following covariates: indicators for whether the student is white, black, or Hispanic, dummies for the student's family income (under $15,000; between $15,000 and $35,000; between $35,000 and $50,000; between $50,000 and $100,000; over $100,000; and a missing income indicator), indicators for whether the student's most highly educated parent is a high school graduate or college graduate, region-of-country indicators (four Census regions), and indicators of central city or suburban residence.[10] In order to reduce the possibility that our tracking estimates are picking up something else about the school environment, we also include a set of school level characteristics in our regressions. These include the fraction of students who are white, the fraction of students who are free lunch-eligible, the number of days in the school year, the student–teacher ratio, the size of the 10th grade cohort, and the highest teacher salary.[11] Our sample excludes students who are missing these variables. Finally, we eliminate students who are missing the instrumental variables that we require for the two stage least squares analysis described in Section IV.[12] This leaves us with a sample of 7676 students with principal-provided information on whether their school tracks in mathematics.[13] When we use the other

[9]Argys *et al.* [2] provide evidence that tracking measures created from these two sets of questions may be quite distinct. In other words, a teacher will not necessarily characterize her class as being both "high achieving" and "advanced," or as both "low-achieving" and "general."

[10]We have experimented with including many other covariates, including measures of personal possessions, measures of family involvement in school, etc., and have found that these variables do not affect our estimates.

[11]Our empirical results are very similar whether or not we include these school level variables in our regressions.

[12]We lost 354 observations due to missing instrumental variables; this is because our instruments are measured at the county level, and we are missing the ability to identify county for some of the schools. There does not appear to be any systematic difference between the included and excluded students. For example, there are no statistically significant differences in their family background variables. The estimates presented in Tables 2 and 3 are not substantively affected by the inclusion of these additional observations.

[13]We arrive at this sample size as follows: there are 15,217 public school observations in the NELS with 10th grade school identification numbers. Of these, 12,817 observations have data on 8th and 10th grade mathematics tests. Of these, 12,008 have principal-reported measures of tracking. Of these, 11,642 have all of the individual level covariates. 354 observations are lost due to missing instrumental variables, and another 3612 observations are lost due to missing school level characteristics. We have also run all of our NELS regressions without the school level variables, which allows us to boost our sample sizes to 11,288 (principal-reported tracking measure) and 9516 (teacher reported tracking measures). The estimates produced by those regressions were virtually identical to those presented in Tables 2, 3, and 5. We prefer to report estimates based on the smaller samples to help reduce possible concerns that our estimates reflect the effect of unobserved school characteristics.

definitions of tracking our sample is reduced to 5948 students because mathematics teachers were not surveyed in every school.

Table 1 provides summary statistics by students' tracking status, using the principal-reported measure of tracking. The proportion of students coming from families with incomes greater than \$50,000 (in 1988 dollars) is nearly 50% higher among tracked students than among untracked students, and the fraction of tracked students with college-educated parents is almost one-third higher than the corresponding fraction of untracked students. Likewise, students attending tracked schools tend to have higher eighth grade test scores and exhibit higher test score growth than their counterparts: average eighth grade test scores are 1.6 points higher in tracked schools, and average test score gains are 0.2 points higher. Table 1 also indicates that relative to untracked students, students attending schools with tracking programs are more likely to live in urban environments. The prevalence of tracking programs also varies by region. These observed differences raise the possibility that tracking programs influence selection into schools.[14] Alternatively, the differences may indicate that student body composition influences a school's decision to track.

IV. DO TRACKING PROGRAMS HARM LOW-ABILITY STUDENTS?

A. *Addressing the Endogeneity of Track Placement*

We start by using Eq. (1) to replicate the results from early studies. Our dependent variable is the 8th to 10th grade change in the student's IRT math score. If the student's math teacher characterizes the student's class as being of "above average" ability then the student is labeled "high track," if the teacher characterizes her class as "average" then we classify the student as "middle track," and if the teacher characterizes her class as "below average" then the student is classified as low track. Individuals whose teachers describe the ability level of their students as "widely varying" are classified as untracked. This is the same classification that Argys *et al.* [2] use. The results produced by this exercise are shown in the first column of Table 2. Like the existing literature, the estimates in column 1 suggest that students placed in low-ability tracks experience smaller achievement gains and students placed in high-ability tracks experience larger achievement gains than students with similar test scores and family background characteristics who are placed in heterogeneous classrooms. The point estimate of −6.0 on low-track placement suggests that students in low-ability tracks experience achievement gains that are three-quarters of a standard deviation lower than similar students in heterogeneous classrooms, and the point estimate of 2.2 on high-track placement suggests that students placed

[14] These patterns are also observed when the other definitions of tracking are used.

TABLE 1
Sample Means of Variables Included in Achievement Equations

Variable	Mean in tracked schools (standard deviation)	Mean in untracked schools (standard deviation)	p-Value of difference
Change in math test score from 8th to 10th grade	7.290 (6.667)	7.117 (6.520)	0.364
8th grade math test score	36.640 (11.880)	35.059 (11.438)	0.000
White	0.729	0.768	0.002
Black	0.096	0.093	0.596
Hispanic	0.108	0.087	0.016
Income under $15,000	0.154	0.208	0.000
Income $15–35,000	0.341	0.384	0.002
Income $35–50,000	0.213	0.186	0.026
Income $50–100,000	0.147	0.110	0.000
Income over $100,000	0.066	0.035	0.000
Parent high school graduate	0.915	0.894	0.011
Parent college graduate	0.293	0.217	0.000
Central city	0.205	0.138	0.000
Suburb	0.453	0.367	0.000
Northeast	0.223	0.125	0.000
South	0.318	0.403	0.000
West	0.187	0.147	0.001
Days in school year	179.664 (3.333)	178.962 (3.018)	0.000
Percent white in school	74.318 (29.309)	79.276 (27.178)	0.000
Percent free lunch eligible in school	18.871 (19.087)	23.664 (20.441)	0.000
Highest teacher salary in school (1000s)	39.803 (8.111)	36.044 (7.707)	0.000
Cohort enrollment (1000s)	0.347 (0.216)	0.267 (0.189)	0.000
Student–teacher ratio	15.881 (3.675)	15.425 (3.583)	0.000

Note: The variables for which standard deviations are not provided are dichotomous variables. Sample size: 7676 observations (5949 in tracked schools, by principal report).

in high-ability tracks gain about one-third of a standard deviation above similar students in heterogeneous classrooms.[15] Of course, the question is whether these observed differences are causal—they are also consistent with inherent,

[15]Using the same dataset and controlling for 8th grade test score, Argys *et al.* [2] find that students placed in high-ability tracks have 10th grade math achievement scores that are about 25% of a standard deviation higher than similar students placed in heterogeneous classrooms, and students placed in low-ability tracks have 10th grade test scores that are about 25% of a standard deviation

unobserved differences in student characteristics that affect their selection into particular tracks.

In the next three columns of Table 2, we attempt to address this selection problem by dividing the sample into the top, middle, and bottom thirds of the 8th grade test score distribution, and then estimating for each subsample the effect of attending a tracked school. In other words we estimate

$$\Delta A_i = \phi + \alpha T_i + \gamma X_i + \varepsilon_i, \tag{2}$$

where T_i indicates whether individual i's principal reports that individual i's school tracks in mathematics and X_i is a vector of control variables. Relative to existing studies, the advantage of this approach is that $\hat{\alpha}$ is identified only from variation in tracking status across schools. Estimates based on Eq. (1) are identified both from variation in tracking status across schools and from variation across student types. As discussed in Section II, some of the test score variation across student types probably results from factors other than the track into which they were placed. As shown in the next three columns of Table 2, the estimates produced by this model lend themselves toward a different conclusion from those in column 1.[16] The estimated coefficient on tracking is negative but trivial in magnitude for high-, middle-, and low-ability students (-0.19, -0.06, and -0.40) and none of the estimates are significantly different from zero. These results suggest that the lower test score gains observed among students in low-ability tracks stem not from their track placement, but rather from unobserved factors correlated with track placement.

Rees *et al.* [18] argue that principal-reported measures of tracking will be unreliable because of ambiguities about what "tracking" means. In the spirit of their argument, we therefore re-estimate Eq. (2) using six alternative definitions of tracking that are constructed from teacher reports. The first three measures are based on teachers' descriptions of their classes as being of "high," "average," "low," or "widely varying" ability, while the others are developed from teachers' descriptions of their classes as being "advanced," "academic," or "general."[17] As a first alternative, we define a school as a "tracked" school if at least one teacher reports that his class is "high-achieving" and at least one teacher reports that his class is "low-achieving." Our second alternative is based on the number of different types of responses ("high-achieving," "average-achieving," and "low-achieving") observed at a given school and can take on values ranging from

---

below similar students placed in heterogeneous classrooms. This result is based on specifications that account for selection into tracks, as described in Section III.

[16]Because our "treatment" variable does not vary below the school level, we adjust our standard error estimates to account for error correlation within schools, using the Huber standard error correction.

[17]Teachers are also given the choice of describing their classes as "vocational." Only 2% of teachers classified their classes in this way, however, suggesting that the existence of vocational classes is not indicative that the school has a tracking policy.

TABLE 2
OLS Estimates of the Effects of Ability Tracking on Mathematics Test Score Gains

Variable	Specification including track placement	Bottom third of 8th grade test distribution	Middle third of 8th grade test distribution	Top third of 8th grade test distribution
Placement in bottom track	−6.000			
	(1.112)			
Placement in middle track	−0.633			
	(0.557)			
Placement in top track	2.226			
	(0.570)			
School ability tracks in mathematics (principal report)		−0.404	−0.063	−0.185
		(0.338)	(0.395)	(0.322)
8th grade math test score	−0.323	0.201	−0.048	−0.249
	(0.027)	(0.041)	(0.040)	(0.018)
White	−0.423	1.180	−1.753	−1.489
	(0.545)	(0.617)	(0.634)	(0.378)
Black	−2.178	0.077	−1.963	−2.488
	(0.958)	(0.616)	(0.836)	(0.794)
Hispanic	−1.238	0.523	−1.774	−2.244
	(0.967)	(0.576)	(0.693)	(0.640)
Income under $15,000	0.183	−0.951	0.679	0.369
	(0.740)	(0.469)	(0.609)	(0.610)
Income $15–35,000	−0.242	−0.213	1.456	0.256
	(0.553)	(0.451)	(0.593)	(0.483)
Income $35–50,000	0.060	0.457	2.129	0.402
	(0.553)	(0.538)	(0.587)	(0.482)
Income $50–100,000	−0.185	−0.360	1.420	0.180
	(0.622)	(0.672)	(0.662)	(0.499)
Income over $100,000	−0.005	0.746	1.745	−0.010
	(0.736)	(0.932)	(0.907)	(0.544)
Parent high school graduate	0.500	0.988	0.836	1.439
	(1.299)	(0.313)	(0.530)	(0.836)
Parent college graduate	1.154	2.126	0.650	0.996
	(0.371)	(0.476)	(0.349)	(0.274)
Days in school year	0.000	−0.075	−0.056	−0.035
	(0.063)	(0.034)	(0.053)	(0.037)
Percent white in school	0.005	0.001	−0.006	0.000
	(0.010)	(0.006)	(0.007)	(0.007)
Percent free lunch in school	−0.020	0.016	−0.025	−0.016
	(0.014)	(0.008)	(0.012)	(0.010)
Highest teacher salary (1000s)	−0.008	−0.024	−0.012	0.025
	(0.030)	(0.022)	(0.023)	(0.020)
Tenth grade enrollment (1000s)	1.598	1.191	2.221	0.296
	(0.935)	(0.922)	(1.013)	(0.668)
Student–teacher ratio	−0.001	0.003	0.003	−0.000
	(0.001)	(0.004)	(0.001)	(0.001)

Note: Standard errors (in parentheses) are corrected to account for within school clustering of errors. All models also include a constant term, region dummies, and central city/suburb dummies, the coefficients of which are omitted due to space constraints. Parent dropout and missing income are reference categories.

zero to three. Our third definition counts a school as "tracked" if at least one teacher reports that his class has "high-achieving" students. The last three measures are similar to the first three except that we replace "high-achieving" with "advanced," "average-achieving" with "academic," and "low-achieving" with "general." In all six cases, we add as an additional covariate the number of unique classes observed in the school because the likelihood of observing a particular type of class is higher as we observe more classes. About two-thirds of schools have at least one teacher reporting that her class is "high-achieving," and about three-quarters of schools have at least one teacher reporting that her class is "advanced." Eighty percent of schools with teachers reporting that their class is "high-achieving" also have a teacher reporting that her class is "low-achieving." Ninety-six percent of schools with teachers reporting that their class is "advanced" also have a teacher reporting that her class is "general." Depending on measure of tracking used, between 64 and 78% of schools are reported to be tracked.

The estimated treatment effects are reported in Table 3. Across the specifications, there is no evidence that tracking harms low-ability children. Compared to the estimates in the first column of Table 2, the coefficient estimates for this group are all small in magnitude, ranging from -0.40 to $+0.06$, and none of them is statistically different from zero. Some of the alternative definitions do produce more economically (and nearly statistically) significant estimates for the high- and medium-ability groups, but the estimates are still substantially smaller than the estimated coefficients on the track placement variables. Taken together, the estimates in Table 3 reiterate the conclusions that we drew in Table 2: much of the estimated effect that has been associated with tracking disappears when we control for unobservable factors correlated with track placement. Once the endogeneity of track placement is addressed in this manner there is no evidence that low-ability children are hurt by tracking.

B. *Assessing the Endogeneity of Tracking Programs*

The results presented in Tables 2 and 3 provide no evidence that tracking programs are harmful to low-ability students, but, like other studies, the estimates are based on a model in which the school's tracking status is assumed to be exogenous with respect to student test scores. Differences in student composition across schools with and without tracking programs, shown in Table 1, raise the possibility that tracking influences school selection.

We investigated this possibility further using a panel of schools that were in all three waves (1987–88, 1990–91, and 1993–94) of the Schools and Staffing Survey. We estimated the correlation between a school's "tracking" status[18] and the fraction of students who are free lunch-eligible, by regressing changes in the

[18]The SASS does not contain explicit information on tracking programs, so we created proxies for tracking using information on whether the school maintains gifted and/or remedial programs.

TABLE 3
Estimated Treatment Effects of Ability Tracking, Using Alternative Measures of School Tracking Status

School is ability tracked if	Fraction of schools tracking using this measure	Bottom third of 8th grade test distribution	Middle third of 8th grade test distribution	Top third of 8th grade test distribution
Principal reports school tracked in mathematics	77.5%	−0.404 (0.338)	−0.063 (0.395)	−0.185 (0.322)
Teachers report both "high-achieving" and "low-achieving" classes	64.6	0.062 (0.489)	0.520 (0.555)	0.751 (0.531)
Number of tracks observed (among "high," "average," and "low-achieving" classes)	49.3% 3 tracks observed 23.1% 2 tracks observed	−0.089 (0.169)	0.007 (0.211)	0.031 (0.169)
Any teacher reports a "high-achieving" class	67.2	−0.091 (0.323)	0.676 (0.399)	0.483 (0.384)
Teachers report both "advanced" and "general" classes	70.4	−0.271 (0.316)	0.280 (0.400)	−0.314 (0.301)
Number of tracks observed (among "advanced," "academic," and "general" classes)	69.0% 3 tracks observed 27.2% 2 tracks observed	−0.274 (0.241)	0.267 (0.318)	−0.314 (0.230)
Any teacher reports a "advanced" class	73.1	−0.099 (0.318)	0.275 (0.413)	−0.398 (0.338)

Note: Standard errors (in parentheses) are corrected to account for within school clustering of errors. All models also include all covariates from Table 2, an indicator of the number of unique classes observed in the school (for teacher reported measures of tracking), and a constant term. The sample size using the principal-reported measure of tracking is 7676, while the sample size using the teacher-reported measures of tracking is 5948.

fraction of free lunch-eligible students on changes in the school's tracking status. This allowed us to control for school fixed effects. Our regression equation also included a school trend. We found that the addition of a tracking program is associated with an 8% reduction in the number of students who are free lunch-eligible.[19] This finding lends weight to the possibility that estimates of the return to tracking based on Eq. (2) may be biased by differential selection into tracked vs untracked schools.

We attempt to address the potential endogeneity of a school's tracking status by conducting a two stage least squares analysis that uses as instruments two and three way interactions between three variables: the number of academic courses required for state graduation, the number of schools in the county (measured in 1987), and the fraction of voters in the county who voted for President Reagan

[19] More details on these analyses are available from the authors.

in the 1984 election.[20,21] While the rationale for using the interactions between these variables as instruments may not be immediately transparent, we argue that these interactions are plausible instruments.[22] When state academic requirements are minimal high school students of different ability levels are likely to select into different classes—for example, if 10th grade math is not required then students taking 10th grade math will be disproportionately of high ability. As the number of academic classes required for high school graduation increases, so does the probability that classrooms, if untracked, will be heterogeneous. We hypothesize that when this occurs parents of high-ability children will be more likely to pursue schooling options that formally separate their children from low-ability peers. As a result, when academic requirements increase schools may have more of an incentive to track, and this incentive should be larger when parents have more schools from which to choose. Presumably, the decision to track will be further influenced by parental tastes: graduation requirements should have a larger influence on tracking policies in communities where parents have less taste for mixing. Finally, the impact of residents' tastes should be larger when schools face more competition. Our proxy for parental tastes is the percent of county voters who voted for President Reagan in 1984.[23]

Table 4 reports our second stage coefficient estimates, together with the p-value of the instruments from the first stage regressions, the first stage partial R^2, and the p-value of the Hausman exogeneity test. The first stage partial R^2 runs from 0.012 to 0.050, and the p-values of the marginal explanatory power of the instruments in the first stage is always less than 0.0004. Because the instruments do not explain a great deal of the variation in tracking status they are not ideal; we expect our IV strategy to yield relatively imprecise estimates.

[20]These variables come from the Digest of Education Statistics, the City and County Data book, and the Common Core of Data.

[21]Note that the interaction between the number of schools in the county in 1987 and the fraction of voters in the county who voted for President Reagan in 1984 is a variable that does not vary over time. Thus, it is subsumed in the school-specific fixed effect.

[22]We do *not* argue that the number of academic courses required for state graduation, the number of schools in the county, or the fraction of voters in the county who voted for President Reagan would individually make good instruments. State graduation requirements may be positively correlated with student achievement, for example. We cannot think of a story in which the impact of state requirements would be larger in countries where there is more competition, however, unless schools in those counties respond differently to the change in state requirements through the adoption of a policy such as tracking. We include all three of the noninteracted components of our instrumental variables in both the first and second stage of our IV procedure.

[23]We have no *a priori* expectations about whether Republican voters are more or less responsive to changes in state requirements, but we do expect that tastes should influence the degree to which changes in state requirements will affect the tendency to track. The first stage coefficient estimate on this variable is 0.0016 but is not statistically significant when the three way interaction is also included in the model. However, when the three way interaction is excluded from the model, the coefficient estimate is 0.004 and is modestly significant at the 11% level.

TABLE 4
Instrumental Variables Estimates of the Effects of Ability Tracking on Mathematics Performance

Tracking measure:	Principal-reported measure of tracking	Number (advanced, academic, general) observed in school	Advanced class observed in school	Both advanced and general classes observed in school
Bottom third of initial score distribution				
Treatment effect	6.844	3.362	4.582	3.781
	(2.710)	(2.155)	(2.750)	(2.624)
p-Value of Hausman exogeneity test	0.468	0.188	0.112	0.121
Partial R^2 in first stage	0.015	0.016	0.012	0.013
First stage p-value	0.000	0.000	0.000	0.000
Middle third of initial score distribution				
Treatment effect	−3.439	1.456	2.369	2.302
	(2.147)	(1.693)	(2.351)	(2.291)
p-Value of Hausman exogeneity test	0.809	0.381	0.414	0.420
Partial R^2 in first stage	0.019	0.019	0.013	0.014
First stage p-value	0.000	0.000	0.000	0.000
Top third of initial score distribution				
Treatment effect	−3.819	−1.425	−1.835	−1.636
	(3.075)	(0.956)	(1.883)	(1.180)
p-value of Hausman exogeneity test	0.575	0.750	0.488	0.661
Partial R^2 in first stage	0.019	0.056	0.026	0.050
First stage p-value	0.000	0.000	0.000	0.000

Note: Standard errors (in parentheses) are corrected to account for within school clustering of errors. All models also include all covariates from Table 2, an indicator of the number of unique classes observed in the school (for teacher-report-based measures of tracking).

We are hopeful, however, that their explanatory power is sufficiently strong so as to avoid the type of 2SLS inconsistency discussed in Bound *et al.* [6].

We find that the effect of tracking is, if anything, *positive* for members of the low-ability group, regardless of whether we employ the principal-reported measure of tracking or one of our teacher-reported measures. The estimates for the middle group are not distinguishable from zero, and the estimates for the high-ability group are actually negative (though usually insignificant) when we apply IV. Using the principal-reported measure of tracking, the coefficient estimate for the low-ability group is 6.8 points, an effect equivalent to moving a student from the 20th percentile of the 8th grade ability distribution to about the 40th percentile of the 10th grade ability distribution. Although this estimate is not precise, the standard error estimate is small enough to reject (at the 2% level) the null hypothesis that tracking has no effect, and the confidence interval around the estimated effect for low-achieving students, does not include the estimates reported in Tables 2 and 3. In other words, our IV estimate is both statistically different from zero and larger in magnitude than the estimates based on regressions that ignore the impact of tracking policies on school choice. Using constructions of tracking measures based on teacher reports yields similar point estimates, but they are less precise: one can reject the null that the effect of tracking is zero at the 9 to 15% level, depending on specification.

V. CONCLUSIONS

The current trend away from ability tracking results largely from the perception that tracking is harmful to low-ability students. Previous studies have concluded that high-ability students gain from tracking at the expense of their less-able schoolmates, but those studies have not adequately addressed the possibility that track placement and tracking programs may be endogenous with respect to student outcomes. Research on tracking effects has also suffered because of ambiguity over the definition of tracking.

We estimate the effects of tracking on students of different ability levels by comparing achievement gains across similar students attending tracked vs untracked schools. By dividing the sample into groups based on their 8th grade test scores instead of using their track assignment we are able to identify the effect of tracking on different ability groups using only variation across types of schools (those with tracking vs those that do not track) rather than using variation across *both* school and student types. This empirical strategy produces results that are strikingly different from those produced by comparing individuals schooled in different tracks—in particular, our estimates provide no evidence that tracking harms low-ability students. This finding is robust to a number of different tracking definitions.

Our comparison of achievement gains across tracked vs untracked students may still produce biased estimates, however, if the existence of tracking programs affects school choice. We therefore supplement our results with a two stage least squares analysis. The results from this exercise suggest that, if anything, low-ability students may actually experience larger test score gains when they are schooled in tracked settings. We can find no evidence that detracking America's schools, as is currently in vogue, will improve outcomes among disadvantaged students.

REFERENCES

1. K. L. Alexander and E. L. McDill, Selection and allocation within schools: Some causes and consequences of curriculum placement, *American Sociological Review* **41**, 963–980 (1976).
2. L. M. Argys, D. I. Rees, and D. J. Brewer, Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management* **15**, 623–645 (1996).
3. J. R. Betts and J. L. Shkolnik, The effects of ability grouping on student math achievement and resource allocation in secondary schools, *Economics of Education Review* **19**, 1–15 (2000).
4. J. Bishop, Is the test score decline responsible for the productivity growth decline? *American Economic Review* **79**, 178–197 (1994).
5. J. Bound and G. Johnson, Changes in the structure of wages in the 1980's: An evaluation of alternative explanations, *American Economic Review* **82**, 371–392 (1992).
6. J. Bound, D. A. Jaeger, and R. M. Baker, Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association* **90**, 443–450 (1995).
7. J. H. Braddock and R. E. Slavin, Why ability grouping must end: Achieving excellence and equity in American education, *The Journal of Intergroup Relations* **20**, 51–64 (1993).
8. D. Epple, E. Newlon, and R. Romano, Ability tracking, school competition, and the distribution of educational benefits, *Journal of Public Economics*, in press.
9. D. N. Figlio and M. E. Page, "School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Equality?" NBER Working Paper 8055, 2000.
10. A. Gamoran, The stratification of high school learning opportunities, *Sociology of Education* **60**, 135–155 (1987).
11. A. Gamoran, Measuring curriculum differentiation, *American Journal of Education* **97**, 129–143 (1989).
12. T. B. Hoffer, Middle school ability grouping and student achievement in science and mathematics, *Educational Evaluation and Policy Analysis* **14**, 205–227 (1992).
13. C. Juhn, K. M. Murphy, and B. Pierce, Wage inequality and the rise in returns to skill, *Journal of Political Economy* **101**, 410–442 (1993).
14. A. C. Kerckhoff, Effects of ability grouping in British secondary schools, *American Sociological Review* **51**, 842–858 (1986).
15. R. J. Murnane, J. B. Willett, and F. Levy, The growing importance of cognitive skills in wage determination, *Review of Economics and Statistics* **77**, 251–266 (1995).
16. J. Oakes, "Keeping Track: How Schools Structure Inequality," Yale Univ. Press, New Haven, CT, 1985.
17. J. Oakes, Can tracking research inform practice? Technical, normative and political considerations, *Educational Researchers* **21**, 12–21 (1992).
18. D. I. Rees, L. M. Argys, and D. J. Brewer, How should we measure the effect of ability grouping on student performance? *Economics of Education Review* **19**, 16–20 (2000).
19. J. E. Rosenbaum, Track misperceptions and frustrated college plans: An analysis of the effects of tracks and track perceptions in the National Longitudinal Survey, *Sociology of Education* **53**, 74–88 (1980).
20. J. E. Rosenbaum, Social implications of educational grouping, *Review of Research in Education* **8**, 361–401 (1980).
21. R. E. Slavin, Achievement effects of ability grouping in secondary schools: A best evidence synthesis, *Review of Educational Research* **60**, 471–499 (1990).
22. C. M. Tiebout, A pure theory of local expenditures, *Journal of Political Economy* 416–424 (1956).
23. B. E. Vanfossen, J. D. Jones, and J. Z. Spade, Curriculum tracking and status maintenance, *Sociology of Education* **60**, 104–122 (1987).
24. A. Wheelock, "Crossing the Tracks: How 'Untracking' Can Save America's Schools," The New Press, New York, 1992.