Pergamon

Economics of Education Review 19 (2000) 21–26

Economics of Education Review

www.elsevier.com/locate/econedurev

Key difficulties in identifying the effects of ability grouping on student achievement

Julian R. Betts [a,*], Jamie L. Shkolnik [b]

[a] *Department of Economics, University of California, San Diego, La Jolla, CA 92093-0508, USA*
[b] *National Opinion Research Center, 1350 Connecticut Avenue, Suite 500, Washington, DC 20036, USA*

Received 25 September 1998; accepted 29 September 1998

Abstract

The paper presents empirical evidence that earlier research may have overstated the impact of ability grouping and tracking on inequality in student achievement. We list six key difficulties facing research on the effects of grouping on student achievement. Each of these difficulties offers opportunities for further research and for collection of more appropriate data sets. Strong conclusions as to the differential effect of ability grouping on high-achieving and low-achieving students are probably not yet warranted. [JEL I21] © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Educational economics; Human capital; Educational finance; Resource allocation

1. Introduction

Ability grouping is a widespread practice in American schools. For well over a decade, researchers have investigated how the grouping of students into classrooms by achievement levels (ability grouping) has affected the average level and the dispersion of achievement. The purpose of this paper is to enumerate some of the major difficulties in distinguishing the impact of ability grouping on student achievement. Section 2 discusses the comments made by Rees, Brewer and Argys (1999) (henceforth RB and A) and presents new evidence that omitted ability bias has likely led to an overstatement of the differential effects of grouping in the previous literature. Section 3 lists six key difficulties that confront all researchers in this area. We conclude that based on the existing evidence it is difficult to make a clear policy prescription as to whether "detracking" America's schools will lead to gains or losses for all, some, or even any students.

2. Formal versus informal grouping, and the problem of omitted ability bias

We begin by summarizing and rebutting the criticisms of our work made in RB and A. We then summarize our criticism of the earlier literature, including papers such as Argys, Rees and Brewer (1996), and present new evidence in favor of our interpretation.

Our data set, the Longitudinal Study of American Youth (LSAY), asks principals whether their schools use "ability grouping or tracking (other than AP courses)" in their math classes. Using this variable, we divide our sample into students at schools that use grouping (73% of student observations) and those at schools that do not (27%). Our data set also includes questions directed to each student's teacher, who reports the average ability of the class on a 1–5 scale. Accordingly, we test whether students in classes of ability level "*n*" in a school with grouping learn at the same rate as students in classes of ability level "*n*" at schools without grouping. We find little effect of formal ability grouping.

RB and A worry that many schools in our sample informally group students, even if the principal claims that no ability grouping takes place. In other words, principals cannot be trusted to provide reliable information

* Corresponding author. Tel.: +1-619-534-3369; fax: +1-619-534-7040; e-mail: jbetts@ucsd.edu

PII: S0272-7757(99)00022-9

about how their schools group students. If this were so, then our comparison of students in high-ability classes in schools with tracking to students in high-ability classes in schools that do not track is not useful—it's simply, as the authors claim, comparing "apples to apples".

We readily concede that one possible interpretation of our results is that we are testing the effects of "formal ability grouping," in which schools admit to grouping, to the effects of "informal ability grouping". In these latter schools, perhaps schools claim not to group but actually do, or students effectively group themselves based on the level of classes that they choose. We make this point repeatedly throughout the abstract and text.

But a second possibility is that the finding of many previous researchers that tracking aggravates the gap in achievement between top and bottom students is overstated, due to omitted ability bias in the test score equation. Our technique of comparing "apples to apples" may greatly reduce this type of bias because it avoids comparing "apples to oranges".

2.1. Comparing apples to apples or apples to oranges?

We compare students in schools that track (according to the principal) to students at schools that do not. We control for "environmental" factors that could affect student achievement, such as family background. In addition, by comparing students in tracked schools who are in math classes of ability "*n*" to students in untracked schools who are in classes of ability "*n*", we compare apples to apples. For each type of class, we derive the effect of tracking by comparing the "treatment" group (that was in a school with tracking) to the control group.

Some earlier literature on ability grouping runs the risk of comparing apples to oranges. For instance, Argys, Rees and Brewer (1996) compare students in "above average" classes to students in "heterogeneous" classes. Hoffer (1992) uses LSAY data to compare students in high, middle and low grouped classes to a control group of *all* students in schools that, according to his metric, do not use ability grouping. This can create omitted ability bias when this highly heterogeneous control group is really of a quite different level of initial achievement than students in the various grouped classes.

This approach is likely to lead to a systematic upward bias in the estimated effects of placing students in above average classes, and a downward bias in the case of below average classes. As we argue in the introduction to Betts and Shkolnik (1999), because test scores measure achievement with error, a lagged test score in the test score equation will not adequately control for initial achievement. Therefore the ability level of the class, when included as a regressor, will be biased upward because it is positively correlated with the student's own imperfectly observed initial level of achievement. This leads to an overstatement of the *differential* effects of ability grouping on student learning in papers that use the "apples versus oranges" approach. (See equation (2a) in our companion paper.)

Both Argys, Rees and Brewer (1996) and Hoffer (1992) run separate regressions for various ability groups. Technically, this changes the problem from one of omitted ability bias in a full-sample regression to one of selectivity bias in the regressions on subsamples. That is, the expected value of the error term in each test score model is unlikely to be zero, if there is any correlation between the error term in that equation and the error term in the equation that determines how each student was assigned to an ability group. The authors attempt to control for this problem using corrections for selectivity bias, but their corrections will be imperfect unless they can perfectly capture the actual class assignments of each student.

Which of these two problems, confusing non-grouped schools for schools that group informally, or upward bias due to comparing apples to oranges, is a greater problem in the literature? We can provide three indirect pieces of evidence that the latter is a greater source of bias.

2.1.1. Mislabeling grouped schools as ungrouped does not explain our results

First, in our analysis, if it's true that we are improperly classifying many schools as non-grouping when in fact they do use ability grouping, then we should be unable to replicate Hoffer's results closely, even though we both use LSAY data. Hoffer's indicator for grouping is based on teacher interviews, school documents, and when necessary, phone calls to the schools.[1] He estimates that 85% of students in Grade 7 math classes are in grouped classes. (Hoffer restricts his analysis to middle school test score data.) For our sample, we estimate that 73% of math students are in grouped classes. Could it be that many of our non-grouped schools, as identified by principals, are in fact grouped, and that this explains why we find little or no effect of grouping?

The answer to this is clearly no. If we had such substantial measurement error, then a replication of Hoffer's method using our grouping and class ability measures should produce quite different results from his. But we can replicate his approach by ignoring the information provided by teachers on the ability level of classes in non-grouped schools, and instead combining *all* students in non-grouped schools into a single control group. We would expect to find little or no effect of grouping if we had misallocated students between grouped and ungrouped schools. In reality, by aggregating all "non-grouped" students and comparing them to our five levels

[1] This variable is not available in the public-use data set.

of "grouped" students, we can replicate his result almost exactly.

Hoffer reports a predicted gap in test scores between the top and bottom math classes of 6.1 points. In our model that aggregates all students in non-grouped schools together into the control group, we find (Table 2, #2), the predicted test score gap is 7.4 points.[2] As we show in Figure 1 of our companion paper, this predicted gap in learning is unreasonably high given the actual distribution of test scores over time in the LSAY.

In our view, questions about whether 15% or 27% of students are in untracked math classes are of second-order importance.[3] The fact of the matter is that we can closely replicate Hoffer's results when we use his method, even though we use a different indicator of whether the school tracks. This suggests that it is the model specification, rather than the way in which we measure whether schools use ability grouping, that leads to the differences in our results.

It seems clear that it is not the measure of which schools track that account for differences between our results and Hoffer's. Our results differ because we use class ability information provided by teachers in schools that do not group in order to find an appropriate comparison group for students in a class of a given ability in grouped schools. Hoffer, using earlier waves of the same data set, instead combines all students in non-grouped schools into the comparison group, ignoring information on the ability level of each student's class in the school without formal grouping. This leads to the risk of comparing apples to oranges.

2.1.2. Evidence of selectivity bias in Argys, Rees and Brewer (1996)

The regression results provided by Argys, Rees and Brewer (1996) have every indication of suffering from bias due to the apple–oranges problem.[4]

Consider first the actual test score gains between Grades 8 and 10 for students placed in various ability groups, as reported by these authors, and shown in the top panel of Table 1. Students in below average classes gained 5.7 points over this period, compared to a gain of 11.1 for students in above average classes. On the face of it, then, ability grouping has caused an 11.1 − 5.7 = 5.4 point widening in the achievement gap between these two types of students.

As we have argued, part of this widening gap reflects the fact that abler students learn more quickly. Thus, the 5.4 point gap is likely an upper bound on the effect of tracking. Once we control for selectivity bias, we would expect the part of the widening gap that is caused by ability grouping to be less than 5.4 points. But as shown in the bottom panel of our Table 1, Argys, Rees and Brewer (1996) predict that after controlling for selectivity, placing two identical students into different eighth grade classes, one "above average" and the other "below average," will lead to a 10.83 point gap between the two students by tenth grade.

After controlling for selection into each type of class, the effect of grouping appears to become stronger, not weaker as expected. How could this be?

Correcting the apple–oranges problem requires good instruments with which to control for the selectivity bias in each of their four test score equations (high ability, average ability, low ability and heterogeneous). The coefficients on the inverse Mills selection terms in the test score equations should meet two criteria. First, they should be highly significant, reflecting the fact that above average students select into above average classes and so on. Second, the coefficients on the selectivity terms should be positive for the high ability class (since students placed in high ability classes are exceptionally able, and lagged test scores can capture this only imperfectly), and negative for the low ability class.

Remarkably, neither of these signs of a proper selectivity correction is met in Argys, Rees and Brewer (1996). The *t*-statistics on the selection term for the test score model are insignificant for "above average" and "average" classes: 1.3 for "above average" classes, 1.0

Table 1
Actual and predicted achievement gains between grades 8 and 10 by track, derived from Tables 1A and 4A from Argys, Rees and Brewer (1996)

Actual test score gains between grades 8 and 10, by track				
Above	Average	Below	Heterog.	Actual gain in gap (Above–below)
11.1	10.3	5.7	7.6	5.4
Predicted scores by grade 10, by track, for initially identical but randomly assigned students				
Above	Average	Below	Heterog.	Predicted gap (Above–below)
68.90	65.19	58.07	63.08	10.83
(0.41)	(0.35)	(0.92)	(0.70)	

Standard errors in parentheses.

[2] One reason why our predicted gap might be slightly larger is that we categorize students into five ability groups, while Hoffer uses three.

[3] We note that an independent data set, the Schools and Staffing Survey, suggests that "about 73%" of high schools tracked their students in 1990. See Figlio and Page (1998). This exactly matches our own measure from the LSAY.

[4] Technically, their approach does not suffer from omitted ability bias but rather from potential selectivity bias. Instead of running a test score equation that includes all students, they run separate test score models for students in each group, after adding an inverse Mills term to control for the fact that, for instance, students in "above average" classes are better-than-average students.

for "average" classes, 2.2 for "below average" classes, and 2.1 for "heterogeneous" classes. The low *t*-statistics indicate that the instruments used to control for selection into each ability group have done little to control for selectivity bias. In addition, the signs of the coefficients on the selection terms are the *opposite* of what we would expect. The coefficient is − 1.83 for the "above average" class, and 4.70 for the "below average" class, suggesting that their selectivity correction fails to correct for the fact that a student's initial achievement is positively correlated with his or her initial placement. Hence, much of the observed gains attributed to being in an "above average" class merely reflect the non-randomness of a subsample composed of highly able students.

The fact that the signs are "wrong" on the selectivity correction suggests that the selectivity correction will *aggravate* the omitted variable bias. This may be why the predicted gain from ability grouping is larger than what appears in the raw data, rather than smaller as expected, after the selectivity correction is performed.[5]

2.1.3. Direct evidence that omitted ability bias leads to overstatement of the differential effect of grouping

Results in Hoffer (1992) provide direct evidence in favor of our hypothesis that part of the gain attributed to being in a "high ability" class simply reflects the student's own imperfectly measured initial achievement. Hoffer, in his Table 3, models Grade 9 test scores in math and science twice, first including Grade 8 test scores and then including *both* Grade 8 and Grade 7 test scores as controls for initial achievement. His results (reproduced in our Table 2) indicate that the addition of further controls for initial achievement (7th grade test scores) diminishes the impact of placement in the high ability level class to statistical insignificance. Just as importantly, the predicted gap in student performance between the high ability group and the low ability group falls by over a third. These findings provide direct evidence that the estimated impact of ability grouping on inequality is overstated, due to positive correlation between the ability level of the class and the part of the student's initial test score that is in the error term. In contrast our approach compares students in high-ability classes in tracked and non-tracked schools to difference out this bias.

[5] Argys, Rees and Brewer (1996) also examine the effects of placing students into four different types of curricula. Similar criticisms apply to those models. In particular, the inverse Mills selection terms are insignificant in all regressions, and the predicted widening in the achievement gap across tracks is much larger than observed in the actual data. For example, the raw data suggest that between Grade 8 and 10 the test score gap between those in the honors track and the general track widens by 3.2 points, while the predicted widening in the gap—10.39 points—is over three times as large.

Table 2
Reproduction of Table 3 from Hoffer (1992), modeling grade 9 science and math test scores as functions of group placement and lagged test score(s)

Test score type	Regressor	Test score(s) included as regressors	
		Grade 8 only	Grades 7 and 8
Science	High group	2.263*	0.823
		(0.550)	(0.526)
	Middle group	−0.071	−0.247
		(0.401)	(0.377)
	Low group	−1.970*	−1.657*
		(0.890)	(0.837)
	Predicted gap (High–low)	4.233	2.480
Math	High group	2.627*	1.431
		(1.216)	(1.150)
	Middle group	−0.462	−0.239
		(1.154)	(1.089)
	Low group	−3.492*	−2.569*
		(1.237)	(1.168)
	Predicted gap (High–low)	6.119	4.000

Standard errors in parentheses. Other regressors included are SES, gender, race, Hispanic ethnicity, school size and school-average SES. * $p < 0.05$.

3. A list of challenges to research ability grouping and tracking

We believe that the research on ability grouping to date has not yet produced clear directions for policymakers on the extent to which grouping aggravates inequality in student achievement. Below we list six problems that all researchers in this area face.

1. Group placement may be a proxy for imperfectly measured ability, leading to overstatements of the impact of ability grouping. (The apples vs. oranges problem). Support for the existence of this problem is provided above and in our companion piece in this issue.
2. Schools that do not use ability grouping officially may nonetheless group students informally. This is the critique offered by RB and A, and discussed and partially rebutted in the previous section.
3. The meaning of "heterogeneous" ability classes is ambiguous. In the National Educational Longitudinal Study (NELS), used by Argys, Rees and Brewer (1996), teachers are asked to identify a class as "above average", "average", "below average", or "heterogeneous". It is far from clear what either the range of ability or the mean ability will be in "heterogeneous" classes. In one school, all classes may be

truly heterogeneous, and therefore of "average ability". But in a second school, students with lower levels of achievement may be grouped together in "below average" classes. Consequently, the classes composed of the remaining students, whom the teacher is likely to view as being "heterogeneously" grouped, are by definition above average. Thus, a survey question of this form can tell us little about peer group effects simply because "heterogeneous" could mean so many different things. (Indeed, in the results of Argys, Rees and Brewer, 1996, in the test score equation for students in "heterogeneous" classes, the coefficient on the selectivity correction is statistically significant, suggesting that these students do not represent a random sample.)

4. Survey instruments typically fail to differentiate between ability grouping and tracking. In the major data sets brought to bear on ability grouping, perhaps too few distinctions have been made between grouping students by ability, on the one hand, and channeling students into different tracks of curriculum on the other. These two practices are no doubt often intricately intertwined, but they should be differentiated. To their credit, Argys, Rees and Brewer (1996) compare not only "above average" to "heterogeneous" classes etc., but they also compare outcomes in honors, academic, general and vocational tracks. The second set of comparisons is much closer to curriculum tracking, while the former is in all likelihood a combination of ability grouping and tracking. In the LSAY data we used, principals are asked to indicate whether "ability grouping or tracking" is used, so that we cannot disentangle the effects. Given these difficulties, the optimal data set would include all students in the classroom so that researchers could measure the actual mean ability as well as the dispersion of abilities in the class. Details on curriculum would also be useful. On this point, we appear to be in strong agreement with RB and A. Unfortunately, with the nationally representative data sets currently in use such as NELS and the LSAY, it is not possible to identify every student in the classroom.
5. We need to know much more about how tracked schools allocate resources. A welcome contribution by RB and A is a set of tables showing mean teacher traits and class size by class ability. Their results are highly consistent with our own results for math classes. Our companion paper shows that in schools with (formal) ability grouping, students in lower-ability classes have smaller class sizes and less educated and less experienced teachers. RB and A's tabulations from NELS agree closely with this conclusion. Our results go beyond this though, and show that the "smaller class size" result holds only within schools that (formally) use ability grouping, and that the "less teacher experience" result is reversed in schools that do not (formally) track. It would be worthwhile to investigate these effects further. Certainly, the results in our work and those of RB and A raise some important questions about the traditional view that tracking must always take resources away from the students with the lowest achievement within each school. Reality appears to be far more complex than any of us imagined.
6. We know little about the extent of ability grouping and curriculum tracking within the classroom. The research cited by RB and A and ourselves typically measures ability grouping at the level of either the school or the classroom. The literature neglects the possibility that within a classroom teachers group students by ability. Similarly, to challenge the top students, teachers might need to teach an extended curriculum.[6] To the extent that schools use grouping *within* the classroom, it suggests that existing studies of ability grouping at the classroom level will not fully measure the impacts of grouping. If teachers are more apt to use grouping within heterogeneous classes, it will lead to an overstatement of how much the high-achieving students' scores will drop if schools are detracked. It also raises serious questions about whether national calls for "de-tracking" can ever succeed fully.

4. Concluding comments

Based on the evidence presented in Section 2, we believe that the estimated effects of ability grouping on inequality in test scores in some of the earlier literature is substantially overstated. At the same time, the six problems listed in Section 3 raise serious questions about whether our earlier research, or any of the earlier research, has completely captured the true effects of ability grouping and tracking. Definitional problems (ability grouping versus curriculum tracking) are compounded by a series of measurement issues. No data set adequately deals with all of these problems. An optimal data set would include data on all students at a school, allowing one to measure both mean achievement and the dispersion of achievement within each classroom. It would also include information on how teachers grouped students within the classroom, and detailed information on curriculum differences between and *within* classrooms. In the absence of such a data set, perhaps our wisest conclusion is that we still do not fully understand what "detracking" schools would do.

[6] One of our colleagues reports that in her son's elementary school, children commonly speak of the "smart table" of students and the "stupid table" within classes. Children can be harsh, but also perceptive.

Acknowledgements

This research was supported by a Dissertation Grant to Shkolnik from the American Educational Research Association, which receives funds for its "AERA Grants Program" from the National Science Foundation and the National Center for Education Statistics (US Department of Education) under NSF Grant #RED-9452861. Betts thanks the Social Sciences and Humanities Research Council of Canada for research support, and the Public Policy Institute of California for research support while he served there as a Visiting Fellow in 1998. We thank Heather Rose for useful suggestions. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies.

References

Argys, L. M., Rees, D. I., & Brewer, D. J. (1996). Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management*, *15*(4), 623–645.

Betts, J. R., & Shkolnik, J. L. (1999). The effects of ability grouping on student math achievement and resource allocation in secondary schools. *Economics of Education Review*, *19*(1), 1–15.

Figlio, D. N. and Page, M. E. (1998). Ability Tracking, School Choice and Social Mobility: Does Separation Increase Equality? University of Florida manuscript.

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis*, *14*(3), 205–227.

Rees, D. I., Brewer, D. J., & Argys, L. M. (1999). How should we measure the effects of ability grouping on student performance? *Economics of Education Review*, *19*(1), 17–20.