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.