4-2-2013

NEPC Review: Does Sorting Students Improve Scores? An Analysis of Class Composition

Carol C. Burris
burriscarol@gmail.com

Katherine E. Allison
University of Colorado Boulder

NEPC
NATIONAL EDUCATION POLICY CENTER

Review of *Does Sorting Students Improve Test Scores?*

Reviewed By

Carol Corbett Burris, Rockville Centre School District
Katherine E. Allison, University of Colorado Boulder

April 2013

Summary of Review

This National Bureau of Economic Research working paper purports to examine the extent and effects of sorting students into classrooms by test scores. It then claims to explore the effect of sorting on overall student achievement as well as on low achievers, high achievers, gifted, special education and Limited English Proficient students. The paper uses standardized Texas state test scores as the measure of learning growth. Based on a comparison between third- and fourth-grade scores, the paper concludes that sorting students by scores is associated with significant learning gains for both lower and higher achievers. It does not, however, find similar effects for the sub-groups. The paper is limited by several important methodological issues. First, it simply assumes, based on test score distributions, that the schools tracked students between classes—and this assumption is highly questionable. Second, it provides no criteria by which students were classified as high or low achievers. Finally, it measures only *relative* standing of students on two proficiency tests given in different years. It does not measure growth. Because of these and other weaknesses, this paper should not be used to inform policy regarding tracking or grouping practices.

Kevin Welner
Project Director

William Mathis
Managing Director

Erik Gunn
Managing Editor

National Education Policy Center

School of Education, University of Colorado
Boulder, CO 80309-0249
Telephone: (802) 383-0058

Email: NEPC@colorado.edu
http://nepc.colorado.edu

Publishing Director: Alex Molnar

This is one of a series of Think Twice think tank reviews made possible in part by funding from the Great Lakes Center for Education Research and Practice. It is also available at http://greatlakescenter.org.

Review of *Does Sorting Students Improve Test Scores?*

Carol Corbett Burris, Rockville Centre School District
Katherine E. Allison, University of Colorado Boulder

I. Introduction

The National Bureau of Economic Research (NBER) publishes "working papers" that allow economists to circulate non-peer-reviewed drafts for discussion and comment purposes. Advocates of policy positions can circulate these papers before they have been critiqued, however. A recent example is *Does sorting students improve scores? An analysis of class composition*, by Courtney A. Collins and Li Gan, which explores the achievement effects of sorting students by test scores into the elementary classrooms of the Dallas public schools during 2004-2005.[1]

The paper hypothesizes the school's sorting practices based on the dispersion of prior student test scores on the Texas Assessment of Knowledge and Skills (TAKS) exam. Based

> *This paper seeks to determine the effect of tracking on elementary students, especially those with lower academic performance. Yet it cannot answer the intended question because it ultimately is not a study of tracking.*

on observed classroom test score distributions, schools are classified as sorting or not-sorting schools. Student scores on the TAKS are converted into standard scores in an attempt to determine learning growth in English and mathematics between the third and fourth grade. In addition to TAKS scores, the paper uses the following indicators in the data set to determine sorting practices (i.e., whether there is a clustering of categorized students in three subgroups): English language learner status, gifted and talented status, and special education status.

The paper's purpose is to examine the effects of sorting on student achievement, including the effects of sorting on high and low achievers and on the three subgroups listed above. It concludes that sorting students, especially by test scores, benefits students. Further, it claims that such positive effects are seen among both high and low achievers. The authors then advise that classroom composition should be determined in accordance with its

findings in order to improve student achievement without incurring the high costs associated with other education reforms.

II. Findings and Conclusions of the Paper

The paper first categorizes elementary schools as either sorting or non-sorting. The overall dispersion of scores in a school is compared with the dispersion of scores in individual classrooms. According to the paper, the narrower the dispersion of classroom scores, the more likely it is that the school sorts students. Similar indices were created to measure sorting by gifted and talented, special education, and English language learner status.

The paper then describes the developed measure of student learning growth. It compares third-grade individual students' TAKS scores to those students' fourth-grade TAKS scores. TAKS scores are not on a vertical scale to show growth, however, having been designed for the purposes of cohort comparison. The researchers responded to this obstacle by converting both years' scores to z-scores for standardization, thus placing each student's score on the resulting bell curve of student scores for comparison purposes. As discussed below, this does not result in valid growth measurements.

In order to determine which schools sort and which do not, the paper compares whether the mean math scores, the mean reading scores, and the scores on the three other dimensions (gifted and talented, special education and LEP status) show significant differences among fourth-grade classes. It determines that 19% of the elementary schools sort by math scores, 24% by reading scores, 28% by gifted and talented status, 57% by LEP status, and 13% by special education status. In some cases there is overlap among the scores and the dimensions, in others not.

The paper includes several regression analyses that control for commonly included factors. It concludes that "sorting students into more homogeneous groups is beneficial, particularly for sorting by previous testing score" (p. 19), but it does not adequately connect any model's findings to such claims around student growth or even interpret the significance of the study's findings. It ranks students according to third-grade scores and examines the sorting effect on different groups of students. It concludes that both high and low achievers benefit from classes that have a narrower range of student achievement.

Finally, the paper explores whether sorting by gifted status, special education status and LEP status confers any advantage on students. Effects that are positive but not significant for gifted students are found. For special education students, effects in mathematics and reading are negative, with significant effects for reading. Estimates for LEP students are reported as positive but not precisely estimated. The authors conclude that they have presented "strong evidence" that creating classes with narrower bands of achievement leads to improved learning outcomes.

III. The Paper's Rationale for Its Findings and Conclusions

The paper begins with a discussion of the student sorting decisions that schools make and how those decisions affect tracking and peer effects. It puts forth the argument that sorting students for instruction is, intuitively, a way to increase student performance. The authors opine that if proven effective, grouping by test scores would be a way to increase test scores without any additional expenditure of district funds.

Unlike the majority of studies of tracking, which identify sorting practices through school reports, this study seeks to identify schools that sort students for instruction through an examination of the spread of scores in any given classroom. It ignores the variety of ways that elementary schools group for instruction,[2] however, such as within-class grouping, flexible grouping, re-grouping for specific subjects and between-class grouping. Principals of elementary schools may also assign particular students to particular teachers based on past effectiveness with those or similar students. In other words, the paper bases its findings on the questionable assumption that if there are significant differences in mean scores among classes, the school is engaged in between-classroom grouping, and, if there is not, no grouping is occurring. The paper does not explain why there was no attempt to ascertain the actual grouping practices of the included schools.

IV. The Paper's Use of Research Literature

The literature review is a discussion of a small slice of the literature on ability grouping and peer effects. It comprises six tracking studies, most at the secondary level, and four studies on peer effects at the secondary and college level. Four of the studies cited are not listed in the references, while many studies listed in the references are not cited in the paper.

The literature on the effects of tracking spans decades and includes several meta-analyses. For example, Hattie[3] conducted a meta-analysis of more than 300 studies of ability grouping that included all grade levels and areas of curriculum. He found that overall there was no effect for ability grouping in reading and that the effect in mathematics was slight. Hattie concluded that "tracking has minimal effects on learning outcomes and profound negative effects on equity outcomes."[4] Hattie also examined the effects on subgroups of students and concluded that "no one profits,"[5] including high achievers, from ability grouping.

The majority of recent studies that focus on ability grouping in the elementary years concur with Hattie's meta-analyses. Lleras and Rangel[6] found that minority first- and third-grade students who were placed in lower reading groups experienced substantially lower reading gains than students who learned in non-grouped settings. High-group placement results in slightly greater learning gains. Macqueen[7] used an experimental grouped/non-grouped design to study the effects of grouping on elementary student growth in reading, writing and mathematics and found that grouping by achievement

produced no learning gains. This finding was consistent whether Macqueen looked at the overall student population or at the subpopulations of lower, average and higher achievers. Using international data, Hanushek and Woessmann found that early tracking increased inequitable learning outcomes while depressing overall student achievement.[8] An exception is a study of elementary school tracking in Kenya,[9] which showed beneficial effects for students from grouping, and which was cited by the paper.

In summary, the paper did a poor job of situating the Dallas research within the vast literature on tracking.

V. Review of the Paper's Methods

The paper seeks to estimate the impact of classroom-level sorting on student achievement. To do so, it attempts to examine the impact of sorting on fourth-grade test scores using a variety of regression models and controls. It also attempts to examine how this effect varies over students who are high or low achieving, as well as over students who are designated as gifted and talented, special education, or LEP. The paper employs an instrumental variable procedure to correct for bias in the estimated effect of sorting[10] by using the fifth-grade sorting index from each school as an instrument for the fourth-grade sorting index. The purpose of this instrumental variable is to control for unobserved behavior that may affect both a school's sorting practice and student achievement.

As noted earlier, the paper makes comparisons across two years of test scores that are not vertically linked. In order to account for this, TAKS scale scores are converted to z-scores, which allows for standardization in relation to mean test scores across all students for each year. This is presented as part of an analysis to yield student growth estimates. To understand why this approach is problematic, it may be helpful to first note that the process of determining student growth is not a simple one, as evidenced by the body of recent literature on estimating student growth. Although the z-score conversion in this new study can allow for comparisons of how each student stands in relation to his or her peers in a given year, standardization alone does not create a mechanism by which one can determine learning *growth* in English and mathematics for proficiency examinations. That is, the paper's discussion of changes in z-scores as student growth in English or mathematics is misguided and ultimately lacks understanding of current research on student growth.

Within the context of describing how distinctions between sorting and non-sorting schools was made, the paper presents an example, taken from the Dallas data set, of a sorting school containing four classes. The school appears to divide students based on test scores, yet students identified as gifted and talented, special education, or LEP are distributed across the four classes, which would be uncommon in sorting schools. Other than mentioning that this particular school does not sort students according to these designations, the paper offers no explanation as to why gifted students, who are generally considered high ability, are not concentrated into the classes with the highest average test

scores. Thus, the example points to a key problem with the assumption that an elementary school with non-random test-score distribution is actually engaged in ability tracking.

Moreover, to allow for different effects of sorting on students of different achievement levels, the paper creates dummy variable indicators for both high-scoring and low-scoring

The authors go beyond their data and analyses when they conclude that schools should sort students as a cost-free method to improve student achievement.

students; it then estimates the effect of sorting for each group separately. The paper does not specify the manner in which students are classified as being high or low achievers, however—whether, for instance, they were assigned based on proficiency cut points or some other value.[11]

Research outcomes are considered more reliable if the findings are shown to be robust over multiple specifications of a model. Are the findings similar no matter what reasonable definition is given to "sorted" schools, or "high" and "low" achievers? Although the paper presents various criteria on which students may be sorted and states that the effects of sorting hold across various specifications of the model, it fails to give evidence of robustness to alternative definitions of sorting or other possible classifications of student achievement. Given the anomalous assignment of gifted and talented students across multiple classes of varying levels of achievement, more evidence of the robustness of these findings would have strengthened the paper's contention that the researchers have in fact determined which schools sort and can make conclusions regarding the effects of sorting.

VI. Review of the Validity of the Findings and Conclusions

This paper seeks to determine the effect of tracking on elementary students, especially those with lower academic performance. Yet it cannot answer the intended question because it ultimately is not a study of tracking. It infers that ability grouping exists based solely on the differences in mean values of scores in grade-level classrooms in the same school. In addition, as noted in the previous section, the findings are not shown to be robust to alternative specifications, thus limiting the reliability of the paper's conclusions.

Further, although the study claims to use student learning growth as the basis for its conclusions, it does *not* use a measure of learning growth. It only compares relative standings on one test versus another. Likewise, the paper presents no explanation of how high and lower achievers were determined in order to draw the conclusions that tracking has positive effects on both groups of learners.

The authors go beyond their data and analyses when they conclude that schools should sort students as a cost-free method to improve student achievement. Furthermore, they

should acknowledge the limitations of the study and acknowledge that their findings contradict a vast, longstanding body of tracking research.

VII. Usefulness of the Paper for Guidance of Policy and Practice

For the reasons stated above, the paper is not a useful resource to inform the literature on tracking and ability grouping. In fact, because the paper fails to identify the grouping practices of the studied schools, it is impossible to know if grouping is occurring within or between the classrooms in the study, or not occurring at all.

The paper also fails to present the models used in a sufficiently clear and consistent manner to allow the reader to evaluate them. This concern is heightened by the lack of specification regarding how the researchers determined high and low achievers. Additional caution is advised simply because this is indeed very much a working paper, as demonstrated by its scant and weak review of the literature and the incongruence between cited sources and references.

Although the paper may inform, to a limited degree, the literature on peer effects, it does not inform the literature on tracking or ability grouping. We advise that policymakers and schools disregard the authors' advice and not make policy decisions on grouping based on their results. Evidence-based policy should always be grounded in the complete research base and should focus on high-quality data and analyses.

Notes and References

1 Collins, C.C. & Gan, L. (2013). *Does sorting students improve scores? An analysis of class composition*. Cambridge, MA: National Bureau of Economic Research. Retrieved March 10, 2013, from http://www.nber.org/papers/w18848.

2 For a discussion of the variety of ways that elementary schools group for instruction see Macquenn, S. (2012). Academic outcomes form between-class achievement grouping: the Australian primary context. *Australian Educational Research, 39,* 59-73.

3 Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

4 Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge, 90.

5 Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge, 90.

6 Lleras, C. & Rangel, C. (2009). Ability grouping practices in elementary school and African American/Hispanic achievement. *American Journal of Education 115*, 279-304.

7 Macqueen, S. (2012) Academic outcomes from between-class achievement grouping: The Australian primary context. *Australian Educational Research, 39*, 59-73.

8 Hanushek, E. & Woesmann. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal, 116*, C63-C76. This study was cited in the report, but it was not included in the references.

9 Dflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review, 101* (5), 739-774.

10 Though the complexities of using instrumental variables are beyond the scope of this article, the report's use of sorting indices within the same school leans on the potentially problematic assumption that a measure of fifth-grade sorting is unassociated with other contextual variables that may also affect student achievement, such as school-specific policies and common practices.

11 Though the report gives various estimates of the effect of sorting on student achievement gains, with coefficients on the sorting index (note that these are not effect sizes and are largely uninterpretable) estimated between 0.2883 and 0.6613 for mathematics, little information is provided to the reader that would allow these findings to be understood in terms of their practical significance.

DOCUMENT REVIEWED:	**Does sorting students improve scores? An analysis of class composition**
AUTHORS:	Courtney A. Collins and Lin Gan
PUBLISHER:	National Bureau of Economic Research
DOCUMENT RELEASE DATE:	February 2013
REVIEW DATE:	April 2, 2013
REVIEWERS:	Carol Corbett Burris, Rockville Centre School District Katherine E. Allison, University Of Colorado Boulder
E-MAIL ADDRESSES:	burriscarol@gmail.com Katherine.allison@colorado.edu
PHONE NUMBER:	(516) 255 8820

SUGGESTED CITATION:

Burris, C. C. & Allison, K. E., (2013). *Review of "Does sorting students improve scores? An analysis of class composition."* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-does-sorting-students.