Is kindergarten ability group placement biased? New data, new methods, new answers

Paul T. von Hippel
Ana P. Cañedo

LBJ School of Public Affairs
University of Texas, Austin
2315 Red River St.
Austin, TX 78712

paulvonhippel.utaustin@gmail.com
(512) 537-8112

Is kindergarten ability group placement biased? New data, new methods, new answers

ABSTRACT

Many kindergarten teachers use ability groups to differentiate instruction in reading and math. Ability group placement should depend primarily on student achievement, but critics charge that placement is biased by socioeconomic status (SES), gender, and race/ethnicity. We predict group placement in the Early Childhood Longitudinal Study of the Kindergarten class of 2010-11, using linear and ordinal regression models with classroom fixed effects. The best predictors of group placement are test scores, but girls, high-SES students, and Asian Americans receive higher placements than test scores alone would predict. One third of students move groups during kindergarten, and high-SES students move up more than score gains would predict. Although group placement depends mainly on test scores, there are signs of bias.

Is kindergarten ability group placement biased? New data, new methods, new answers

INTRODUCTION

Every fall, many teachers sort young elementary students into higher and lower "ability groups" for instruction in math and reading. Ability groups—less often called "skill groups" or "achievement groups"[1] (Condron, 2007, National Center for Educational Statistics, 2010)—assign each child to work, for part of the day, in a group of students with similar skill levels. Higher ability groups receive more advanced materials and instruction than lower groups.

Within-classroom ability grouping, which is common in elementary school, has some elements in common with between-classroom "tracking," which becomes common in middle and high school (Gamoran, 2010). The difference is that tracking does not assign students to temporary groups within the same classroom. Instead, tracking routes students into different classes with different levels (e.g., "honors English" vs. "regular English") or even different content (e.g., trigonometry vs. calculus). Tracking is rare in elementary school, where the curriculum is less differentiated and schools are smaller than middle and high schools, with fewer separate classrooms to track students into. In kindergarten—the focus of this article—tracking is often practically impossible, because little or nothing is known about children's achievement levels when they are assigned to classrooms in the summer before kindergarten begins.[2] Ability grouping, however, is common in kindergarten , as we will see.

Like the purpose of tracking, the social and educational purpose of ability grouping has been debated for decades. According to a benign, functional interpretation, ability grouping helps teachers to differentiate instruction—giving each group instruction that is neither too hard nor too easy, but tailored to their current achievement level, challenging every student and maximizing every

[1] The phrase "ability grouping" may sound a little distasteful. In psychometrics "ability" is just a synonym for a student's current skill level, but in economics and sociology "ability" may connote a trait that is fixed or innate. We use the phrase "ability grouping" anyway, because alternatives have not caught on. A Google search for "ability grouping" and "education" returned approximately 35,000 hits, while searches for "achievement grouping" and "education," or "skill grouping" and "education," returned less than 1,000 hits each.

[2] In the ECLS-K, less than 1 percent of the within-school variance in initial test scores lies between classrooms, suggesting that fall achievement levels had little or no influence on children's classroom assignments.

student's opportunity to learn (Tieso, 2003). According to a more nefarious, conflict perspective, though, ability grouping exacerbates inequality, giving more opportunity to children in higher groups, and denying children in lower groups the chance to realize their potential (Oakes & Lipton, 1990). There are several mechanisms by which ability grouping might exacerbate inequality. Teachers might lavish attention on students in higher groups, while neglecting, disparaging, or failing to adequately challenge students in lower groups (Rist, 1970). Students in lower groups might come to identify with lower-achieving peers, and might be distracted from learning if some low-achieving peers have disruptive behaviors (Saleh et al., 2005). The very act of assigning students to ability groups may publicly label students as higher and lower achieving in ways that affect what their teachers, parents, and classmates expect of them, and what they expect of themselves (Pallas et al., 1994). To avoid labeling effects, some teachers give ability groups neutral names like "crickets" and "grasshoppers." However, it is not clear whether such names neutralize the labeling effect or exacerbate it by suggesting there is something to hide.

A key question about ability grouping is which students get placed in higher groups, and why. According to the functional interpretation, each child is assigned to a group that offers them just the right level of challenge. If that is true, then the assignment of students to math and reading ability groups should depend primarily on students' prior reading and math skills. According to the conflict interpretation, higher and lower group placements are the start of a strategy to advance students from higher-status families and hold back students from lower-status families. If that is true, then the assignment of students to higher and lower groups will depend on students' SES, race, ethnicity, or gender. Assignment of equally skilled students to different ability groups on the basis of SES, race, ethnicity, or gender constitutes a form of discrimination or bias. Bias may originate with teachers, or it may be a result of higher-status parents lobbying teachers to give their children higher placement than their skills alone would merit.

Assessing bias in ability group placement is challenging, because even at the start of kindergarten, reading and math skills vary with SES and race/ethnicity, and reading skills vary with gender (Duncan & Magnuson, 2011; von Hippel et al., 2018; von Hippel & Hamrock, 2019). The prevalence of high-SES, white, Asian, or female children in higher kindergarten ability groups may

or may not be a sign of bias. It is not bias if those students are placed in higher groups because of their reading and math skills. It is bias if they get high placements because of their race, ethnicity, gender, or SES.

A third possibility is that certain classroom behaviors or non-cognitive skills may help children get higher group placements than their reading and math skills alone would warrant. Whether this constitutes bias is subject to different interpretations. According to a benign, functional perspective, children who are attentive, respectful, non-disruptive, motivated, and tolerant of frustration may be able to handle more challenging material. Recognizing their readiness to learn, teachers may place such children in higher groups than worse-behaved children with similar reading and math skills. But according to a conflict perspective, teachers' assessments of children's behavior are themselves biased; indeed, the very norms that are set for classroom behavior may be shaped by racial, ethnic, class, or gender bias.

It is debatable how much behaviors and non-cognitive skills should influence group placements. When initial achievement controlled, most social, emotional, and non-cognitive skills—with the important exception of attention deficit—have little predictive value for later learning and achievement in reading and math (Duncan et al., 2007) In addition, we often rely on parent or teachers to report children's behaviors, and it is hard to know to what degree these reports are accurate or biased.

Efforts to assess bias in ability group placement have a long history. An ethnography conducted in 1967 claimed that kindergartners were assigned to ability groups on the basis of social class and deportment, such as speaking standard English, wearing clean new clothes, and keeping a neat desk (Rist, 1970). But that ethnography was limited to a single all-black classroom in St. Louis, Missouri, and did not measure or control for kindergartners' initial reading and math skills.[3] Later quantitative research revisited the question, using larger samples and employing regression analysis to assess whether group placement was better predicted by test scores or by SES, race/ethnicity, and gender. Early quantitative studies sampled hundreds of students from handfuls or dozens of schools

[3] The children in Rist's (1970) study were not tested at the start of kindergarten, when achievement groups were assigned. They did take an IQ test at the end of kindergarten and a reading test at the end of first grade. Neither test appeared to affect their group placements.

in selected states (Gamoran, 1989; Haller, 1985; Haller & Davis, 1980; Hallinan & Sorenson, 1983). More recent quantitative studies analyzed thousands of students from hundreds of schools in the nationally representative Early Childhood Longitudinal Study of the Kindergarten class of 1998-99 (ECLS-K:1999) (Catsambis et al., 2012; Condron, 2007; Jean, 2016; Tach & Farkas, 2006).

Quantitative studies have agreed that reading and math scores are by far the best predictor of ability group placement, while SES remains a significant, though weaker predictor (Condron, 2007; Haller & Davis, 1980; Tach & Farkas, 2006). Results have disagreed with respect to race, with some studies reporting that black students get lower placements than their test scores would predict (Condron, 2007; Hallinan & Sorenson, 1983), while others report that race and ethnicity do not predict group placement, at least once test scores and SES are controlled (Haller, 1985; Tach & Farkas, 2006). In recent studies of the ECLS-K:1999, girls got slightly higher placements than boys with similar test scores, and this was partly but not entirely explained by teachers' higher ratings of girls' classroom behaviors (Catsambis et al., 2012; Condron, 2007; Tach & Farkas, 2006).

Limitations of past studies: Data and methods

Although past studies shed light on the question of why children get placed into higher or lower groups, the data used in past studies of ability group placement had important limitations. First, although teachers first assign students to ability groups near the start of the school year, in fall, most prior studies recorded students' ability group placements toward the end of the school year, in spring. Using spring groups complicates interpretation of the association between test scores and group placement. In the fall, the finding that higher-scoring students are concentrated in higher groups can only mean that test scores, or something correlated with them, have affected group placement. In the spring, though, it is hard to tell whether scores have affected group placement or placement has affected scores. The ECLS-K:1999 only collected ability group placement in the spring, and that limited very study that used it (Catsambis et al., 2012; Condron, 2007; Jean, 2016; Tach & Farkas, 2006). While Hallinan and Sorenson (1983) recorded ability group placement repeatedly in both fall and spring, Haller (1985) only recorded ability group placement in spring, and Haller & Davis (1983) did not report when ability group placement was recorded.

A further complication is that about one-third of students change ability groups between the beginning and end of the year (see our Results). Ideally, one would like to predict group placement in the fall and then predict group mobility from fall to spring of the same school year—but past data rarely offered the detail to support such an analysis. Some studies have predicted changes in ability group from one school year to the next (Gamoran, 1989; Jean, 2016), but only one relatively small study predicted mobility between the beginning and end of a single school year (Hallinan & Sorenson, 1983).

In addition to data limitation, studies predicting ability group placement face several challenges to statistical modeling. A key challenge, often overlooked, is that any model of ability group placement must focus solely on within-classroom variation. Teachers decide group placements by comparing students within the same classroom, so prediction should be limited to within-classroom variation in group placement and predictors such as test scores, SES, race, ethnicity, gender, and behavior. Between-classroom variation in predictors is irrelevant to teachers' decisions about group placement, and analyses that include between-classroom variation can produce misleading results.

To see the potential for between-classroom variation to distort results, imagine a completely segregated, pre-*Brown* school system in which black and white students never share a classroom. Because race does not vary within classrooms, it would be impossible to know if any teacher would place a black student above or below a white student with similar scores. Any analysis that includes between-classroom variation will draw misleading conclusions about race's influence on group placement. For example, a model might suggest that race does not predict placement because black students are just as likely as white students (in different classrooms) to be placed in a high group. If black students have lower average scores, a between-classroom analysis may even conclude that black race predicts *higher* placement, because black students are more likely to get high group placements than are white students—in different classrooms—with comparable scores.

This is a simplified example and does not imply that analyses including between-classroom variation will always underestimate bias against black children. In general, whether bias against black children is under- or under-estimated will depend how black and white children are distributed

across classrooms and on the correlation between race and other predictive variables, both within and between classrooms.

Models can be limited to within-classroom variation by incorporating classroom fixed effects, which can be incorporated into a linear models by including a dummy variable for each classroom or, equivalently, centering every variable around its classroom mean (Allison, 2009). Yet no studies of ability group placement have used classroom dummies, and only one study has used classroom-centered variables (Tach & Farkas, 2006), though it did not describe its model as one with classroom fixed effects. Some studies, especially older ones, used pooled linear regression analysis, which did not distinguish within- from between-classroom variation (Condron, 2007; Haller, 1985; Haller & Davis, 1980; Hallinan & Sorenson, 1983). Recent studies have favored hierarchical linear models with random effects at the classroom and/or school level (Catsambis et al., 2012; Jean, 2016; Tach & Farkas, 2006), but coefficients from random effects models still give some weight to between-classroom variation (Greene, 1999; Wooldridge, 2001)—unless they center variables around classrooms means, which again only one study has done (Tach & Farkas, 2006). Some models have even included school- or classroom-level predictors, such as percentage of students in poverty (Condron, 2007; Tach & Farkas, 2006), but because these variables do not vary within classrooms, they cannot predict ability group placement and should be omitted from models that try to do so.

How much might classroom fixed effects change predictions about ability grouping? There are examples where different analyses of the same data have reached different conclusions. One study of the ECLS-K:1999, which omitted classroom fixed effects, concluded that black children received lower placements than white children with similar scores and other characteristics (Condron, 2007). But another study of the same data, which implicitly included classroom fixed effects through classroom centering, concluded that there was no significant difference between the placements of black and white children with similar scores (Tach & Farkas, 2006).

Another methodological challenge is that ability group placement is a tricky dependent variable to model. Ability groups are ordinal, and the number of groups varies across classrooms: some classrooms have two groups, some have three, four, five or more. Authors have coded ability

groups in different ways, whose implications for estimation are unknown. Some authors have transformed group placements into a standard score (Tach & Farkas, 2006) or quantile score (Condron, 2007; Gamoran, 1989) that has a similar but not identical distribution regardless of the number of groups in a classroom. Transformation puts classrooms with different numbers of groups on a common scale, but it ignores the fact that ability groups are ordinal and not interval. Some authors have collapsed the ability group variable down to two categories (lowest group vs. other, or highest group vs. other) or three categories (lowest group vs. highest group vs. other) which are analyzed using logistic regression (Catsambis et al., 2012; Hallinan & Sorenson, 1983; Jean, 2016). This approach keeps the variable ordinal, but discards variation, reduces power, and makes classrooms less comparable since the meaning of the lowest or highest group depends on how many groups a classroom has. In a classroom with two groups, some students in the lower group may be just a little below the average for their classroom, but in a classroom with five groups, students in the lowest group are likely struggling.

Our contributions

In this article, we use new data and improved methods to update and reassess the question of whether ability group placements are biased. We predict group placements in a relatively new and nationally representative dataset that records students' ability groups near both the beginning and end of kindergarten. We predict initial group placement in fall and then predict group mobility, or which students move up or down between the beginning and end of kindergarten. All our models include classroom fixed effects. We use mathematical statistics to compare the properties of different ways to code and model the ability group variable, including a new approach that preserves the variable's ordinal character and still allows the number of groups to vary across classrooms. We also evaluate the influence of group coding and other decisions on our empirical results.

DATA

Our data come from the Early Childhood Longitudinal Study of the Kindergarten Class of 2010-11 (ECLS-K:2011), which began in the fall of 2010 with a nationally representative probability

sample of 15,088 US kindergarteners. The ECLS-K:2011 was a two-stage cluster sample, with children clustered in schools, and schools clustered in primary sampling units (PSUs), each of which was either a large county or a group of similar and contiguous small counties. The sample did not cluster by teacher or classroom, so within a school, the sampled students could be scattered across different classrooms taught by different teachers. On average, the ECLS-K:2011 sampled 18 children per school, which worked out to 5 children per classroom. Six percent of teachers had two kindergarten classrooms, one in the morning and one in the afternoon.

Kindergarten teachers were asked about ability grouping in fall, near the beginning of the school year, and again in spring, near the end. We used these questionnaires to define a *fall sample*, which we used to predict initial placement, and a *fall-spring sample*, which we used to predict group mobility between fall and spring. We also defined a *restricted fall-spring sample*, which was smaller but perhaps had higher data quality. We will define these samples shortly.

The ECLS-K:2011 continued to follow children through fifth grade, and teachers were asked about ability grouping again in first and second grade. In first and second grade, however, teachers were only asked about ability grouping in spring, not in fall, so it would be impossible to tell which group students were placed in initially. We therefore restrict our analysis to kindergarten. While our focus on kindergarten was really dictated by the data, and we would gladly have analyzed first and second grade if we could, our results are not terribly different from an earlier analysis of first grade ability grouping carried out by Tach & Farkas (2006). That said, there are reasons to think that ability grouping in kindergarten, particularly in the fall, may be distinctive in some ways. First, because kindergarten is the start of formal schooling, children's achievement levels at the start of kindergarten are shaped entirely by out-of-school factors and not by their responsiveness to school. Second, kindergartners learn basic reading skills more quickly in kindergarten than at later ages (von Hippel & Hamrock, 2019), and may be exceptionally sensitive to classroom practices such as ability grouping. Third, modern kindergarten seems to serve a compensatory or preparatory function of getting children to a minimum standard, so that achievement gaps between advanced and less advanced children shrink during kindergarten, at least in the ECLS-K cohorts (von Hippel, Workman, and Downey 2018). The same may not be true in later grades.

Fall sample

We defined the fall sample using the fall kindergarten version of the "Teacher Questionnaire (Child Level)," called Questionnaire T1. Questionnaire T1 asked teachers the following question separately for reading and math:[4]

> *"How many achievement groups in [reading/math] do you currently have in this child's class?"*

The question was multiple choice; possible answers were 2, 3, 4, and "5 or more," and "I do not use achievement groups for [reading/math]." According to this question, about half of kindergarten classrooms used ability grouping in reading, and about one in six used ability grouping in math. We restricted the sample to these classrooms.

Teachers who reported using ability groups were asked, for each sampled child,

> *"In which [achievement] group is this child currently placed? Use 1 for the highest achievement group."*

Following convention, we reverse coded the answer so that 1 was the lowest group.

We had to drop about 2 percent of children from ability-grouped classrooms because the information about their group placement was inconsistent or ambiguous. For example, it was possible for a teacher to report using, say, 3 ability groups but then report that some child was in group 4 or higher. We dropped such children. In addition, if a teacher reported using "5 or more" ability groups, it was not clear whether a student in group 5 was in the highest group. Had the data included every child in the classroom, we would have known whether group 5 was the highest, but because the data were only a sample, it was possible there was a child in group 6 (or higher) whom we didn't know about. So we dropped children with group numbers greater than 5.

Table 1 reports the exact number of children, classrooms, teachers, and schools in our fall sample. It shows the number in the full ECLS-K:2011 and the number left after each restriction, giving the size of the fall sample in row 4. Table 2 describes the distribution of ability groups in the fall sample. In reading, classrooms with 3, 4, or 5 or more groups were common, and classrooms

[4] On the fall questionnaire (Questionnaire T1), this was question 4 for reading and question 6 for math. On the spring questionnaire (Questionnaire T2), this was question 20 for reading and 22 for math.

with only 2 groups were rare. In math, though, classrooms with 2, 3, or 4 groups were common, and classrooms with 5 or more groups were rare.

Fall-spring sample

We next defined a fall-spring sample to use in analyzing children's group mobility between fall and spring. To do this, we first defined a spring sample using the spring version of the "Teacher Questionnaire (Child Level)," called Questionnaire T2, which asked the same ability grouping questions as Questionnaire T1. As we did on Questionnaire T1, on Questionnaire T2 we restricted the sample to students in ability groups 1 through 5 whose ability group number did not exceed the number of groups that the teacher reported using.

We then defined a fall-spring sample by merging the fall and spring sample and restricting the merged data for consistency. Specifically, we restricted the fall-spring sample to children who were in the same classroom with the same teacher in both fall and spring. We further restricted to teachers who reported using the same number of ability groups in fall and spring. There was a surprising amount of disagreement on this question. In math, about a quarter of teachers who reported using ability groups in fall did not report using ability groups in spring, and a further quarter reported using a different number of ability groups in fall than in spring. In reading, nearly all teachers who reported using ability grouping in fall also reported using ability grouping in spring, but nearly a third reported using a different number of groups. It is not clear whether these disagreements reflect data quality issues or actual changes in ability grouping practice between fall and spring. Either way, we dropped classrooms with inconsistent ability grouping data.

The penultimate row of Table 1 gives the size of the fall-spring sample after restrictions. The fall spring sample is one-third smaller than the fall sample in reading, and only half as large as the fall sample in math. According to Table 3, though, the fall and fall-spring samples are similar with respect to the number of ability group classrooms, as well as gender and race/ethnicity. Our models also produced very similar results whether they were fit to the fall sample or to the fall-spring sample.

Restricted fall-spring sample

We also defined a *restricted fall-spring sample* that used further but lower-quality data about ability grouping. This sample relied on a "Teacher Questionnaire," called Questionnaire A1 in fall and A2 in spring, that teachers filled out in addition to Questionnaires A1 and A2. In spring, but not in fall, the Teacher Questionnaire included this question about ability grouping:[5]

B4. On days when you use achievement grouping, how many groups do you have in your class or classes? How many minutes are your class or classes usually divided into achievement groups for reading and math activities or lessons?

IF YOU HAVE MORE THAN ONE CLASS, WRITE THE AVERAGE FOR YOUR CLASSES. IF YOU DO NOT USE ACHIEVEMENT GROUPING IN THE SUBJECT LISTED, PLEASE WRITE "0" IN THE NUMBER BOX AND SKIP TO THE NEXT QUESTION.

	Number of achievement groups	1-15 minutes/ day	16-30 minutes/ day	31-60 minutes/ day	More than 60 minutes/ day
a. Reading	☐☐	☐	☐	☐	☐
b. Mathematics	☐☐	☐	☐	☐	☐

This is a poorly worded, multi-barreled question. It really combines three questions: Do you use ability grouping? If so, how many groups? And for how many minutes? Adding to the confusion, the question does not specify whether the number of minutes is per day or per week.

The response options are also confusing. The box for the number of achievement groups permits a two-digit, free response answer, but the number of groups is almost sure to be a one-digit number, and when Questionnaires T1 and T2 asked the number of groups, they allowed just four discrete choices (2, 3, 4, and 5 or more). By contrast, the number of minutes is a continuous variable, but the question allows just four discrete choices.

Perhaps because of the confusing question design, about a quarter of teachers in the fall-spring sample reported a different number of groups on Questionnaire A2 than they did on Questionnaires T2 and T1—even though half of teachers filled out Questionnaires T2 and A2 within

[5] This is part of question B4 on Questionnaire A2. A version of the same questionnaire was administered in the fall of kindergarten (Questionnaire A1), but it omitted this question.

a day of one another, and three quarters of teachers filled out those Questionnaires within ten days. Among the teachers who gave discrepant answers, a few answers to Questionnaire A2 simply did not make sense. One percent of teachers said they used only one ability group in reading, and two percent said they used only one ability group in math—but the number of ability groups must be at least two. In addition, one percent of teachers said that they used between 10 and 33 ability groups in reading, perhaps confusing the number of groups with the number of minutes.

Notwithstanding our misgivings about Questionnaire A2, it does offer another source of information about the number of ability groups. We therefore defined a restricted fall-spring sample containing teachers from the fall-spring sample who reported the same number of ability groups on questionnaire A2 as on Questionnaires T1 and T2. The number of students, classrooms, teachers, and schools in the restricted fall-spring sample is given in the last row of Table 1. Despite the restrictions, the restricted fall-spring sample was demographically similar to the other samples, and models produced similar estimates when fit to any of the three samples, although statistical power decreased with sample size.

Test scores

Children took reading and math assessments in the fall and spring of kindergarten. Evaluation started with an oral "screener"—consisting of a "Simon Says" game and a child-directed "Art Show"—to evaluate whether the child had a basic command of spoken English. Children who scored above a threshold on the screener proceeded to a two-stage assessment. In the first stage, they took English language "routing tests" in reading and mathematics, which contained items of varying difficulties. The results of the routing tests determined whether they got an easy, medium, or hard test in the second stage. Children who scored below the threshold on the screener proceeded to take a test of English basic reading skills, consisting of the just the easiest questions on the English reading routing tests. Children who spoke Spanish then took two stage assessments of reading and math skills in reading and mathematics.

As a result of this assessment procedures, all tested students received a score of their reading ability in English, and all tested students who spoke English or Spanish received a score assessing

their ability in math. These are the scores that we used in our analyses. Math scores were missing for students who spoke neither English nor Spanish at a basic level, and either math or reading scores could be missing for students who were not available for assessment.

Children's ability scores, or "theta" scores, were estimated using an item response theory (IRT) model with three parameters to control for the difficulty, discrimination, and guessability of test items. According to the IRT model, the fall reading score had a reliability of 95 percent and the fall math score had a reliability of 92 percent (Najarian et al. 2018, Tables 5-7 and 5-14).

Fall tests were most often given in October, and spring tests were most often given in April. Although different schools took the tests on different days, within schools students almost always took the tests within a few days of each other. So test scores compare the skill levels of students in the same classrooms at very nearly the same time. Our fixed effects models control for differences in test dates between students in different classrooms.

Student demographics

Student gender, race/ethnicity, and socioeconomic status (SES) were recorded from parent questionnaires and data provided by school administrators. We collapsed race and ethnicity into five categories: Hispanic and four non-Hispanic groups—white, black/African American, Asian, and other. Within these five broad categories, the data did offer some smaller groups, but the groups were so small that an analysis using them would have lacked the power to make meaningful distinctions.

The ECLS-K:2011 coded SES by averaging standardized variables measuring family income, parents' occupational status, and parents' years of education. We standardized the SES measure to facilitate interpretation. This was necessary because an average of standardized variables is not itself standardized, as proven in this footnote.[6]

[6] Consider two standardized variables Z_1 and Z_2, each with mean 0 and variance 1 Their average $Z=(Z_1+Z_2)/2$ also has a mean of 0, but its variance is not one. Instead, the variance of the average is $\text{Var}(Z)=(\text{Var}(Z_1)+\text{Var}(Z_2)+2\text{Cov}(Z_1,Z_2)/4=(2+\text{Corr}(Z_1,Z_2))/4$, which is less than 1—in fact less than ¾. So Z is not standardized.

In both reading and math, about half the students were white non-Hispanics, and most of the remainder were either black or Hispanic. The fraction of Asian Americans was relatively small—6 percent of the fall sample and 7 percent of the fall-spring sample—but we still had enough statistical power to show that being Asian American was a significant predictor of children's group placement—as we will show.

Teacher-rated behavior

In the fall and spring of kindergarten, teachers answered a number of items asking how often children displayed certain behaviors and social skills (Tourangeau et al., 2018). Available responses to each item ranged from never (1) to very often (4), except for items describing children's attentional focus, where available responses ranged from "extremely untrue" (1) to "extremely true" (7). Responses were reduced to seven scales, each constructed by averaging responses across four to seven items. Each scale ranged from 1 to 4, except for the scale for attentional focus, which ranged from 1 to 7. To facilitate interpretability and comparison, we standardized every scale to a mean of 0 and an SD of 1.

Two scales were adapted from the Short Form of the Children's Behavior Questionnaire (Putnam & Rothbart, 2006):

1. *Attentional focus*. This scale consisted of six items that "measure the child's tendency to maintain attention on a task" (National Center for Education Statistics, 2010). It had a reliability of 0.87 (Tourangeau et al., 2018).
2. *Inhibitory control*. This scale consisted of six items that describe the child's "capacity to plan and to suppress inappropriate approach responses under instructions or in novel or uncertain situations" (Rothbart et al., 2001, p. 1406). It had a reliability of 0.87 (Tourangeau et al., 2018).

Four scales were adapted from the Social Skills Rating System (Gresham & Elliott, 1990; NCS Pearson, 1990). The component items are masked due to copyright, but they measure the following constructs:

3. *Interpersonal skills*. This scale consisted of four items describing whether "the child interacted with others in a positive way" (Tourangeau et al., 2018). It had a reliability of 0.81 (Tourangeau et al., 2018, Table 3-9).
4. *Externalizing problem behaviors*. This scale consisted of five items describing behaviors such as "fighting, arguing," and "impulsiveness" (National Center for Education Statistics, 2010). It had a reliability of 0.88 (Tourangeau et al., 2018, Table 3-9).
5. *Internalizing problem behaviors*. This scale consisted of four items describing emotional and cognitive conditions such as "depression, low-self-esteem" (National Center for Education Statistics, 2010). It had a reliability of .79 (Tourangeau et al., 2018, Table 3-9).
6. *Self-control*. This scale consisted of four items. Its interpretation was not defined in the ECLS-K:2011 documentation, but its name is somewhat self-explanatory, and it had a reliability of 0.81 (Tourangeau et al., 2018, Table 3-9).

The final scale was

7. *Approaches to learning*. Constructed specifically for the ECLS-K, this scale consisted of seven items: "keeps belongings organized; shows eagerness to learn new things; works independently; easily adapts to changes in routine; persists in completing tasks; pays attention well; and follows classroom rules" (Tourangeau et al., 2018). Its reliability was not specified in the ECLS-K:2011 documentation, so we assumed a reliability of 0.85, which was typical for the other behavioral measures.

Note that the behavior scales represent teachers' subjective impressions of behavior, which, though highly reliable and guided by a rubric, were not necessarily objective or unbiased. For example, some subjectivity or bias may affect whether a teacher indicates that a student displays a certain behavior "somewhat often" or "very often."

METHODS

We use different types of regression models to predict which children were placed in higher and lower groups within their classrooms. We carefully considered various modeling issues, including the coding of the dependent variable (ability groups), measurement error in some

independent variables (especially test scores, socioeconomic status, and teacher-reported behavior), and the inclusion of classroom fixed effects to ensure that children were only compared to other children in the same classroom.

Coding of ability groups

When a classroom had *K* ability groups, we assigned them values *k*=1,…,*K*, where group *k*=1 was the lowest group and group *k*=*K* was the highest. Because the number of groups *K* varied across classrooms, the distribution of the variable *k* was different in different classrooms. We used two different methods to transform *k* so that it had a similar distribution in every classroom. But before transforming *k*, we need to understand the distribution of *k* itself.

Within a classroom with *K* groups, the groups may not have equal numbers of children, but if they do, then the variable *k* has a discrete uniform distribution, with the following mean and standard deviation (SD):

$$E(k) = \frac{K+1}{2}$$

$$SD(k) = \sqrt{\frac{K^2-1}{12}}$$

Notice that the mean and SD increase with the number of groups *K*. This can cause inconsistent estimates because *K* varies across classrooms. For example, when the number of groups *K* is 2, individual student group numbers *k* have a mean of 1.5 and an SD of 0.5. But when *K*=5, individual student group numbers *k* have a mean of 3 and an SD of 1.4.

We tried two different methods to recode *k* so that it had the same mean and SD in different classrooms, at least approximately.

Standardization

The simplest approach, used in at least one prior study (Tach & Farkas, 2006), was to standardize *k* within each classroom *c* by subtracting the classroom mean $\overline{k}_c$ and dividing by the

classroom standard deviation s_c. The result is a standardized variable z which, within each classroom,[7] had a mean of 0 and an SD of 1:

$$z = \frac{k - \overline{k}_c}{s_c}$$

The ECLS-K:2011 sample typically did not include all the children in classroom c, and so we had to estimate $\overline{k}_c$ and s_c from a sample. In about 10 percent of classrooms, the sample contained only one child or all sampled children were in the same group; then s_c was either undefined or zero, so z could not be calculated.

Percentile coding

A slightly more complicated approach, used in at least two prior studies (Condron, 2007; Gamoran, 1989), was to transform k into an implied quantile q, such as a percentile.[8] The basic idea is this: If there are K=2 ability groups, you treat the lower group (k=1) as though its members are between percentiles 0 and 50 and the upper group (k=2) as though its members between percentiles 50 and 100. You then assign each group the midpoint of its percentile range, so that the lower group is coded as q=25 and the upper group is coded as q=75. Likewise, if there are three ability groups, you code the low group (k=1) as q=16 2/3 (the midpoint of 0 and 33 1/3), the middle group (k=2) as q=50 (the midpoint of 33 1/3 and 66 2/3), and the high group (k=3) as q=83 1/3 (the midpoint of 66 2/3 and 100). More generally, if there are K groups, then the percentile q corresponding to group k=1,…,K is

$$q = 100\frac{k - 1/2}{K}$$

Percentile coding makes the most sense when there is an equal number of children in each group, but it is also used when groups are unequal in size.

[7] The total SD of z is somewhat less than the within-classroom SD of 1, because the between-classroom sum of squares is constrained to zero.

[8] Both Gamoran (1989) and Condron (2007) used deciles d, but we use percentiles q because they are easier to interpret. The two are related by d=q/10.

Since q is just a linear transformation of k, the mean and variance of q can be derived from the mean and SD of k. If the K groups are equal in size, then k has the mean and SD given above, and the mean and SD of q are

$$E(q) = 50$$

$$SD(q) = 100\sqrt{\frac{K^2 - 1}{12K^2}}$$

Table 4 summarizes the distribution, mean and SD of q for classrooms with K equal-sized groups, where K can take any value from 2 to 5. The mean is 50 regardless of K. The SD increases with K, but only slightly; it rises from SD=25 when K=2 to SD=28.3 when K=5. Therefore q behaves much like a variable that has been standardized to a mean of 50 and a standard deviation somewhere between 25 and 28.3.

Because the SD of q is a little more than 25 times the SD of the standardized variable z, the slopes of a linear regression that uses q will be a little more than 25 times the slopes of a linear regression that uses z. Besides that, the regression results should be very similar. A small advantage of q over z is that q is defined when the sample includes only one child from a classroom, or when all sampled children are in the same group. This slightly increases the available sample size in some analyses.

Ordinal models

The q and z transformations assume that, for a given value of K, k is an interval variable with an equal distance between groups. But this is not necessarily the case. In a classroom with K=3 groups, or example, suppose that groups k=1 and 2 differ only a little in achievement level, while group k=3 is a "gifted" group with a much higher achievement level. Or suppose that groups k=2 and 3 differ only a little, while group k=1 is a remedial group that starts far behind. In either case, k would be an ordinal variable, but not an interval one.

Because k is an ordinal variable, it is natural to model it using ordinal logistic regression. In classroom c with K ability groups, the probability that student i is placed in group k is

$$P(i \in k) = P(\tau_{k-1} < y^* \leq \tau_k), where\ y^* = \alpha_c + \beta X_i + e$$

Here y^* is an unobserved latent variable. α_c is a class-specific fixed effect, β is a vector containing the slopes of the child variables X_{ic}, and e is an unobserved residual with a standard logistic distribution. τ_k and τ_{k-1} are thresholds; to identify the model, the top and bottom threshold, τ_1 and τ_K, are set equal to $-\infty$ and $+\infty$, respectively.

As usually implemented, ordinal regression models assume that the number of groups *K* is the same for every classroom. To overcome this limitation, we fit an ordinal logistic regression model separately to classrooms with *K*=2, 3, 4, and 5 groups,[9] and then averaged the results[10] across the five regressions, giving more weight to the regressions that have smaller standard errors (because of larger sample sizes). To do this, we adopted a formula widely used in meta-analyses that average heterogeneous effects across multiple studies (DerSimonian & Laird, 1986).

Specifically, suppose $\hat{\beta}_K$ and s_K estimate the coefficient β_K and standard error for a particular *X* variable in an ordinal logistic regression fit to classrooms with *K* groups. Then there are five coefficients $\hat{\beta}_K$, *K*=2,3,4,5,6, and their weighted average is

$$\bar{\beta} = \frac{\sum_{K=2}^{6} \hat{\beta}_K w_K}{\sum_{K=2}^{6} w_K}$$

with standard error $SE(\bar{\beta}) = \sqrt{\sum_{K=2}^{6} w_K}$. Here the weights are

$$w_K = 1/(s_K^2 + \hat{\tau}^2)$$

and $\hat{\tau}^2$ is an estimate of how much the true coefficients β_K vary across the five regressions (DerSimonian & Laird, 1986).[11]

[9] In reading, the sample of classrooms with *K*=2 groups was quite small and the ordinal logistic regression model did not converge. Our reading results therefore average only the results for classrooms with *K*=3,4, or 5 groups only.

[10] It is appropriate to average coefficients because the coefficients of the five regressions are on the same standard logistic scale.

[11] This formula, implemented by the user-developed Stata command *admetan, re* (Fisher, 2018), is appropriate for "random-effects" meta-analysis, which allows for the possibility that the true coefficients *K* are different for classrooms with different numbers of groups *K*. Our use of a formula from random-effects meta-analysis should not be confused with the use of fixed effects in the underlying regressions. The meaning of the terms fixed vs. random effects is different in random-effects meta-analysis than in fixed-effects regression.

Interpretability

As it will turn out, the three codings of ability grouping produce similar results. Coefficients that use percentile coding are approximately 25 times the coefficients that coefficients that use standardized coding. The ordinal model does not have such a tidy relationship to the linear models, but nevertheless produces coefficients that are similar in direction, significance, and relative size.

The question then arises which coding produces the most interpretable results. Logit results are famously hard to interpret (von Hippel 2015), and standardized coefficients are often interpreted according to arbitrary thresholds, such as effects of less than 0.2 standard deviations being classified as "small" (Cohen 1988).

By contrast, the percentile coding is easy to interpret, since the distance between groups can be defined in percentile terms. With K=5 groups, for example, the distance between the group midpoints is 20 percentile points, so an effect size of 10 percentile points may be considered large. More specifically, an effect size of 10 percentile point is half the average difference between groups. Equivalently, for half of children, an increase of 10 percentile points would be enough move them into a higher group—if they are not in the highest group already.

Note that the interpretation of effect size depends on how many groups there are in a classroom. In a classroom with 5 groups, an effect size of 10 percentile points may seem large. But in a classroom with just two groups, where the average difference between groups is 50 percentile points, an effect size of 10 percentile points would be smaller. Increasing the percentile of a randomly selected child from the lower group by 10 points would have only a one in five chance of moving them to the high group.

Classroom fixed effects

As mentioned in the introduction, any model of ability group placement should include classroom fixed effects, which limit estimates to within-classroom variation. Classroom fixed effects eliminate between-classroom variation in group placement, observed predictors, and unobserved confounders. In the ECLS-K:2011, including classroom fixed effects also has the benefit of controlling away between-classroom variation in test dates. Nearly all the variation in test dates lay

between classrooms, in fact between schools. Within classrooms, test dates rarely differed by more than one day, or three days if a weekend intervened.

In a linear regression model, it is straightforward to include classroom fixed effects—either by adding classroom dummies or, equivalently, by centering all variables around their classroom means (Allison 2009). We fit linear regression with classroom fixed effects using the Stata command **xtreg, fe**. Standard errors were clustered at the PSU level.

Including fixed effects in ordinal logistic regression models is a little harder, since dummies or centering cannot produce consistent fixed effects estimates with a logit model (Allison 2009). Instead, we use the "blow up and cluster" (BUC) estimator (Baetschmann et al., 2017). The intuition behind the BUC estimator is that, when there are K=2 groups, consistent fixed effects estimates can be obtained using conditional logistic regression (Chamberlain 1979), and when there are more than two groups, the data can be reduced to two groups by dichotomizing—for example by defining groups 3 and higher as 1 and lower groups as 0. The problem with dichotomizing is that it sacrifices efficiency by ignoring variation in the dependent variable. The BUC estimator restores efficiency by dichotomizing the data at every possible level—"blowing it up"—and then using clustered standard errors to adjust for the fact that several dichotomized observations come from the same child. By default, standard errors are clustered at the child level, but they can also be clustered at higher levels; we cluster at the PSU level.

Measurement error in predictors

Several readers[12] raised concerns that test scores are measured with error, and regression models with error in regressors can produce biased estimates. This is true, and in fact test scores are not the only regressors that contain measurement error. Teacher-reported behaviors and SES have measurement error, too.

To compensate for measurement error, we fit a fixed effects model that allowed for errors in variables (EIV). We did this by applying Stata's **eivreg** command to data that had been classroom-

[12] We thank Doug Downey, Adam Gamoran, and an anonymous reviewer for raising this point.

centered using Stata's **xtdata** command. We accounted for measurement error in test scores and behaviors using their estimated reliabilities (reported above), since reliability, according to classical measurement theory, is the percentage of variance that is not due to measurement error (Allen & Yen, 2002). There is no estimate for the reliability of the SES measure used in the ECLS-K:2011, but SES measures used in other studies have been found to be 81 to 91 percent reliable (Cirino et al. 2002). So we conducted two versions of our errors-in-variables regression, one assuming that SES was 81 percent reliable and one assuming that it was 91 percent reliable. (Gender and race/ethnicity were assumed to be measured without error.) The errors-in-variables model is only estimable using linear regression; our ordinal regression cannot account for errors in regressors. We use standard errors bootstrapped at the PSU level, because **eivreg**'s analytic standard errors are biased (Lockwood & McCaffrey 2020).

While the errors-in-variables model helps to test the sensitivity of our results, it is debatable whether correcting for measurement error improves the validity of the results. Analysts typically correct for measurement error when they believe the true value, and not the error-prone measurement, influenced the outcomes. However, kindergarten teachers do not base their decisions about ability grouping on students' true abilities, true SES, or true behaviors. Instead, teachers' rely on their own formal or informal assessments of students' abilities and SES, which, while different from the measures in the ECLS-K, are likely at least as prone to error. Teachers' also rely on their own assessments of students' behaviors, which are prone to error, as reflected in the imperfect reliability of the behavior measures in the ECLS-K—which may not be the same as the more informal behavior judgments that teachers use to make grouping decisions.

In short, it is unclear whether correcting for measurement error would produce a more accurate picture of why students are placed into higher and lower ability groups. Fortunately, the question is moot because correcting for measurement error turns out to make very little difference to the estimates. We could foresee this even before seeing the results, because the reliability of the variables is relatively high—between 79 and 95 percent—and fairly similar across variables. With little measurement to correct, and most variables being corrected by a similar amount, correcting for measurement error does not materially change the estimates.

Modeling mobility

In addition to predicting initial group placement in the fall sample, we also used the fall-spring sample to model mobility in group placement between fall and spring. We did this by regressing spring group placement on fall group placement, student gender, SES, and race/ethnicity, and changes in test scores and behaviors. In the linear models, groups were coded the same way in spring as in fall—either as a percentile or as a standardized score. In the ordinal logit models, spring groups (the dependent variable) were coded as an ordinal variable, but fall groups (as independent variables) were standardized.

Interpreting the relationship between changes in group and changes in scores and behaviors is debatable because causality could flow either way. Teachers might promote children to a higher group because their scores and behaviors have improved more quickly than other children's—or children's scores and behaviors might have improved more quickly because the teacher promoted them. The relationship between group mobility and other predictors is less debatable because gender, SES and race/ethnicity do not change between fall and spring. But if the slopes of changing scores and behaviors are biased, then the slopes of correlated predictors that do not change may be biased as well. Despite these uncertainties, the mobility results hold substantive interest, so we present them with qualifications.

RESULTS

Initial group placement in the fall of kindergarten

Table 5 and Table 6 use the fall sample to predict the initial placement of kindergartners into higher and lower ability groups in reading and math, respectively. The tables show results from fixed effects linear models using both standardized and percentile coding for groups, as well as fixed effects ordinal logit models. The linear models explain about half the within-classroom variance in group placement. In general, the results are quite similar across the three model specifications; the percentile coefficients are a little more than 25 times the standardized coefficients, as we predicted, and the coefficients of the ordinal logit are approximately 3 to 5 times the coefficients that used

standardized group numbers. We will report all three results, but emphasize the percentile coefficients because they are the most interpretable.

We will start by summarizing the models that omit teacher assessments of students' behavior, then report results for models that include those assessments.

Models without teacher-reported behaviors

By far the strongest predictor of group placement was standardized test scores. Reading scores were about twice as important as math scores in predicting reading group placement, but reading and math scores are about equally important in predicting math group placements. In reading, when SES, gender, and race/ethnicity, and classroom were held constant, a one SD increase in reading scores predicted a group placement that was 14 points higher on the percentile scale, 0.5 SD higher on the standardized scale, and 1.7 logits higher on the logit scale. In math, a one SD increase in math scores predicted an increase of 11 percentile points, 0.38 SD, or 1.2 logits in group placement. These are sizable coefficients; in a five-group classroom, a coefficient of 11 to 14 percentile points is more than half the average difference between adjacent groups, such as group 4 and group 5.

Net of test scores, both SES and gender were significant predictors of group placement, suggesting that ability group placements were biased in favor of girls and higher-SES children. The coefficients of gender and standardized SES were about 2 to 4 points on the percentile scale, 0.1 SD on the standardized scale, or 0.2 to 0.3 logits. These are small coefficients, but not trivial if the number of groups in the classroom is large. In a five-group classroom, for example, a coefficient of 2 to 4 points is 10 to 20 percent of the difference between adjacent ability group. In other words, among girls and high-SES students whose scores alone are not high enough to put them in the highest group, 10 to 20 percent receive higher placements than boys and average-SES students with similar test scores.

There was also evidence that ability group placement was biased in favor of Asian Americans. The coefficient for Asian Americans was statistically significant and comparable in size to the coefficient for girls. Asian Americans' reading group placements were approximately 4 percentile points higher than white children in the same classrooms with similar test scores and SES.

In math, the Asian American coefficient was smaller and not statistically significant, though the sample size was smaller and there was less power to detect an effect, especially for a group as small as Asian Americans.

There was no evidence of bias against African American or Hispanic children. In both reading and math, the coefficients for black and Hispanic children were 1 percentile point or less, far from statistically significant, and as likely to be positive as negative. Because Hispanic and African American children tend to have lower test scores and lower SES, they did tend to get placed in lower ability groups, but net of test scores and SES, there was little evidence that their race or ethnicity *per se* predicted their group placement.

Models with teacher-reported behaviors

Adding teacher-reported behaviors increased explained variance by just 2 to 6 percentage points, but several of the behaviors were significant predictors.

In reading, net of other variables, teachers gave significantly higher placements to students who in the teacher's judgment had good approaches to learning and strong attentional focus; the coefficients were 4 to 5 percentile points in both reading and math—small effects, but not trivial in a five-group classroom. Teachers gave significantly lower reading group placements to students who displayed more internalizing problem behaviors (e.g., depression, self-esteem), but surprisingly they gave significantly higher reading group placements to students who displayed more externalizing problem behaviors (e.g., fighting, arguing). Students' interpersonal skills, self-control, and inhibitory control did not predict their reading group placements.

Including behavioral variables in the reading group model substantially changed the coefficients for some of the other predictors. Most conspicuously, the coefficients for female gender shrank by more than half in reading and became non-significant in math. This suggests that the fact that girls are placed a little higher than boys with similar scores has primarily to do with girls' better behavior—at least as rated by kindergarten teachers, who are almost all female themselves. Adding behavioral variables also shrank the coefficients of SES and Asian ethnicity, but only by 10 to 20 percent.

Table A 1 and ***p<0.001, **p<0.01, *p<0.05.

Table A 2 in the Appendix give results from models that correct for measurement error in the predictors. The results, as we predicted, are very similar.

Group mobility from fall to spring of kindergarten

Table 7 summarizes mobility in group placement between fall and spring. By spring, about a third of math students and nearly half of reading students had moved groups—more often upward than downward. There was more mobility in classrooms with more groups, where the differences between groups are smaller, and there are more groups to move to. For example, in classrooms with 2 reading or math groups, only 10 to 20 percent of students changed groups between fall and spring, but in classrooms with 5 groups, more than half of students changed groups.

Mobility in reading groups

Why did students change groups between fall and spring? Did students move up by improving their test scores and behaviors, relative to other students, or was mobility predicted by students' SES, race/ethnicity, and gender? We addressed this question by regressing spring group placement, controlling for fall group placement, on student demographics and changes in test scores and behaviors. Statistics come from the fall-spring sample.

Table 8 gives the results for reading in the fall-spring sample.

In reading, score gains were one of the strongest predictors of upward mobility. Teachers tended to promote children with above average reading gains. Relative to a similar student with average reading gains, a student whose reading scores improved by 1 SD more than other students could expect to move up in reading group placement by 0.12 SD, 3.5 percentile points, or 0.34 logits. These coefficients changed very little when behavioral variables were added to the model.

Improvements in approaches to learning were even more predictive of upward mobility than reading gains. Relative to an otherwise similar students, a student whose approaches to learning improved by 1 SD, according to their teacher, could expect to move up in reading group placement by 0.2 SD, 5 percentile points, or 0.5 logits.

After gains in reading scores and approaches to learning, though, the next strongest predictor of reading group mobility was SES. Compared to lower-SES students with similar test scores and

behaviors, students with higher SES were more likely to get promoted into a higher reading group. This occurred in addition to the fact that higher-SES children received higher initial placements than their scores and behaviors alone would predict. Relative to a similar student whose SES was average for their classroom, a student whose SES was 1 SD above average could expect to move up by 0.1 SD, 2.5 percentile points, or 0.3 logits in reading group. The SES coefficients changed very little when behavioral variables were added to the model.

The next strongest predictor was Hispanic ethnicity. Relative to white students with similar score gains, similar changes in behavior, and similar SES, Hispanics were less likely to move up and more likely to move down. Compared to similar children of other ethnicities, Hispanic children's chances of promotion were lower by 0.1 SD, 3 to 4 percentile points, or 0.1 to 0.2 logits. The Hispanic coefficients were significant in the linear models but not in the logit models.

The weakest significant predictors were changes in interpersonal skills and internalizing problem behaviors. The coefficients for internalizing problem behaviors was negative, suggesting that students whose internalized problems worsened were slightly less likely to move up. The coefficient for interpersonal skills was also negative, suggesting that students who improved their interpersonal skills were less likely to move up. This result may seem somewhat surprising, unless teachers were reluctant to promote gregarious children who distracted their classmates with friendly chatter. In any case, the coefficients for interpersonal skills and internalizing problem behaviors were quite small, suggesting that changes in these behaviors predict group placement by only about 0.07 SD, 2 percentile points, or 0.1-0.2 logits.

Mobility in math groups

Table 9 gives results for mobility between math groups. The results are quite different than they were for reading groups. Variables that were significant predictors of reading group mobility, including score gains and SES, are not significant predictors of math group mobility. This is not just because the math sample is smaller than the reading sample, so that larger coefficients are needed to achieve statistical significance. The coefficients of score gains and SES are not just less significant but substantially smaller for math than they were for reading.

There are two exceptions, two ways in which the results for math group mobility resemble those for reading group mobility. First, like the results for reading mobility, the results for math mobility show a small and sometimes significant negative coefficients for Hispanic ethnicity. Compared to otherwise similar white students, Hispanic students' chances of promotion were lower by 0.1 SD, 3 to 4 percentile points or 0.8 logits, with or without adjustment for behavior. Second, like the results for reading mobility, the results for math mobility suggest a small and sometimes effect of improvements in approaches to learning. Net of other variables, a child whose approaches to learning improve by 1 SD more than other children, can expect to move up by 0.1 SD, 3 percentile points, or 0.5 logits. The coefficients for approaches to learning are marginally significant at best ($p<.10$), but they are comparable in size to the coefficients in the reading mobility model, and might have achieved statistical significance in the math mobility model if the sample were larger.

***p<0.001, **p<0.01, *p<0.05.

The mobility models in Table 8 and Table 9 were estimated on the full fall-spring sample. The next two table
the restricted fall-spring sample.

Table A 3 and Table A 4 in the appendix give results from the same model estimated on the restricted fall-spring sample. The results are very similar.

CONCLUSION

Is kindergarten ability group placement biased?

Is kindergarten ability group placement biased? Our results suggest that it is. While test scores remain far and away the best predictor of initial group placement in fall—as they should be—girls, Asian Americans, and high-SES students get higher placements than their test scores alone would justify. When children change groups between fall and spring, bias is present again. High-SES children are more likely to move up, and Hispanic children are less likely to move up, than their score gains alone would justify. Biases appear to be larger in reading than in math.

Not every group that we might expect to receive biased placement does. African American children, in particular, do not appear to receive biased group placement. Although African

Americans do receive, on average, lower placement than white students in the same class, when we adjust for test scores and SES, African Americans are placed in approximately the groups that their scores would predict. While this finding may seem surprising, past results on black children's ability group placements have been mixed, and ours is not the first study to find no sign of bias. In fact, the study that was most like ours—an analysis of the ECLS-K:1999 with classroom fixed effects coded by mean-centering—also found no bias against African Americans (Tach & Farkas, 2006). Hispanic children also did not receive biased placement in the fall, but there was evidence of bias against moving them up between fall and spring.

The biases are not enormous, but they are not trivial, and they can accumulate. In fall reading groups, for example, girls are placed 3.6 percentile points higher, Asian Americans are placed 4.2 percentile points higher, and high-SES children are placed 2.5 percentile points higher than their scores alone would predict. By spring, high-SES children have moved up by an additional 2.5 percentile points, on average, relative to average-SES children with similar scores and score gains. Putting these estimates together, high-SES Asian-American girls are placed, on average, 10.3 percentile points higher in fall, and 12.8 points higher in spring, than white boys of average SES with the same test scores in the same classrooms. This example is somewhat extreme, since high-SES Asian American girls constitute less than 3 percent of the sample, but it is a real example which illustrates what can happen when several coefficients, which are individually not terribly large, are combined.

How much do these biases affect group placement? The answer depends, to some degree, on the number of groups in the classroom. In a classroom with 4 or 5 groups, the difference between groups is 20 to 25 percentile points, and a bias of 10.3 to 12.8 percentile points is approximately half of that. In other words, for a high-SES Asian-American girl whose scores alone are not enough to put her in the highest group, there is about a 50 percent chance that she will be in a higher reading group than an average-SES white boy with the same test scores. In a classroom with just two groups, though, the difference between groups is 50 percentile points, and there would be a smaller chance—20 to 25 percent—that a high-SES Asian American girl whose scores alone would put her in the low group would be placed in the high group instead.

Skills other than reading and math can predict ability group placement as well, but the interpretation of this finding is debatable because we rely on teachers to report children's behaviors, and teacher reports may themselves be biased. That said, one of the most predictive skills is "attentional focus," which is provocative, because attentional focus (or its opposite, attention deficit) is one of the only skills that has been shown to predict later reading and math achievement, when earlier reading and math achievement are controlled (Duncan et al., 2007). Are teachers then correct to give students with poor attentional focus lower group placements than their scores would otherwise warrant? Or are lower placements one of the reasons that these students fall behind?

Whether the patterns documented here constitute bias is open to debate. Our results show that gender, SES, and ethnicity predict group placement after test scores are taken into account. This certainly looks like bias, but there could be a benign explanation for it. We have accounted for the most obvious confounders, in the form of classroom behaviors, and we have also accounted for random measurement error in test scores, behaviors, and SES. Yet it is always possible that other, unobserved confounders might explain away the apparent bias. On the other hand, it is possible that we have understated the bias by controlling for student behaviors, because those behaviors are reported by teachers, and teachers might display some of the same biases when they evaluate student behaviors as when they place students in ability groups.

Possible mechanisms of bias

If the patterns do indeed represent bias, what are some of the social or psychological mechanisms that might explain the bias? The data offer little opportunity to test mechanisms directly, but there are several possibilities, some of which seem more compatible with the results than others.

One possibility, motivated by the conflict perspective introduced earlier, is that teachers assign students to groups in ways that reinforce or reproduce the advantages of children from already dominant social groups. Biases toward high-SES children and against promoting Hispanic children are consistent with conflict theory, but our other findings are not. If teachers were trying to reproduce the social order, then why would they give girls higher placements than boys with similar

test scores? Why would they give Asian Americans higher placements than white students with similar scores? And why wouldn't they give white students higher placements than black students with similar scores?

Another possible mechanism is that teachers engage in "in-group favoritism," favoring students who resemble themselves (Jhangiani, et al., 2014). This might explain the bias in favor of girls. Many studies of gender bias examine organizational settings where men have the power to make decisions, but in kindergarten classrooms 98 percent of teachers are women (according to the ECLS-K:2011). In experimental settings, women appear at least as likely as men to display in-group gender favoritism when given the chance (Rudman & Goodwin, 2004; Lynch et al., 2018).

In-group favoritism might also explain the bias in favor of high-SES children, since teachers tend to have higher than average SES, at least as measured by their educational attainment. Since about three-quarters of teachers are white non-Hispanics, in-group favoritism might also explain their reluctance to promote Hispanic children, but it would fail to explain the bias toward Asian Americans, or the lack of bias against African Americans,

Another possible mechanism is that teachers, having limited information about their students' actual abilities at the start of kindergarten, engage in "statistical discrimination" (Phelps, 1972), giving higher placements to members of groups who are higher scoring on average—which at the start of kindergarten includes Asian Americans, high-SES children, and (in reading) girls (von Hippel et al., 2018). Statistical discrimination would be "accurate" (Bohren et al., 2019) if groups' average group placements were consistent groups' average test scores. Empirically, though, it appears that statistical discrimination in group placements is "inaccurate," since girls, Asian Americans, and high-SES children receive higher average placements than their average scores would warrant.

Alternatively, it may be that teachers place students more or less appropriately—that is, consistently with actual reading and math ability—but then succumb to lobbying by Asian American and high-SES parents who believe that their children belong in a higher group. This explanation agrees with evidence that high-SES parents are more involved in and more likely to intervene at their child's school (Lareau, 1989; Cheadle & Amato, 2011), but it contradicts evidence that Asian

American parents are less involved in their child's schools, on average, than white parents with similar SES and family structure (Cheadle & Amato, 2010).

Methodological issues

In the introduction, we highlighted several methodological concerns about past studies. Our results show that some of these concerns mattered more than others. The coding of ability group placement as a dependent variable appeared to have little effect on the results. We got very similar results whether we coded group placement as an ordinal variable or treated it as a continuous variable which we transformed into a standard score z or a percentile score q. That said, there are other approaches to modeling ability group placement that we would not recommend. In particular, we would not recommend dichotomizing group placements, which discards informative variation. And because the number of ability groups varies across classrooms, we would also not recommend modeling group placements as a continuous variable without some kind of transformation.

The inclusion of classroom fixed effects was more important. Models of ability group placement clearly require classroom fixed effects because ability group placement is entirely a within-classroom process. Models that omit classroom fixed effects confound within- and between-classroom variation, and the between-classroom variation can bias the results. In previous research on the ECLS-K:1999, models that did include classroom fixed effects sometimes returned different results than models that did not. One study that omitted classroom fixed effects alleged bias against African American students (Condron, 2007), but a study that included classroom fixed effects found none (Tach & Farkas, 2006). In our own study, we also noticed that omitting classroom fixed effects could change some results. Our final writeup did not include results without classroom fixed effects, because those results were incorrect, and we had limited space.

Implications for learning and policy

What are the consequences of biased ability group placement for student learning? Empirically, this question is beyond the scope of our analyses, but if higher group placement accelerates learning (Gamoran, 1992; Lleras & Rangel, 2009; Oakes & Lipton, 1990), we might

expect biased placement to accelerate the learning of high-SES students, Asian Americans, and girls, while suppressing the learning of boys and Hispanic children. However, the effect of ability grouping on learning is contentious. Concerns have been raised about the methods used in older studies (Betts & Shkolnik, 2000), and a recent summary of meta-analyses concluded that students in lower and higher groups benefited equally (Steenbergen-Hu et al., 2016).

If placement in higher groups does accelerate learning on average, it is not clear whether the benefits extend to children who, because of bias, are placed in higher groups than their scores would predict. According to a conflict perspective, such misplaced students should benefit because the instruction, materials, and peers in higher groups are more conducive to learning. But according to a functional perspective, the learning of misplaced students will suffer because the instruction and materials in their group are poorly matched to their actual achievement level.

Even if biased placement does not affect learning, it may be that bias in group placements affects how students see themselves and others. White, Asian-American, female, and high-SES students make up a disproportionate share of higher reading and math groups, and that would be true even if ability group placement were based entirely on measured achievement. Biases in group placement exaggerate these social differences between ability groups, and may exaggerate students' ideas about differences between ethnic groups, genders, and social classes, and where each student belongs.

The simplest way to avoid bias in ability group placement would be not to use ability grouping at all. When ability grouping is used, one way to avoid biases would be to base placements and mobility exclusively on objective criteria, such as scores on a formative test given in fall, winter, and spring (cf. Gamoran 2011).

REFERENCES

Allen, M.J., & Yen, W. M. (2002). Introduction to Measurement Theory. Long Grove, IL: Waveland Press.

Allison, P. D. (2009). *Fixed Effects Regression Models* (1st ed.). Sage Publications, Inc.

Baetschmann, G., Staub, K. E., & Winkelmann, R. (2017). Consistent estimation of the fixed effects ordered logit model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*,

685–703. https://doi.org/10.1111/rssa.12090@10.1111/(ISSN)1467-985X.TOP_SERIES_A_RESEARCH

Betts, J. R., & Shkolnik, J. L. (2000). Key difficulties in identifying the effects of ability grouping on student achievement. *Economics of Education Review*, *19*(1), 21–26. https://doi.org/10.1016/S0272-7757(99)00022-9

Bohren, J. A., Haggag, K., Imas, A., & Pope, D. G. (2019). *Inaccurate statistical discrimination* (No. w25935). National Bureau of Economic Research.

Catsambis, S., Mulkey, L. M., Buttaro, A., Steelman, L. C., & Koch, P. R. (2012). Examining Gender Differences in Ability Group Placement at the Onset of Schooling: The Role of Skills, Behaviors, and Teacher Evaluations. *The Journal of Educational Research*, *105*(1), 8–20. https://doi.org/10.1080/00220671.2010.514779

Chamberlain, G. (1979). *Analysis of covariance with qualitative data* (No. w0325). National Bureau of Economic Research.

Cheadle, J. E., & Amato, P. R. (2011). A quantitative assessment of Lareau's qualitative conclusions about class, race, and parenting. *Journal of family Issues*, *32*(5), 679-706.

Cirino, P. T., Chin, C. E., Sevcik, R. A., Wolf, M., Lovett, M., & Morris, R. D. (2002). Measuring socioeconomic status: reliability and preliminary validity for different approaches. *Assessment*, 9(2), 145-155.

Cohen, Jacob. (1988). *Statistical Power Analysis for the Behavioral Sciences. Routledge.*

Condron, D. J. (2007). Stratification and Educational Sorting: Explaining Ascriptive Inequalities in Early Childhood Reading Group Placement. *Social Problems*, *54*(1), 139–160. https://doi.org/10.1525/sp.2007.54.1.139

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Sexton, H. (2007). School readiness and later achievement. Developmental psychology, 43(6), 1428. Duncan, G. J., & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. In G. J. Duncan & R. J. Murnane, *Whither Opportunity?: Rising Inequality, Schools, and Children's Life Chances* (pp. 47–69). Russell Sage Foundation.

Fisher, D. (2018). *ADMETAN: Stata module to provide comprehensive meta-analysis*. Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458561.html

Gamoran, A. (1989). Rank, Performance, and Mobility in Elementary School Grouping. *The Sociological Quarterly*, *30*(1), 109–123. https://doi.org/10.1111/j.1533-8525.1989.tb01514.x

Gamoran, A. (1992). Synthesis of research: Is ability grouping equitable? *Educational Leadership*, *50*, 11–17.

Gamoran, A. (2010). Tracking and inequality. *The Routledge international handbook of the sociology of education*, 213-228.Greene, W. H. (1999). *Econometric Analysis* (4th ed.). Prentice Hall.

Gamoran, A. (2011). Designing instruction and grouping students to enhance the learning of all: new hope or false promise?. In Frontiers in sociology of education (pp. 111-126). Springer, Dordrecht.

Gresham, F. M., & Elliott, S. N. (1990). Social skills rating system: Manual. American Guidance Service.

Haller, E. J. (1985). Pupil Race and Elementary School Ability Grouping: Are Teachers Biased Against Black Children? *American Educational Research Journal*, *22*(4), 465–483. https://doi.org/10.3102/00028312022004465

Haller, E. J., & Davis, S. A. (1980). Does socioeconomic status bias the assignment of elementary school students to reading groups? *American Educational Research Journal*, *17*(4), 409–481.

Hallinan, M. T., & Sorenson, A. B. (1983). Effects of race on assignment to ability groups. In *The Social Context of Instruction: Group Organization and Group Processes* (pp. 85–103). Academic Press.

Jean, M. (2016). *Can you "work your way up?"—Ability grouping and the development of academic engagement* [Ph.D. thesis, The University of Chicago]. https://search.proquest.com/docview/1837431780/abstract/5614B121957A4545PQ/1

Jhangiani, R., Tarry, H., & Stangor, C. (2014). "Ingroup Favoritism and Prejudice." Section 11.2 in *Principles of social psychology*, 1st international edition. https://opentextbc.ca/socialpsychology/chapter/ingroup-favoritism-and-prejudice/

Lleras, C., & Rangel, C. (2009). Ability grouping practices in elementary school and African American/Hispanic achievement. *American Journal of Education*, *115*(2), 279-304.

Lockwood, J. R., & McCaffrey, D. F. (2020). Recommendations about estimating errors-in-variables regression in Stata. *The Stata Journal*, *20*(1), 116-130.

Lynch, R., Wasielewski, H., & Cronk, L. (2018). Sexual conflict and the Trivers-Willard hypothesis: Females prefer daughters and males prefer sons. *Scientific reports*, *8*(1), 1-13.

National Center for Education Statistics. (2010). Fall 2010 Kindergarten Teacher Questionnaire (Child Level), Early Childhood Longitudinal Study, Kindergarten Class of 2010-11. U.S. Department of Education.

NCS Pearson. (1990). *Social skills rating system*. NCS Pearson.

Oakes, J., & Lipton, M. (1990). Tracking and ability grouping: A structural barrier to access and achievement. In *Access to knowledge: An agenda for our nation's schools* (pp. 187–204). College Entrance Examination Board.

Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Stluka, M. F. (1994). Ability-group effects: Instructional, social, or institutional? *Sociology of Education*, *67*, 27–46.

Phelps, Edmund S. (1972). "The Statistical Theory of Racism and Sexism". *American Economic Review*. 62: 659–661.

Putnam, S. P., & Rothbart, M. K. (2006). Development of Short and Very Short Forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment*, *87*(1), 102–112. https://doi.org/10.1207/s15327752jpa8701_09

Rist, R. C. (1970). Student Social Class and Teacher Expectations: The Self-Fulfilling Prophecy in Ghetto Education. *Harvard Educational Review*, *40*(3), 411–451.

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. Child development, 72(5), 1394-1408.

Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men?. *Journal of personality and social psychology*, 87(4), 494.

Saleh, M., Lazonder, A. W., & De Jong, T. (2005). Effects of within-class ability grouping on social interaction, achievement, and motivation. *Instructional Science*, *33*(2), 105-119.

Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K–12 students' academic achievement: Findings of two second-order meta-analyses. *Review of Educational Research*, *86*(4), 849-899.

Tach, L. M., & Farkas, G. (2006). Learning-related behaviors, cognitive skills, and ability grouping when schooling begins. *Social Science Research*, *35*(4), 1048–1079. https://doi.org/10.1016/j.ssresearch.2005.08.001

Tieso, C. L. (2003). Ability Grouping Is Not Just Tracking Anymore. *Roeper Review*, *26*(1), 29–36. https://doi.org/10.1080/02783190309554236

Najarian, M., Tourangeau, K., Nord, C., and Wallner-Allen, K. (2018). Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), Kindergarten Psychometric Report (NCES 2018-182). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved July 1, 2020 from http://nces.ed.gov/pubsearch.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L., & Najarian, M. (2018). *User's Manual for the ECLS-K:2011 Kindergarten–Third Grade Data File and Electronic Codebook, Public Version* (NCES 2018034; p. 316). National Center for Education Statistics, U.S. Department of Education. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018034

von Hippel, P. T., & Hamrock, C. (2019). Do Test Score Gaps Grow Before, During, or between the School Years? Measurement Artifacts and What We Can Know in Spite of Them. *Sociological Science*, *6*(3). http://dx.doi.org/10.15195/v6.a3

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in Reading and Math Skills Forms Mainly before Kindergarten: A Replication, and Partial Correction, of "Are Schools the Great Equalizer?" *Sociology of Education*, *91*(4), 323–357. https://doi.org/10.1177/0038040718801760

von Hippel, P.T. (2015). Linear vs. logistic probability models: Which is better, and when. *Statistical Horizons*. https://statisticalhorizons.com/linear-vs-logistic

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data* (1st ed.). The MIT Press.

TABLES

Table 1. Sample restrictions. Each sample inherits the restrictions of the samples above it.

a. Reading

Analytic sample	Restrictions	Students	Classrooms	Teachers	Schools	PSUs
Full ECLS-K:2011	None	15,088	3,238	3,057	851	97
	Restrict to classrooms that use ability groups, according to fall questionnaire T1.	7,447	1,794	1,712	590	96
	Restrict to children with ability group numbers on fall questionnaire T1.	7,366	1,779	1,697	586	96
Fall sample	Restrict for consistency between child group number and number of groups in classroom, according to questionnaire T1.	7,220	1,768	1,686	585	96
	Require child to be in the same classroom at rounds 1 and 2.	7,202	1,766	1,685	584	96
	Restrict to classrooms that use ability groups in spring, according to spring questionnaire T2.	6,880	1,691	1,611	568	96
	Restrict to children with ability group numbers on spring questionnaire T2.	6,869	1,690	1,610	568	96
	Restrict for consistency between the child group number and the number of groups in the classroom, according to questionnaire T2.	6,748	1,686	1,606	568	96
Fall-spring sample	Require number of groups in classroom to agree between questionnaires T1 and T2.	4,675	1,168	1,117	499	95
Restricted fall-spring sample	Require number of groups in classroom to also agree with questionnaire A2.	3,502	840	800	420	91

b. Math

Analytic sample	Restrictions	Students	Classrooms	Teachers	Schools	PSUs
Full ECLS-K:2011	None	15,088	3,238	3,057	851	97
	Restrict to classrooms that use ability groups, according to fall questionnaire T1.	2,685	671	656	342	83
	Restrict to children with ability group numbers on fall questionnaire T1.	2,656	662	647	337	83
Fall sample	Restrict for consistency between child group number and number of groups in classroom, according to questionnaire T1.	2,607	660	645	337	83
	Require child to be in the same classroom at rounds 1 and 2.	2,595	659	644	336	83
	Restrict to classrooms that use ability groups in spring, according to spring questionnaire T2.	2,019	500	490	277	81
	Restrict to children with ability group numbers on spring questionnaire T2.	2,015	499	489	276	81
	Restrict for consistency between the child group number and the number of groups in the classroom, according to questionnaire T2.	2,002	499	489	276	81
Fall-spring sample	Require number of groups in classroom to agree between questionnaires T1 and T2.	1,355	337	332	217	74
Restricted fall-spring sample	Require number of groups in classroom to also agree with questionnaire A2.	964	240	236	174	69

Table 2. Fall sample. Used to analyze initial group placement.

	Reading sample	Math sample
Sample size		
Students	7,220	2,607
Classrooms	1,768	660
Teachers	1,686	645
Schools	585	337
Gender		
Female	52%	52%
Male	48%	48%
Race		
White, Non-Hispanic	50%	42%
Black/African American, Non-Hispanic	14%	16%
Hispanic	23%	29%
Asian, Non-Hispanic	6%	7%
Other	6%	6%
Number of ability groups in classroom		
2	5%	28%
3	29%	36%
4	40%	26%
5 (or more)	26%	10%

Note: Percentages refer to students rather than classrooms, teachers, or schools. For example, in the reading sample, 5 percent of students were in classrooms with 2 ability groups. Measures of SES and teacher-reported behavior are not summarized because they were all standardized to have a mean of 0 and a standard deviation of 1.

Table 3. Fall-spring sample. Used to analyze group mobility.

	Reading sample	Math sample
Sample size		
Students	4,675	1,355
Classrooms	1,168	337
Teachers	1,117	332
Schools	499	217
Gender		
Female	49%	48%
Male	51%	52%
Race		
White, Non-Hispanic	49%	41%
Black/African American, Non-Hispanic	15%	18%
Hispanic	26%	28%
Asian, Non-Hispanic	6%	6%
Other	6%	6%
Number of ability groups in classroom		
2	2%	23%
3	27%	39%
4	41%	24%
5 (or more)	30%	14%

Note. Percentages refer to students rather than classrooms. For example, in the reading sample, 2 percent of students in the reading sample were in classrooms with 2 ability groups. Measures of SES and teacher-reported behavior are not summarized because they were all standardized to have a mean of 0 and a standard deviation of 1.

Table 4. Transforming group numbers (k) into percentiles (q)

Group number (k)	Number of groups (K)			
	K=2	3	4	5
k=1	q=25	16.67	12.5	10
2	75	50	37.5	30
3		83.33	62.5	50
4			87.5	70
5				90
Mean of q	50	50	50	50
SD of q	25.0	27.2	28.0	28.3

Note: All calculations assume that the K groups are equal in size.

Table 5. Reading ability groups in fall: Predictors of initial group placement.

	Without teacher-reported behaviors			With teacher-reported behaviors		
	Linear models		Ordinal	Linear models		Ordinal
Predictors	Standardized	Percentile	logit	Standardized	Percentile	logit
Reading score	0.51***	14.45***	1.66***	0.46***	13.07***	1.57***
	(0.02)	(0.56)	(0.10)	(0.02)	(0.64)	(0.11)
Math score	0.30***	8.37***	0.81***	0.18***	5.07***	0.54***
	(0.02)	(0.55)	(0.08)	(0.02)	(0.60)	(0.08)
Socioeconomic status (SES)	0.08***	2.53***	0.27***	0.07***	2.02***	0.22***
	(0.01)	(0.37)	(0.04)	(0.01)	(0.37)	(0.05)
Female	0.13***	3.60***	0.31***	0.05*	1.41*	0.13
	(0.02)	(0.56)	(0.07)	(0.02)	(0.58)	(0.10)
Race/ethnicity (ref. white non-Hispanic)						
Black, non-Hispanic	0.01	0.15	0.02	0.03	0.54	0.02
	(0.05)	(1.20)	(0.13)	(0.04)	(1.22)	(0.14)
Hispanic	0.02	0.81	0.09	0.02	0.66	0.09
	(0.04)	(1.15)	(0.19)	(0.04)	(1.10)	(0.21)
Asian, non-Hispanic	0.17***	4.20**	0.66***	0.15**	3.41*	0.56**
	(0.05)	(1.36)	(0.18)	(0.05)	(1.41)	(0.19)
Other	0.00	-0.09	-0.09	0.03	0.21	0.07
	(0.06)	(1.88)	(0.22)	(0.07)	(1.99)	(0.24)
Teacher reported behaviors						
Approaches to learning				0.20***	5.23***	0.45***
				(0.03)	(0.93)	(0.11)
Self-control				-0.05+	-1.28+	-0.15
				(0.03)	(0.69)	(0.15)
Interpersonal skills				-0.01	-0.18	-0.03
				(0.03)	(0.75)	(0.12)
Externalizing problem behaviors				0.11***	2.74***	0.26*
				(0.03)	(0.70)	(0.10)
Internalizing problem behaviors				-0.07***	-1.70***	-0.22***
				(0.01)	(0.38)	(0.04)
Attentional focus				0.14***	4.43***	0.38***
				(0.02)	(0.62)	(0.08)
Inhibitory control				-0.00	-0.37	0.08
				(0.03)	(0.73)	(0.09)
Children	5,960	5,960		5,273	5,273	
Classrooms	1,364	1,364		1,290	1,290	
R^2 (within classrooms)	0.48	0.46		0.50	0.52	

***p<0.001, **p<0.01, *p<0.05. Test scores, SES, and teacher-reported behaviors were standardized. All models include classroom fixed effects. Standard errors were clustered at the level of the primary sampling unit, defined as either a large county or a group of adjacent and similar small counties. Sample sizes are smaller than in Table 2 because of missing values on test scores, SES, and teacher-reported behaviors.

Table 6. Math ability groups in fall: Predictors of initial group placement.

	Without teacher-reported behaviors			With teacher-reported behaviors		
	Linear models		Ordinal	Linear models		Ordinal
Predictors	Standardized	Percentile	logit	Standardized	Percentile	logit
Reading score	0.38***	10.76***	1.22***	0.33***	9.23***	1.13***
	(0.04)	(0.85)	(0.14)	(0.04)	(0.86)	(0.15)
Math score	0.42***	11.23***	1.28***	0.31***	8.22***	1.03***
	(0.04)	(0.98)	(0.16)	(0.04)	(0.99)	(0.19)
SES	0.08**	2.21***	0.30**	0.07*	1.89**	0.25*
	(0.02)	(0.58)	(0.10)	(0.03)	(0.62)	(0.12)
Female	0.09*	2.07*	0.20+	0.01	-0.07	-0.02
	(0.03)	(0.90)	(0.11)	(0.04)	(0.95)	(0.18)
Race/ethnicity (ref. white non-Hispanic)						
Black, non-Hispanic	0.05	0.81	0.24	0.13	2.52	0.66*
	(0.09)	(2.07)	(0.20)	(0.09)	(2.07)	(0.28)
Hispanic	-0.03	-1.05	-0.01	-0.05	-1.57	-0.11
	(0.06)	(1.56)	(0.38)	(0.06)	(1.60)	(0.52)
Asian, non-Hispanic	0.08	2.48	0.25	0.04	1.29	0.17
	(0.08)	(2.24)	(0.36)	(0.09)	(2.37)	(0.36)
Other	0.18*	4.22+	0.85	0.26**	6.44**	1.21*
	(0.09)	(2.20)	(0.62)	(0.09)	(2.31)	(0.60)
Teacher reported behaviors						
Approaches to learning				0.15*	4.43*	0.40
				(0.06)	(1.69)	(0.32)
Self-control				-0.05	-1.68	-0.12
				(0.04)	(1.22)	(0.22)
Interpersonal skills				0.02	1.15	0.21
				(0.05)	(1.32)	(0.23)
Externalizing problem behaviors				0.10**	2.44*	0.37**
				(0.03)	(0.95)	(0.14)
Internalizing problem behaviors				-0.04*	-0.97+	-0.15+
				(0.02)	(0.53)	(0.08)
Attentional focus				0.16***	4.00***	0.45**
				(0.04)	(1.15)	(0.17)
Inhibitory control				0.01	0.01	0.10
				(0.04)	(1.20)	(0.17)
Children	2,069	2,069		1,857	1,857	
Classrooms	483	483		460	460	
R^2 (within classrooms)	0.47	0.46		0.50	0.51	

***p<0.001, **p<0.01, *p<0.05. Test scores, SES, and teacher-reported behaviors are standardized. All models include classroom fixed effects. Standard errors are clustered at the level of the primary sampling unit, defined as either a large county or a group of adjacent and similar small counties. Sample sizes are smaller than in Table 3 because of missing values on test scores, SES, and teacher-reported behaviors.

Table 7. Group mobility between fall and spring

	Reading		Math	
	Children	%	Children	%
Mobility type	All classrooms			
Downward	800	17%	192	14%
None	2,714	58%	876	65%
Upward	1,161	25%	287	21%
Total	4,675		1,355	
	Classrooms with 2 groups			
Downward	3	3%	21	7%
None	101	89%	253	80%
Upward	9	8%	41	13%
Total	113		315	
	Classrooms with 3 groups			
Downward	137	11%	68	13%
None	842	68%	352	67%
Upward	268	21%	104	20%
Total	1,247		524	
	Classrooms with 4 groups			
Downward	313	16%	63	19%
None	1,140	60%	178	54%
Upward	460	24%	87	27%
Total	1,913		328	
	Classrooms with 5 groups			
Downward	347	25%	40	21%
None	631	45%	93	49%
Upward	424	30%	55	29%
Total	1,402		188	

Statistics come from the fall-spring sample.

Table 8. Reading ability groups in spring: Predicting mobility from fall.

	Without teacher-reported behaviors			With teacher-reported behaviors		
	Linear models		Ordinal	Linear models		Ordinal
Predictors	Standardized	Percentile	logit	Standardized	Percentile	logit
Fall group placement	0.67***	0.66***	1.66***	0.68***	0.67***	1.73***
	(0.02)	(0.02)	(0.12)	(0.02)	(0.02)	(0.16)
Reading gains	0.12***	3.52***	0.34***	0.10***	2.91***	0.33**
	(0.02)	(0.63)	(0.10)	(0.03)	(0.69)	(0.10)
SES	0.09***	2.48***	0.33***	0.09***	2.45***	0.32***
	(0.02)	(0.51)	(0.07)	(0.02)	(0.48)	(0.08)
Female	0.02	0.43	0.13	0.02	0.64	0.09
	(0.03)	(0.67)	(0.08)	(0.03)	(0.73)	(0.10)
Race (ref. non-Hispanic white)						
Black, Non-Hispanic	-0.07	-1.50	-0.02	-0.03	-0.66	0.06
	(0.07)	(1.69)	(0.20)	(0.06)	(1.70)	(0.21)
Hispanic	-0.09+	-3.45**	-0.12	-0.10*	-4.02***	-0.17
	(0.04)	(1.11)	(0.14)	(0.05)	(1.17)	(0.16)
Asian, Non-Hispanic	-0.09	-2.44	-0.34	-0.10+	-2.96*	-0.36
	(0.05)	(1.48)	(0.23)	(0.06)	(1.44)	(0.23)
Other	0.02	-0.03	0.03	-0.02	-0.96	-0.15
	(0.06)	(1.88)	(0.22)	(0.07)	(1.99)	(0.24)
Changes in teacher reported behaviors						
Approaches to learning				0.19***	5.25***	0.51***
				(0.03)	(0.74)	(0.13)
Self-control				-0.00	0.31	-0.05
				(0.03)	(0.62)	(0.10)
Interpersonal skills				-0.07**	-2.17***	-0.14
				(0.02)	(0.61)	(0.11)
Externalizing problem behaviors				0.01	0.54	-0.05
				(0.03)	(0.68)	(0.09)
Internalizing problem behaviors				-0.07***	-1.63***	-0.22***
				(0.02)	(0.41)	(0.06)
Attentional focus				0.06*	2.06*	0.08
				(0.03)	(0.80)	(0.12)
Inhibitory control				0.00	-0.02	-0.01
				(0.02)	(0.63)	(0.10)
Children	3,833	3,901		3,319	3,374	
Classroom	868	902		814	845	
R^2 (within classrooms)	0.48	0.49		0.50	0.52	

***p<0.001, **p<0.01, *p<0.05. Fall and spring group placement are measured in the same way, except in the ordinal logit model, where spring group placement is ordinal and fall group placement is standardized.

Table 9. Math ability groups in spring: Predicting mobility from fall

	Without teacher-reported behaviors			With teacher-reported behaviors		
	Linear models		Ordinal	Linear models		Ordinal
Predictors	Standardized	Percentile	logit	Standardized	Percentile	logit
Fall group placement	0.63***	0.61***	1.83***	0.65***	0.63***	2.14***
	(0.03)	(0.03)	(0.17)	(0.04)	(0.03)	(0.30)
Reading gains	0.02	1.06	-0.19	0.00	0.50	-0.13
	(0.06)	(1.57)	(0.17)	(0.07)	(1.77)	(0.26)
SES	0.06+	1.83*	0.14	0.04	1.29	-0.19
	(0.04)	(0.79)	(0.16)	(0.04)	(0.82)	(0.25)
Female	0.03	1.08	0.20	0.02	1.04	-0.03
	(0.05)	(1.23)	(0.40)	(0.05)	(1.29)	(0.42)
Race (ref. non-Hispanic white)						
Black, Non-Hispanic	-0.17	-4.58+	-0.27	-0.18	-5.61*	-0.60
	(0.11)	(2.49)	(0.44)	(0.12)	(2.81)	(0.51)
Hispanic	-0.11	-3.19	-0.80*	-0.14	-4.19+	-0.82*
	(0.10)	(2.34)	(0.38)	(0.10)	(2.37)	(0.37)
Asian, Non-Hispanic	0.12	2.18	7.99+	0.09	0.94	8.06*
	(0.11)	(2.80)	(4.30)	(0.10)	(2.54)	(3.73)
Other	0.05	-0.14	4.31*	-0.02	-2.11	3.65
	(0.10)	(2.54)	(2.11)	(0.10)	(2.56)	(2.60)
Changes in teacher reported behaviors						
Approaches to learning				0.08	2.85+	0.50**
				(0.07)	(1.67)	(0.19)
Self-control				0.03	0.44	-0.21
				(0.05)	(1.30)	(0.16)
Interpersonal skills				0.08	1.83	0.21
				(0.05)	(1.22)	(0.26)
Externalizing problem behaviors				0.05	1.49	-0.02
				(0.04)	(1.11)	(0.19)
Internalizing problem behaviors				-0.01	-0.62	-0.11
				(0.03)	(0.76)	(0.12)
Attentional focus				0.08+	2.70*	0.30+
				(0.05)	(1.14)	(0.17)
Inhibitory control				-0.04	-1.25	-0.22
				(0.05)	(1.22)	(0.16)
Children	1,056	1,101		939	982	
Classroom	241	260		228	247	
R^2 (within classrooms)	0.43	0.45		0.44	0.46	

***p<0.001, **p<0.01, *p<0.05. Fall and spring group placement are measured in the same way, except in the ordinal logit model, where spring group placement is ordinal and fall group placement is standardized.

APPENDIX

The linear models of initial group placement in Table 5 and Table 6 had measurement errors in the predictors. The next two tables re-fit these models with correction for measurement error.

Table A 1. Reading groups. Predictors of initial placement in fall sample. Linear models re-estimated with correction for measurement error (unreliability) in predictors.

	Without teacher-reported behaviors				With teacher-reported behaviors			
	if SES 91% reliable		if SES 81% reliable		if SES 91% reliable		if SES 81% reliable	
Predictors	Standardized groups	Percentile groups	Standardized groups	Percentile groups	Standardized groups	Percentile groups	Standardized groups	Percentile groups
Reading score	0.54***	15.25***	0.54***	15.19***	0.50***	14.09***	0.50***	14.04***
(95% reliable)	(0.02)	(0.67)	(0.03)	(0.55)	(0.02)	(0.58)	(0.02)	(0.67)
Math score	0.30***	8.39***	0.30***	8.34***	0.18***	4.96***	0.17***	4.93***
(92% reliable)	(0.02)	(0.68)	(0.03)	(0.69)	(0.02)	(0.66)	(0.03)	(0.63)
SES	0.08***	2.46***	0.09***	2.81***	0.07***	2.04***	0.08***	2.34***
(91% or 81% reliable)	(0.02)	(0.44)	(0.02)	(0.52)	(0.02)	(0.43)	(0.02)	(0.46)
Female	0.13***	3.66***	0.13***	3.68***	0.05*	1.47*	0.05*	1.49**
	(0.02)	(0.54)	(0.02)	(0.56)	(0.02)	(0.58)	(0.03)	(0.48)
Race/ethnicity (ref. white non-Hispanic)								
Black, non-Hispanic	0.03	0.35	0.03	0.43	0.03	0.56	0.03	0.62
	(0.04)	(1.36)	(0.03)	(1.23)	(0.04)	(1.14)	(0.04)	(1.10)
Hispanic	0.06	1.70+	0.06	1.81*	0.03	0.89	0.03	0.98
	(0.04)	(0.95)	(0.04)	(0.86)	(0.03)	(0.99)	(0.04)	(1.02)
Asian, non-Hispanic	0.19***	4.47***	0.19***	4.45***	0.15**	3.35*	0.15**	3.34*
	(0.06)	(1.26)	(0.04)	(1.26)	(0.05)	(1.52)	(0.05)	(1.33)
Other	0.03	0.11	0.03	0.19	0.03	0.18	0.04	0.25
	(0.05)	(1.29)	(0.05)	(1.24)	(0.05)	(1.17)	(0.04)	(1.19)
Teacher-reported behaviors								
Approaches to learning					0.18***	4.87***	0.18***	4.88***
					(0.02)	(0.91)	(0.03)	(0.67)
Self-control					-0.05+	-1.25+	-0.05*	-1.27*
					(0.03)	(0.67)	(0.02)	(0.57)
Interpersonal skills					-0.01	-0.17	-0.01	-0.19
(81% reliable)					(0.02)	(0.52)	(0.02)	(0.58)
Externalizing problem behaviors					0.10***	2.59***	0.10***	2.58***
(88% reliable)					(0.02)	(0.55)	(0.02)	(0.52)
Internalizing problem behaviors					-0.07***	-1.67***	-0.07***	-1.67***
(79% reliable)					(0.01)	(0.33)	(0.01)	(0.35)
Attentional focus					0.14***	4.24***	0.14***	4.23***
(87% reliable)					(0.02)	(0.60)	(0.02)	(0.63)
Inhibitory control					-0.00	-0.34	-0.00	-0.35
(87% reliable)					(0.02)	(0.73)	(0.02)	(0.59)
Children	5,273	5,273	5,273	5,273	5,273	5,273	5,273	5,273
R^2 (within classrooms)	0.47	0.49	0.47	0.49	0.51	0.53	0.51	0.46

***p<0.001, **p<0.01, *p<0.05.

Table A 2. Math groups. Predictors of initial placement in fall sample. Linear models re-estimated with correction for measurement error (unreliability) in predictors.

	Without teacher-reported behaviors				With teacher-reported behaviors			
	if SES 91% reliable		if SES 81% reliable		if SES 91% reliable		if SES 81% reliable	
Predictors	Standardized groups	Percentile groups	Standardized groups	Percentile groups	Standardized groups	Percentile groups	Standardized groups	Percentile groups
Reading score	0.37***	10.21***	0.36***	10.15***	0.32***	9.09***	0.32***	9.04***
(95% reliable)	(0.03)	(1.18)	(0.04)	(0.98)	(0.04)	(1.14)	(0.04)	(1.42)
Math score	0.47***	12.57***	0.47***	12.52***	0.36***	9.44***	0.36***	9.42***
(92% reliable)	(0.03)	(1.22)	(0.05)	(1.14)	(0.05)	(1.31)	(0.04)	(1.41)
SES	0.08***	2.40***	0.10***	2.76***	0.07**	1.91**	0.07*	2.19**
(91% or 81% reliable)	(0.02)	(0.66)	(0.03)	(0.67)	(0.02)	(0.69)	(0.03)	(0.72)
Female	0.10*	2.28**	0.10**	2.29**	0.02	0.15	0.02	0.17
	(0.04)	(0.79)	(0.03)	(0.87)	(0.03)	(1.03)	(0.04)	(0.88)
Race/ethnicity (ref. white non-Hispanic)								
Black, non-Hispanic	0.11	1.93	0.11	2.02	0.13*	2.64	0.13*	2.70
	(0.07)	(1.77)	(0.07)	(1.47)	(0.06)	(1.73)	(0.06)	(1.68)
Hispanic	0.00	-0.09	0.01	0.06	-0.03	-1.06	-0.03	-0.93
	(0.06)	(1.73)	(0.06)	(1.62)	(0.05)	(1.40)	(0.05)	(1.58)
Asian, non-Hispanic	0.09	2.36	0.09	2.35	0.04	1.28	0.04	1.27
	(0.09)	(2.05)	(0.07)	(1.93)	(0.08)	(2.35)	(0.08)	(2.06)
Other	0.25**	6.17**	0.25**	6.24***	0.26**	6.40**	0.26**	6.45**
	(0.08)	(2.19)	(0.09)	(1.69)	(0.08)	(2.03)	(0.08)	(2.21)
Teacher-reported behaviors								
Approaches to learning					0.13**	3.99**	0.13**	4.00***
					(0.05)	(1.30)	(0.05)	(1.12)
Self-control					-0.05	-1.60+	-0.05	-1.60
					(0.04)	(0.90)	(0.04)	(1.10)
Interpersonal skills					0.02	1.11	0.02	1.10
(81% reliable)					(0.04)	(0.93)	(0.04)	(1.07)
Externalizing problem behaviors					0.09***	2.30**	0.09**	2.28*
(88% reliable)					(0.02)	(0.71)	(0.03)	(0.97)
Internalizing problem behaviors					-0.04*	-0.91	-0.04+	-0.91+
(79% reliable)					(0.02)	(0.57)	(0.02)	(0.53)
Attentional focus					0.15***	3.84***	0.15***	3.83***
(87% reliable)					(0.04)	(1.17)	(0.03)	(1.00)
Inhibitory control					0.01	-0.06	0.01	-0.09
(87% reliable)					(0.04)	(1.03)	(0.04)	(1.20)
Children	1,857	1,857	1,857	1,857	1,857	1,857	1,857	1,857
R^2 (within classrooms)	0.48	0.49	0.48	0.49	0.51	0.52	0.51	0.52

***p<0.001, **p<0.01, *p<0.05.

The mobility models in Table 8 and Table 9 were estimated on the full fall-spring sample. The next two tables re-estimate the models on the restricted fall-spring sample.

Table A 3. Reading ability groups in spring: Predicting mobility from fall (restricted fall-spring sample)

	Without teacher-reported behaviors			With teacher-reported behaviors		
	Linear models		Ordinal	Linear models		Ordinal
Predictors	Standardized	Percentile	logit	Standardized	Percentile	logit
Fall group placement	0.68***	0.68***	1.66***	0.69***	0.69***	1.73***
	(0.02)	(0.02)	(0.12)	(0.02)	(0.02)	(0.16)
Reading gains	0.13***	4.04***	0.37***	0.13***	3.87***	0.37***
	(0.03)	(0.74)	(0.10)	(0.03)	(0.80)	(0.10)
SES	0.09***	2.33***	0.32***	0.08***	2.19***	0.30***
	(0.02)	(0.55)	(0.07)	(0.02)	(0.54)	(0.07)
Female	0.01	0.28	0.09	-0.00	0.23	0.06
	(0.03)	(0.72)	(0.09)	(0.03)	(0.81)	(0.10)
Race (ref. non-Hispanic white)						
Black, Non-Hispanic	-0.05	-0.29	-0.03	-0.02	0.54	0.05
	(0.07)	(1.85)	(0.20)	(0.07)	(1.99)	(0.21)
Hispanic	-0.07	-3.01*	-0.16	-0.08	-3.14*	-0.21
	(0.05)	(1.29)	(0.15)	(0.06)	(1.40)	(0.18)
Asian, Non-Hispanic	-0.12*	-3.74*	-0.37+	-0.11+	-3.73*	-0.38
	(0.06)	(1.76)	(0.21)	(0.06)	(1.74)	(0.23)
Other	-0.01	-0.42	0.04	-0.06	-1.73	-0.17
	(0.08)	(2.23)	(0.22)	(0.09)	(2.48)	(0.25)
Changes in teacher reported behaviors						
Approaches to learning				0.19***	5.35***	0.54***
				(0.04)	(0.96)	(0.13)
Self-control				-0.00	0.06	-0.02
				(0.03)	(0.73)	(0.10)
Interpersonal skills				-0.05+	-1.65*	-0.15
				(0.03)	(0.76)	(0.10)
Externalizing problem behaviors				-0.01	0.04	-0.05
				(0.03)	(0.76)	(0.09)
Internalizing problem behaviors				-0.07***	-1.46**	-0.21**
				(0.02)	(0.49)	(0.06)
Attentional focus				0.03	1.38	0.04
				(0.04)	(1.02)	(0.12)
Inhibitory control				-0.01	-0.27	-0.00
				(0.03)	(0.76)	(0.10)
Children	2,902	2,944		2,518	2,548	
Classroom	0.49	0.50		0.51	0.52	
R^2 (within classrooms)	645	667		608	627	

***p<0.001, **p<0.01, *p<0.05.

Table A 4. Math ability groups in spring: Predicting mobility from fall (restricted fall-spring sample)

	Without teacher-reported behaviors			With teacher-reported behaviors		
	Linear models		Ordinal	Linear models		Ordinal
Predictors	Standardized	Percentile	logit	Standardized	Percentile	logit
Fall group placement	0.67***	0.63***	1.80***	0.68***	0.66***	2.26***
	(0.04)	(0.03)	(0.17)	(0.04)	(0.04)	(0.39)
Reading gains	-0.02	0.33	-0.19	-0.02	0.02	-0.13
	(0.07)	(1.96)	(0.17)	(0.07)	(2.09)	(0.30)
SES	0.02	1.12	0.15	0.02	0.89	-0.20
	(0.04)	(1.00)	(0.16)	(0.05)	(1.07)	(0.28)
Female	0.03	1.28	0.18	0.02	1.42	-0.02
	(0.05)	(1.39)	(0.40)	(0.06)	(1.51)	(0.42)
Race (ref. non-Hispanic white)						
Black, Non-Hispanic	-0.21	-5.57	-0.28	-0.19	-5.66	-0.75
	(0.15)	(3.37)	(0.37)	(0.14)	(3.50)	(0.49)
Hispanic	-0.19+	-4.20	-0.78*	-0.19	-4.45	-0.76*
	(0.11)	(2.70)	(0.38)	(0.12)	(2.85)	(0.36)
Asian, Non-Hispanic	0.03	0.10	7.85+	-0.05	-1.91	8.07*
	(0.17)	(4.30)	(4.28)	(0.16)	(3.70)	(3.87)
Other	-0.04	-2.31	4.22*	-0.19	-6.11+	3.18
	(0.13)	(3.34)	(2.05)	(0.12)	(3.31)	(2.19)
Changes in teacher reported behaviors						
Approaches to learning				0.14*	4.15**	0.48*
				(0.05)	(1.35)	(0.19)
Self-control				-0.02	-0.59	-0.18
				(0.06)	(1.47)	(0.17)
Interpersonal skills				0.03	1.16	0.19
				(0.05)	(1.14)	(0.25)
Externalizing problem behaviors				0.01	0.26	-0.01
				(0.05)	(1.40)	(0.21)
Internalizing problem behaviors				-0.02	-0.42	-0.11
				(0.04)	(0.90)	(0.12)
Attentional focus				0.08	2.65*	0.23
				(0.05)	(1.32)	(0.18)
Inhibitory control				-0.11*	-2.49+	-0.24
				(0.05)	(1.42)	(0.16)
Children	755	789		671	703	
Classroom	0.47	0.48		0.48	0.49	
R^2 (within classrooms)	177	191		167	181	

***p<0.001, **p<0.01, *p<0.05.