EdWorkingPaper No. 20-204

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

Paul T. von Hippel
University of Texas, Austin

Ana P. Cañedo
University of Texas, Austin

Many kindergarten teachers place students in higher and lower "ability groups" to learn math and reading. Ability group placement should depend 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 their test scores alone would predict. One third of students move groups during kindergarten, and some movement is predicted by changes in test scores, but high-SES students move up more than score gains would predict, and Hispanic children move up less. Net of SES and test scores, there is no bias in the placement of African American children. Differences in teacher-reported behaviors explain the higher placement of girls, but do little to explain the higher or lower placement of other groups. Although achievement is the best predictor of ability group placement, there are signs of bias.

VERSION: February 2020

Suggested citation: von Hippel, Paul T., and Ana P. Cañedo. (2020). Is ability group placement biased? New data, new methods, new answers. (EdWorkingPaper: 20-204). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/d33y-nh50

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

ABSTRACT

Many kindergarten teachers place students in higher and lower "ability groups" to learn math and reading. Ability group placement should depend 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 their test scores alone would predict. One third of students move groups during kindergarten, and some movement is predicted by changes in test scores, but high-SES students move up more than score gains would predict, and Hispanic children move up less. Net of SES and test scores, there is no bias in the placement of African American children. Differences in teacher-reported behaviors explain the higher placement of girls, but do little to explain the higher or lower placement of other groups. Although achievement is the best predictor of ability group placement, there are signs of bias.

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

INTRODUCTION

Every fall, many teachers sort young children into higher and lower "ability groups" for instruction in math and reading. Ability grouping—less often called skill grouping or achievement grouping[1]—lets teachers differentiate instruction by giving more advanced material to higher groups and less advanced materials to lower groups. Unlike "tracking," which becomes common in middle and high school, ability grouping does not sort students into different classrooms. Instead, for some part of the day, ability grouping sorts children into different groups within the same classroom.

The social and educational purpose of ability grouping has been debated for decades. According to a benign, functional interpretation, ability grouping helps teachers give each group instruction that is neither too hard nor too easy, but tailored to their current achievement level, so that the practice maximizes each student's opportunity to learn (Tieso, 2003). According to a more nefarious, conflict interpretation, though, opportunity to learn is actually greater in the higher groups, and students in lower groups are denied the opportunity to realize their potential (Oakes & Lipton, 1990). In addition, ability grouping 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). The effects of labeling and denial of opportunity are particularly disturbing if the students in lower groups tend to come from historically disadvantaged racial and ethnic groups and families of lower socioeconomic status (SES).

[1] The phrase "ability grouping" may sound a little distasteful. The meaning of "ability" is ambiguous; in psychometrics "ability" is simply a synonym for a student's current skill level, but in other fields "ability" sometimes connotes something fixed or innate. We use the phrase "ability grouping" mainly 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.

A key question about ability grouping is which students get placed in higher groups and why. If ability group placement is benign and functional, then the assignment of students to math and reading ability groups should depend entirely on students' reading and math skills. But if group placement represents an attempt to advance students from higher-status families, 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 are correlated with SES and race/ethnicity, and reading skills are associated 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 good classroom behavior helps some children get higher group placements than their reading and math skills alone would merit. Whether this constitutes bias is subject to different interpretations. According to a benign, functional perspective, children who are attentive, respectful, non-disruptive, 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 more nefarious, conflict perspective, teachers' assessments of children's behavior are themselves biased, and the norms that are set for classroom behavior are biased as well. In that case, allowing teachers' assessments of behavior to influence ability group placement can be another mechanism for bias. Distinguishing these two explanations can be difficult in practice, since we often rely on teachers to report children's behaviors, and it is hard to know to what degree teacher reports are accurate or biased. Relying on parents' reports of behavior invites bias of a different kind.

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 (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' reading and math skills. Later, quantitative research revisited the question, using larger, more representative 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 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 studies, including the ECLS-K:1999, measured students' ability groups toward the end of the school year, in spring. This complicates interpretation of the association between test scores and group placement. In the fall, the association can only mean that test scores, or something correlated with them, have affected group placement. But by spring, it could also be that the groups students were placed in have affected their scores. One study avoided this problem by lagging scores by a year, using scores from the spring of kindergarten to predict group placement in the spring of first grade (Condron, 2007). Lagging scores by a year, though, may attenuate the relationship between scores and group placement, and this could exaggerate the predictive values of other variables that are correlated with test scores, including race, ethnicity, gender, and SES.

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 lacked 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 none have predicted mobility between the beginning and end of a single school year.

Studies predicting ability group placement also have several modeling issues. A key methodological issue is that any model of ability group placement must focus on within-class variation. Teachers decide group placements by comparing students within the same classroom, so analysis should be limited to within-classroom variation in predictors (test scores, SES, race, etc.). Between-classroom variation is irrelevant to teachers' decisions about group placement, and analyses that include between-classroom variation can produce misleading results.

To see the potential for bias, consider a completely segregated two-race school system in which black and white students never share a classroom. Because race does not vary within classrooms, it is impossible to know if any teacher would place a black student above or below a white student. An analysis of between-classroom variation will draw misleading conclusions. For example, it may conclude 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 including classroom fixed effects, which in linear models is equivalent to 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). 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.

How much might classroom fixed effects matter for models of 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 ability group placement was biased against black children (Condron, 2007). But another study of the ECLS-K:1999, which implicitly included classroom fixed effects through classroom centering, concluded that there was no bias against black children (Tach & Farkas, 2006).

Another methodological issue 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 converted group placements to a within-classroom standard score (Tach & Farkas, 2006) or quantile score (Condron, 2007; Gamoran, 1989). This approach 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 in both the fall and spring of kindergarten. We predict group placement in fall and then predict group mobility from fall to spring. Our models include classroom fixed effects, and we compare several different ways to code the dependent variable, including a new approach that preserves the variable's ordinal character and still allows the number of groups to vary across classrooms.

DATA

Sample and longitudinal design

Our data come from the relatively recent 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 kindergarteners. The design of the ECLS-K:2011 was a cluster sample, with children clustered in 851 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 or stratify by teacher or classroom, so within a school, the sampled students could come from several different classrooms. On average, the sample included 18 children per school, which worked out to approximately 4 per classroom. Six percent of teachers had two kindergarten classrooms, one that met in the morning and one that met in the afternoon.

The ECLS-K:2011 has followed children for six consecutive years, during which most children progressed from kindergarten through fifth grade. We limited our study to the fall and spring of the kindergarten year and to children who had the same teacher throughout kindergarten. Data were also collected in the fall and spring of first and second grade, but in those grades. the fall data were limited to a subsample comprising one-third of schools, which reduced statistical power available to identify significant predictors of ability group placement. After second grade, data were only collected in spring.

Ability groups and ability group placement

Three different questionnaires, filled out at different times during kindergarten, asked teachers about ability grouping. Teachers' answers did not always agree across questionnaires, and we dropped teachers who gave inconsistent answers. This left us with core analytic samples of 879 teachers who gave consistent answers about their reading ability groups, and 241 teachers who gave consistent answers about their math ability groups.

More specifically, the "Teacher Questionnaire (Child Level)," which teachers answered in both fall (Questionnaire T1) and spring (T2), asked the following question separately for reading and math:[2]

1. *"How many achievement groups…do you currently have in this child's class?"*

This question was multiple choice; possible answers were "none," 2, 3, 4, and "5 or more." We dropped teachers who answered "none" and teachers whose answers disagreed between the fall and spring questionnaires.

In addition, on the "Teacher Questionnaire" administered in spring (Questionnaire A2), teachers were asked the following question:[3]

2. *On days when you use achievement grouping, how many groups do you have in your class or classes? 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"….*

This question was free response, and teachers could answer with any one- or two-digit number. A few teachers gave an answer of 1, which makes no sense since achievement grouping requires

[2] 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.

[3] 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.

at least 2 groups, and a few teachers gave unrealistically high answers, as high as 33. We dropped teachers who gave answers that were less than 2 or greater than 6, as well as teachers whose answers to question 2 disagreed with their answers to question 1. We defined an answer of "5 or more" on question 1 as agreeing with an answer of 5 or 6 on question 2.

Two of the three questionnaires also asked about the group placement of individual students. Specifically, the fall and spring administrations of the "Teacher Questionnaire (Child Level)" (Questionnaires T1 and T2) asked teachers the following questions about both math and reading:[4]

> 3. *In which [achievement] group is this child currently placed? Use 1 for the highest achievement group.*

This question was free response; teachers could give any one- or two-digit number. We dropped teachers whose answers seemed impossible in light of the number of groups that they had reported using on question 2. For example, we dropped a teacher if they reported having only 2 reading groups, but placing a sampled child in reading group 3.

Test scores

Children took reading and math tests in the fall and spring of kindergarten. Tests followed a two-stage adaptive format. Children began by taking a "routing tests" with items of various difficulties; the results of the routing tests determined whether children would take an easy, medium, or hard second-stage test. Children's ability scores, or "theta" scores, were estimated using an item response theory (IRT) model with parameters to control for the difficulty, discrimination, and guessability of test items.

Test dates varied across schools, but fall tests were most often given in October, and spring tests were most often given in May. Between-school variation in test dates can cause

[4] On the fall questionnaire (Questionnaire T1), this was question 6 for reading and question 5 for math. On the spring questionnaire (Questionnaire T2), this was question 21 for reading and 23 for math.

trouble for some models, but our models eliminated between-school variation by using classroom fixed effects.

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.[5]

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 six 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

[5] Consider two standardized variables $Z1$ and $Z2$ each with mean 0 and variance 1 Their average $Z=(Z1+Z2)/2$ also has a mean of 0, but its variance is not one. Instead, the variance of the average is $\mathrm{Var}(Z)=(\mathrm{Var}(Z1)+\mathrm{Var}(Z2)+2\mathrm{Cov}(Z1,Z2)/4=(2+\mathrm{Corr}(Z1,Z2))/4$, which is less than 1—in fact less than ¾. So Z is not standardized.

standardized every scale to a mean of 0 and an SD of 1. The reliability of the scales ranged from 0.83 to 0.96 (Tourangeau et al., 2018).

The first two scales were as follows:

1. *Approaches to learning*. This scale, constructed specifically for the ECLS-K, 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).
2. *Attentional focus*. This scale, adopted from the Short Form of the Children's Behavior Questionnaire (Putnam & Rothbart, 2006), consists of six items that "measure the child's tendency to maintain attention on a task" (National Center for Education Statistics, 2010).

The last four scales come 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. *Self-control*.
4. *Interpersonal skills*. "The child interacted with others in a positive way" (Tourangeau et al., 2018).
5. *Externalizing problem behaviors*—e.g., "fighting, arguing, ...impulsiveness" (National Center for Education Statistics, 2010).
6. *Internalizing problem behaviors*—e.g., "depression, low-self-esteem" (National Center for Education Statistics, 2010).

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

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, then 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; in particular, *K* can take values of 2, 3, 4, or 5, 6. When *K*=2, it has a mean of 1.5 and an SD of 0.5, but when *K*=6, it has a mean of 3.5 and an SD of 1.7.

We tried two different methods for re-coding *k* so that it had the same mean and SD in different classrooms, at least approximately.

<u>Standardization</u>

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 $\bar{k}_c$ and dividing by the classroom standard deviation s_c. The result is a standardized variable *z* which, within each classroom,[6] has a mean of 0 and an SD of 1:

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

The ECLS-K:2011 sample typically did not include all the children in classroom *c*, and so we had to estimate $\bar{k}_c$ and s_c from a sample. In rare classrooms, the sample contained only one child; then s_c could not be estimated and *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.[7] The basic idea is this: If there are *K*=2 ability groups, you treat the lower group (*k*=1) as though the students in it are between percentiles 0 and 50 and the upper group (*k*=2) as though the students in it are 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.

[6] The total SD is somewhat less than the within-classroom SD because the between-classroom SD is constrained to zero.

[7] 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 1 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 the *K*. The SD increases with *K*, but only slightly; it rises from SD=25 when *K*=2 to SD=28.5 when *K*=6. Therefore *q* behaves much like a variable that has been standardized to a mean of 50 and a standard deviation somewhere between 25 and 28.5.

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—but this occurs very rarely in the ECLS-K:2011.

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

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

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[8], 3, 4, 5, and 6 groups, and then averaged the results[9] 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

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

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

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

[8] 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=3,4,5, or 6 groups.

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

and $\hat{\tau}^2$ is an estimate of how much the true coefficients β_K vary across the five regressions.[10]

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. 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. We centered variables in our linear models of the q and *z* transformations of ability groups.[11] Including fixed effects in ordinal logistic regression models is a little trickier, but several estimators have recently been developed and evaluated. We use the simple "blow up and cluster" estimator, which is known to provide consistent estimates (Baetschmann et al., 2017).

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

RESULTS

Sample description

Table 2 describes the analytic samples used for reading and math. In reading, the analytic sample had 3,920 students in 923 classrooms taught by 879 teachers in 449 schools. This

[10] 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 β_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 meta-analysis than in regression modeling.

[11] Stata's *xtreg, fe* command does this automatically.

analytic sample represented about one-quarter of the students in the ECLS-K:2011; the other three-quarters either did not use ability groups or had teachers whose answers to questions about ability groups were inconsistent. There were slightly more classrooms (923) than teachers (879) because some kindergarten teachers taught two classrooms—one in the morning and one in the afternoon. In math, the analytic sample was somewhat smaller, with 1,011 students in 245 classrooms taught by 241 teachers in 177 schools. Evidently, ability grouping was rarer in math than in reading.

In the reading sample, 86 percent of students were in classrooms that used 3, 4, or 5 groups. In math, the number of groups was typically smaller; 86 percent of students were in classrooms that used 2, 3, or 4 groups.

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 Asians and other students were relatively small, but we still had enough statistical power to show that being Asian influenced children's group placement—as we will show. Our sample also included variables measuring SES, teacher-reported behavior, and test scores, but Table 2 does not report summary statistics because all those variables were standardized to have a mean of 0 and an SD of 1.

Initial group placement in the fall of kindergarten

Table 3 and Table 4 predict the fall 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. In general, the results were quite similar across the three model specifications; the coefficients that used percentiles were approximately 25 times the coefficients that used standardized group numbers,[12] and the coefficients of the ordinal logit were approximately 3

[12] We predicted this. See our Methods for the relationship between group percentiles and standardized group numbers.

times the coefficients that used standardized group numbers. We will report all three results to offer different interpretations of effect size.

In general, the predictive accuracy of the models is moderately good; the linear models explain about half of the within-classroom variation in group placement. 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. In reading, when SES, gender, and race/ethnicity, and classroom were held constant, a one SD increase in reading scores predicted a 14 percentile point increase in group placement on the percentile score, a 0.5 SD increase on the standardized scale, and a 1.7 point increase on the logit scale. In math, likewise, a one SD increase in math scores predicted an increase of 9 percentile points, 0.33 SD, or 1 logit in group placement. Reading scores were more important than math scores in predicting reading group placement, but less important than math scores in predicting math group placement.

Net of test scores, there was evidence that ability grouping was biased toward higher-SES children. Although SES predicted group placement significantly, it was not nearly as strong a predictor as test scores. In predicting reading group placement, for example, SES was about one fifth as predictive as reading scores and about one quarter as predictive as math scores. Among children with similar test scores, gender, and race/ethnicity, a one SD increase in SES predicted an increase of just 2 percentile points, 0.08 SD, or 0.28 logits in reading group placement. Results for math were similar.

There was also evidence that ability grouping was biased in favor of girls. In fact, gender predicted group placement about as well as SES, though not nearly as well as test scores. Among children in the same classroom with the same test scores, SES, and race/ethnicity, girls were placed 3 percentile points, 0.11 SDs, or 0.38 logits higher than boys in reading, and 2.5

percentile points, 0.11 SD, or 0.48 logits higher than boys in math. So, in group placement, being female was comparable to being 1 SD higher on SES.

There was also evidence that ability group placement was biased in favor of Asian Americans. In fact, the bias toward Asian Americans was comparable to the bias toward girls. Compared to white children in the same classrooms, with the same test scores, and SES, Asian Americans' reading group placements were approximately 8 percentile points, 0.3 SD, or 1 logits higher. The difference between Asian American and white students is statistically significant in reading. In math, the difference was similar in size, but not statistically significant, perhaps because we had a smaller sample size of classrooms that used ability grouping in math than in reading.

There was little evidence of bias against African Americans. In reading, one of the three models had a coefficient for black children that was negative and significant (p<.05), suggesting a bias against African Americans, but the other two models had coefficients that were non-significant and positive, suggesting that any bias, if it existed, more likely favored African Americans. In math, all three black coefficients were positive, and one was borderline significant (p<.10), again suggesting that any bias was more likely to favor African Americans. Because African Americans tend to have lower test scores and lower SES, they do tend to get placed in lower ability groups, but net of test scores and SES, there is little evidence that their race per se influences group placement.

Likewise, there was little evidence of bias either for or against Hispanics or children of other races and ethnicities.

Models with teacher-reported behaviors

Adding teacher-reported behaviors increased explained variance by just 4 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. They 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). Also surprisingly, they gave slightly lower group placements to students who displayed better self-control. Students' interpersonal skills did not predict their reading group placements.

In predicting reading group placements, behavioral variables were comparable to demographic variables. Self-control, externalizing, and internalizing problem behaviors were each about as predictive as SES. Having strong approaches to learning was about three times as predictive as SES, and approximately as predictive as being Asian American or having high math scores.

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 a factor of three and became non-significant. 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 100 percent female themselves. Adding behavioral variables also halved the predictive value of math scores to predict reading group, and shrank other coefficients by 10 to 20 percent.

The model predicting math group placements had generally similar results, although some of the details were different. The coefficients for approaches to learning and attentional focus were similar to what they were in the reading model, but less statistically significant in math, presumably because of the smaller math sample. In math, interpersonal skills predicted slightly higher group placement, but higher self-control predicted lower group placement, as it did in reading. In math, unlike reading, externalizing and internalizing problem behaviors had no predictive value.

In math, as in reading, including behavioral variables made the gender coefficients turn non-significant—suggesting that girls' higher placement was due in part to their better behavior, as reported by teachers. Including behavior variables also made two of the coefficients for black race turn significant and positive—suggesting that black children were placed in slightly higher math groups than white children with similar test scores, SES, and behavior. This result is limited to the linear models, though. In the ordinal logit model, the coefficient for black race was also positive, but not statistically significant.

Group mobility from fall to spring of kindergarten

Table 5 summarizes mobility in group placement between fall and spring. Fall group placements are not, in general, permanent. About a third of students move groups between fall and spring, and more of them move upward than downward. There is more mobility in classrooms with larger numbers of groups. For example, in classrooms with 2 reading or math groups, only 10 to 20 percent of students change groups between fall and spring, but in classrooms with 6 groups, about 60 percent of students change groups. This makes sense since with more groups, the differences between groups are smaller, and there are more groups to move to. For example, in a class with 2 groups, the only possible movement is to go up from the lower group or down from the upper group, but in a class with 6 groups, any student can be moved to any of 5 other groups.

Models without teacher-reported behaviors

Why did students change ability groups between fall and spring? Did students move up by improving their test scores and behaviors, relative to other students, or was mobility influenced by students’ SES, race/ethnicity, and gender? Table 6 addresses this question for reading groups.

In reading, reading 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 0.14 SD, 4 percentile points, or 0.4 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. The Hispanic coefficients were not consistently significant in the models that omitted behavioral variables, but became consistently significant when behavior was included. With behavior controlled, being Hispanic reduced chances of promotion by 0.08 SD, 4 percentile points, or 0.3 logits.

The weakest significant predictors were changes in interpersonal skills and internalizing problem behaviors. The coefficients for internalizing problem behaviors was negative, suggesting, as expected, that students who increased their internalizing problems more than others were 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 are reluctant to promote children who distract their classmates with friendly chatter. In any case, the coefficients for interpersonal skills and internalizing problem behaviors are quite small, suggesting that changes in these behaviors affect group placement by only about 0.05-0.06 SD, 1.3-1.5 percentile points, or 0.16-0.17 logits.

Table 7 gives results for mobility between math groups. The results are quite different than they were for mobility between 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 substantially smaller in the model of math group mobility than they were in the model of reading group mobility.

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 but significant negative effect of Hispanic ethnicity. Net of other variables, being Hispanic reduced chances of promotion by 0.20 SD, 5 percentile points or 1.3 logits. Second, like the results for reading mobility, the results for math mobility suggest a small 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 0.10 SD, 3.3 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 achieve statistical significance in the math mobility model if the sample were larger.

CONCLUSION

Is 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 slightly higher placements than their test scores alone would justify. When children's 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.

The biases are not terribly large, but they are biases nonetheless. In the fall, being female or having high SES raises group placement by about 2 to 3 percentile points, and being Asian American raises group placement about twice as much. In the spring, being high SES raises group placement by an additional 2 to 3 percentile points, at least in reading, and being Hispanic lowers group placement by almost twice that amount.

Not every group that we might expect to suffer from bias does. African Americans, in particular, do not appear to be singled out for bias. Although African Americans do receive, on average, lower groups 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 we would expect. In mathematics, in fact, they may be placed a little higher. 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 similar to ours—an analysis of the ECLS-K:2011 with implicit classroom fixed effects—also found no bias against African Americans (Tach & Farkas, 2006).

Children's social and behavioral skills—especially attentional focus and approaches to learning—also affect ability group placement. Girls get higher ratings for social and behavioral skills than boys, and this explains why girls get higher placements than boys with the same reading and math scores. But social and behavioral skills do very little to explain why Asian and high-SES children get higher placements. Remember that we rely on teachers to report children's behaviors, and teacher reports may themselves be biased. So the fact that reported behaviors explain why girls get higher placements than boys does not necessarily mean that those placements are fair. It could be that teachers give girls higher average behavior ratings than they deserve.

The patterns of bias are only partially consistent with conflict theory, which would predict that children from dominant social groups would get higher placement than their test scores merit. The bias toward high-SES children and against Hispanic children is consistent with that prediction, but lack of bias against African Americans is not. The biases toward Asian Americans and girls are also hard to reconcile with conflict theory.

What are some of the social or psychological mechanisms that lead to biased group placement? There are several possibilities. It may be that kindergarten teachers, who are almost universally woman, favor girls who remind them of themselves. It may be that teachers have stereotypes about girls, Asian Americans, or high-SES children which lead teachers to overrate those children's abilities relative to what they have actually achieved in their young lives. It may also be that teachers place students more or less fairly, but then succumb to lobbying by Asian American and high-SES parents who believe their children belong in a higher group. While our data provide little opportunity to contrast potential mechanisms, it is provocative that Asian Americans receive higher-than-deserved group placements in reading, not just in mathematics. Given the common stereotype that Asians are good at math, if only stereotypes were at work, we would expect to see a stronger bias toward Asian Americans in math than in reading. The fact that we do not suggests that other mechanisms are at work.

In the Introduction, we highlighted several methodological concerns about past studies. Our results suggest that some of these concerns mattered more than others. In particular, 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 is a more vital methodological issue. 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. In particular, a 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.

What are the consequences of bias in ability group placement? To the degree that placement in higher groups accelerates learning (Gamoran, 1992; Oakes & Lipton, 1990), we might expect the practice to accelerate the learning of high-SES students, Asian Americans, and girls, while suppressing the learning of boys and Hispanic children. But the effect of ability grouping on inequality in learning is contentious; results are somewhat mixed (Slavin, 1987), and concerns have been raised about the quality of older studies (Betts & Shkolnik, 2000). In addition, 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 they deserve—groups where they are lower achieving than the other students, and some of the curriculum seems out of reach.

Even if bias does not directly affect the learning of the students who are affected by it, it may be that bias in group placements affect how students see themselves and others. Whites, Asian-Americans, girls, 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 ability. Biases in group placement exaggerate these social differences between ability groups, and may exaggerate students' ideas about achievement differences between ethnic groups, genders, and social classes, and where each student belongs.

REFERENCES

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

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

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., & 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.

Greene, W. H. (1999). *Econometric Analysis* (4th ed.). Prentice Hall.

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., The University of Chicago]. https://search.proquest.com/docview/1837431780/abstract/5614B121957A4545PQ/1

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.

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.

Slavin, R. E. (1987). Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis. *Review of Educational Research*, *57*(3), 293–336. https://doi.org/10.3102/00346543057003293

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

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

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

TABLES

Table 1. Transforming group numbers (*k*) into percentiles (*q*)

	Number of groups (*K*) in classroom				
Group number (*k*)	*K*=2	3	4	5	6
1	*q*=25	16.67	12.5	10	8.33
2	75	50	37.5	30	25
3		83.33	62.5	50	41.67
4			87.5	70	58.33
5				90	75
6					91.67
Mean of *q*	50	50	50	50	50
SD of *q*	25.0	27.2	28.0	28.3	28.5

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

Table 2. Characteristics of analytic samples in reading and math

	Reading sample	Math sample
Sample size		
Students	3,920	1,011
Classrooms	923	245
Teachers	879	241
Schools	449	177
Gender		
Female	49%	53%
Male	51%	47%
Race		
White, Non-Hispanic	51%	45%
Black/African American, Non-Hispanic	15%	18%
Hispanic	22%	25%
Asian, Non-Hispanic	6%	6%
Other	6%	6%
Number of ability groups in classroom		
2	3%	20%
3	25%	41%
4	42%	26%
5	19%	8%
6	11%	6%

Note: Percentages refer to students rather than classrooms, teachers, or schools. For example, in the reading sample, 3 percent of students were in classrooms with 2 ability groups.

Table 3. Reading ability groups: Predictors of initial group placement, fall of kindergarten

	Without teacher-reported behaviors			With teacher-reported behaviors		
	Linear models		Ordinal	Linear models		Ordinal
Predictors	Standardized	Percentile	logit	Standardized	Percentile	logit
Reading score	0.52***	14.72***	1.73***	0.48***	13.38***	1.71***
	(0.01)	(0.34)	(0.08)	(0.02)	(0.52)	(0.10)
Math score	0.30***	8.39***	0.97***	0.16***	4.71***	0.55***
	(0.02)	(0.55)	(0.23)	(0.03)	(0.70)	(0.10)
SES	0.08***	2.29***	0.28***	0.06***	1.84***	0.25***
	(0.01)	(0.19)	(0.08)	(0.01)	(0.21)	(0.07)
Child Race: Black, Non-Hispanic	0.03	0.28	-2.02*	0.05	0.77	0.09
(vs. white reference)	(0.05)	(1.82)	(0.89)	(0.06)	(1.97)	(0.21)
Child Race: Hispanic	0.04	1.63	-0.01	0.01	0.82	0.06
	(0.04)	(1.13)	(0.24)	(0.03)	(0.87)	(0.19)
Child Race: Asian, Non-Hispanic	0.28***	7.74***	1.00***	0.24***	6.26**	0.59**
	(0.04)	(1.25)	(0.20)	(0.05)	(1.99)	(0.19)
Child Race: Other	-0.00	-0.47	1.76**	0.05	0.67	0.15
	(0.03)	(1.03)	(0.65)	(0.04)	(1.46)	(0.20)
Child Gender: Female (vs. Male reference)	0.11**	3.00**	0.38***	0.04	0.88	0.19+
	(0.03)	(0.84)	(0.11)	(0.03)	(0.69)	(0.11)
Teacher reported behaviors						
Approaches to Learning				0.23***	5.87***	0.54***
				(0.04)	(1.02)	(0.12)
Self-Control				-0.09***	-2.44***	-0.30***
				(0.02)	(0.35)	(0.08)
Interpersonal skills				0.01	0.51	0.06
				(0.03)	(0.90)	(0.06)
Externalizing Problem Behaviors				0.08***	2.36***	0.21+
				(0.01)	(0.36)	(0.12)
Internalizing Problem Behaviors				-0.07**	-1.90***	-0.27***
				(0.02)	(0.40)	(0.04)
Attentional Focus				0.13***	3.83***	0.36***
				(0.02)	(0.63)	(0.09)
R^2 (within classrooms)	0.49	0.51		0.53	0.55	

****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—either a large county or a group of adjacent and similar small counties.

Table 4. Math ability groups: Predictors of initial group placement, fall of kindergarten

	Without teacher-reported behaviors			With teacher-reported behaviors		
	Linear models		Ordinal	Linear models		Ordinal
Predictors	Standardized	Percentile	logit	Standardized	Percentile	logit
Reading score	0.33**	9.37**	1.02***	0.30*	8.32**	1.32*
	(0.10)	(2.42)	(0.17)	(0.10)	(2.32)	(0.55)
Math score	0.48**	13.22**	2.04***	0.34**	9.26**	2.61*
	(0.12)	(3.22)	(0.47)	(0.09)	(2.53)	(1.20)
SES	0.11*	3.03*	0.69*	0.11*	3.00**	1.25
	(0.04)	(0.96)	(0.33)	(0.04)	(0.81)	(0.93)
Child Race: Black, Non-Hispanic	0.07+	0.42	0.86	0.16***	3.19***	1.24
(vs. white reference)	(0.03)	(1.63)	(0.62)	(0.04)	(0.58)	(0.80)
Child Race: Hispanic	0.10	2.98	0.87	0.09	2.51	1.74
	(0.09)	(2.40)	(0.62)	(0.05)	(1.41)	(1.39)
Child Race: Asian, Non-Hispanic	0.23	5.43	6.20+	0.23	4.79	4.56*
	(0.15)	(4.38)	(3.41)	(0.16)	(4.14)	(1.80)
Child Race: Other	0.24	6.55	4.49	0.26	7.49+	3.97
	(0.19)	(4.22)	(3.71)	(0.17)	(3.52)	(3.26)
Child Gender: Female (vs. Male reference)	0.11**	2.52**	0.48***	-0.04	-1.42	0.71
	(0.02)	(0.73)	(0.14)	(0.05)	(1.33)	(0.81)
Teacher reported behaviors						
Approaches to Learning				0.20+	5.19	0.53*
				(0.10)	(3.32)	(0.25)
Self-Control				-0.17***	-6.30***	-0.84*
				(0.02)	(0.88)	(0.42)
Interpersonal skills				0.10*	3.23**	0.71**
				(0.04)	(0.99)	(0.25)
Externalizing Problem Behaviors				0.01	-0.49	-0.18
				(0.02)	(0.82)	(0.76)
Internalizing Problem Behaviors				-0.02	-0.52	0.09
				(0.01)	(0.36)	(0.23)
Attentional Focus				0.12+	3.82+	0.59*
				(0.05)	(1.77)	(0.29)
R^2 (within classrooms)	0.49	0.49		0.54	0.55	

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

Table 5. Group mobility between fall and spring of kindergarten

	Reading		Math	
	Children	%	Children	%
Mobility type	**All classrooms**			
Downward	680	17%	133	13%
None	2,243	58%	654	65%
Upward	965	25%	220	22%
Total	3,888		1,007	
	Classrooms with 2 groups			
Downward	2	2%	12	6%
None	88	89%	165	80%
Upward	9	9%	28	14%
Total	99		205	
	Classrooms with 3 groups			
Downward	111	11%	49	12%
None	666	67%	278	68%
Upward	211	21%	81	20%
Total	988		408	
	Classrooms with 4 groups			
Downward	269	17%	50	19%
None	975	60%	145	55%
Upward	386	24%	67	26%
Total	1,630		262	
	Classrooms with 5 groups			
Downward	167	22%	16	21%
None	352	47%	43	57%
Upward	232	31%	16	21%
Total	751		75	
	Classrooms with 6 groups			
Downward	131	31%	6	11%
None	162	39%	23	40%
Upward	127	30%	28	49%
Total	420		57	

Table 6. Reading ability groups, spring kindergarten: Predictors of final group placement

	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***	0.67***	1.55***	0.69***	0.68***	1.64***
	(0.01)	(0.01)	(0.13)	(0.01)	(0.01)	(0.13)
Reading gains	0.14***	4.25***	0.38***	0.13***	3.89***	0.40***
	(0.03)	(0.88)	(0.05)	(0.03)	(0.82)	(0.06)
SES	0.09***	2.49***	0.28***	0.09***	2.33**	0.23***
	(0.01)	(0.41)	(0.02)	(0.02)	(0.61)	(0.07)
Child Race: Black, Non-Hispanic	-0.04	-0.64	-0.11	-0.01	0.06	0.08
(vs. white reference)	(0.06)	(1.47)	(0.11)	(0.07)	(1.73)	(0.16)
Child Race: Hispanic	-0.06	-3.06*	-0.15	-0.08*	-3.76*	-0.28***
	(0.04)	(1.14)	(0.12)	(0.04)	(1.22)	(0.08)
Child Race: Asian, Non-Hispanic	-0.08+	-2.99+	-0.30+	-0.07	-3.00	-0.33***
	(0.04)	(1.43)	(0.18)	(0.05)	(1.68)	(0.10)
Child Race: Other	0.04	0.57	-0.11	-0.00	-0.46	-0.24
	(0.05)	(1.48)	(0.12)	(0.06)	(2.08)	(0.18)
Child Gender: Female (vs. Male reference)	0.01	0.12	-0.03	-0.00	0.16	-0.05
	(0.01)	(0.49)	(0.12)	(0.02)	(0.39)	(0.12)
Increases in teacher reported behaviors						
Approaches to Learning				0.19***	5.21***	0.48***
				(0.04)	(0.77)	(0.08)
Self-Control				-0.01	-0.21	0.01
				(0.02)	(0.77)	(0.06)
Interpersonal skills				-0.05*	-1.53***	-0.17*
				(0.02)	(0.34)	(0.09)
Externalizing Problem Behaviors				-0.01	0.01	0.05
				(0.02)	(0.55)	(0.11)
Internalizing Problem Behaviors				-0.06***	-1.34***	-0.16**
				(0.01)	(0.27)	(0.05)
Attentional Focus				0.05	1.83+	0.12*
				(0.03)	(0.87)	(0.06)
R^2 (within classrooms)	0.49	0.50		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 7. Math ability groups, spring kindergarten: Predictors of final group placement

	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***	0.64***	2.04***	0.70***	0.67***	2.85**
	(0.04)	(0.05)	(0.29)	(0.05)	(0.05)	(0.91)
Math gains	-0.00	0.74	-0.01	-0.02	0.06	-0.27
	(0.03)	(0.87)	(0.16)	(0.02)	(0.74)	(0.23)
SES	0.02	1.13	-0.04	0.01	0.77	-0.27
	(0.02)	(0.72)	(0.20)	(0.02)	(0.62)	(0.28)
Child Race: Black, Non-Hispanic	-0.15	-3.71	-0.28+	-0.13	-3.80	-0.39
(vs. white reference)	(0.11)	(2.50)	(0.16)	(0.14)	(2.76)	(0.25)
Child Race: Hispanic	-0.20+	-4.79*	-0.91*	-0.20+	-4.86*	-1.29*
	(0.10)	(1.96)	(0.37)	(0.11)	(2.13)	(0.58)
Child Race: Asian, Non-Hispanic	0.08	1.35	5.73*	0.00	-0.55	5.45*
	(0.14)	(4.98)	(2.61)	(0.12)	(3.98)	(2.29)
Child Race: Other	-0.04	-1.69	3.07*	-0.18+	-4.96	2.49
	(0.05)	(1.40)	(1.40)	(0.10)	(3.23)	(1.91)
Child Gender: Female (vs. Male reference)	0.01	0.91	-0.05	-0.00	1.00	-0.26
	(0.06)	(1.41)	(0.31)	(0.07)	(1.46)	(0.45)
Increases in teacher reported behaviors						
Approaches to Learning				0.10	3.28+	0.46+
				(0.08)	(1.80)	(0.26)
Self-Control				0.00	-0.16	-0.18
				(0.05)	(0.76)	(0.23)
Interpersonal skills				0.03	0.72	0.27
				(0.04)	(0.69)	(0.32)
Externalizing Problem Behaviors				0.05	1.15	0.68
				(0.04)	(1.11)	(0.72)
Internalizing Problem Behaviors				-0.04	-0.86	-0.35
				(0.04)	(0.73)	(0.24)
Attentional Focus				0.06	2.42+	0.08
				(0.05)	(1.28)	(0.11)
R^2 (within classrooms)	0.48	0.49		0.49	0.50	

***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.