FEATURE

End the Birthday Bias

Age allowances in high-stakes tests are a proven boost for fairness

OLDER SCHOOLCHILDREN typically perform better on academic achievement tests than younger students in the same classes. Time and again, studies looking at an array of countries, grade spans, and subjects have found that age differences of even a few months do matter.

Consider this example of how age affects academic performance. I look at scores on a standardized math test in Mexico for students in grades 3–9 and group results by students' birth month (see Figure 1). In every grade, the oldest students, those born in January, perform better than their youngest classmates, who were born 11 months later, in December. These age-based differences mean that, in places with academic tracking, students who are older for their grade are more likely to end up in the more demanding and more academically oriented programs. In comprehensive systems, relatively older students are more likely to attend more selective institutions than younger students—particularly within disadvantaged groups.

Relative age introduces an arbitrary bias that favors older students. And while states and school systems in the United States have mostly ignored this problem, parents consistently step in to try to correct for this bias through "academic redshirting," or intentionally delaying kindergarten entry by a year (see "Is Your Child Ready for Kindergarten?," *features,* Summer 2017). Widespread worries about the practice inspired proposals in Illinois and New Jersey that would ban redshirting, which delays enrollments of an estimated 6 percent of kindergarten students nationwide.

One can only wonder, are relative-age effects on test scores a new trend? Or are they simply a new finding? It turns out that they are neither. These effects are a well-established fact as old as standardized testing itself—and they have been addressed head-on in the past. To see the path forward toward greater fairness in testing, we must first look back at its history.

A New "Mental Test"

On a Friday in June 1921, close to 3,000 students in the primary schools of rural Northumberland, England, were given a new test. It had been developed in the previous months for the purpose of measuring their intellectual abilities. The sheets with the answers were gathered the following Monday, and, two days later, they had all been graded.

Less than three decades later, more than one million such tests were given in Great Britain in 1949 alone. The mind behind this new measure of intelligence was Godfrey Thomson, a towering figure in psychology.

Born in England in 1881, Thomson was of modest means but attended top universities after winning multiple school scholarships based on competitive exams. He trained as a teacher and a scientist, and then entered the field of psychology

BY PABLO A. PEÑA

COLLAGE BY SARAH HANSON / PICTURE OF GODFREY THOMSON: CENTRE FOR RESEARCH COLLECTIONS, UNIVERSITY OF EDINBURGH when he took on the responsibility of training teachers at Armstrong College, Newcastle. One of his lecture topics was the measurement of intelligence.

Meanwhile, about 25 miles north of Newcastle, an intelligence-measurement challenge was vexing leaders in Northumberland. Officials were looking for a fair way to determine which 11-year-old primary-school students would earn what was then the privilege of free secondary-school education. Thomson was invited to help devise a solution.

"It was a problem which had a personal interest for me," Thomson explained in *A History of Psychology in Autobiography,* "for I would myself have had no education beyond the primary school had I not won a free place in a secondary school in a competitive examination."

Competitive examinations had been used to select the region's secondary-school students for years, and nearly all the spots went to students from a handful of schools near Newcastle. Students who attended primary schools in poor or isolated areas rarely scored high enough to earn a seat. Local educational authorities, who attributed the pattern to differences in students' home and school environments, sought a new type of test, one that would assess intelligence rather than academic achievement.

"But intelligence tests, it was hoped, might discover in those schools some children of potential secondary school ability even if their environment and their poorer primary schooling had handicapped them in the existing kind of examination," Thomson wrote.

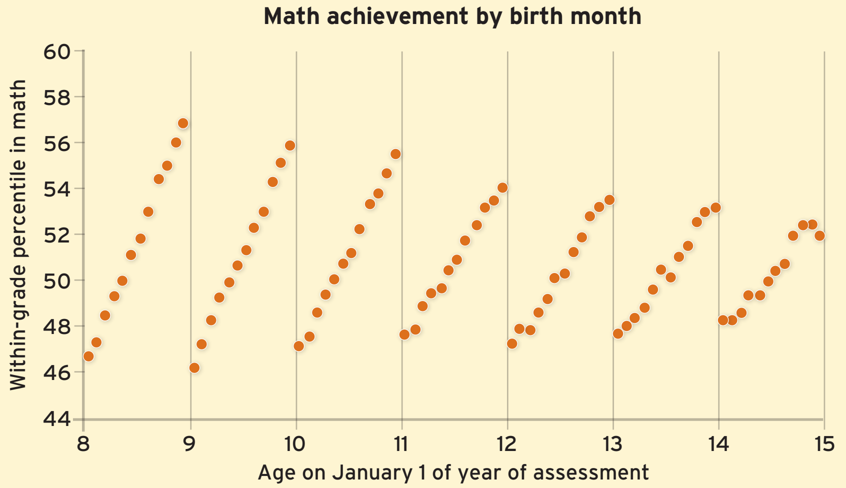
With this in mind, he created the Northumberland Mental Test to assess students' verbal and mathematical reasoning ability. Using its results, he selected about a dozen students and gave them free spots in secondary schools "as an experiment." Those students were followed through the years and, in Thomson's view, their performance justified the choice.

"Two, alas, died in an influenza epidemic, and two or three failed to complete a good secondary school course, though more I think for social and economic reasons than for lack of intelligence. Others, however, went on and did very well," he wrote. "Those Northumberland tests of mine were the beginning of a lifelong task, which I have felt bound to persevere in for the sake of intelligent children."

Word of Thomson's new exam spread, and soon he received requests from other regions in England to help them with secondary-school student selection. "For these they paid me fees," he recounted. "I decided that I would safeguard myself from the temptation to make money out of this activity, and I devised a committee to receive all these fees and royalties from my tests." By 1925, Thomson had become an educational psychologist at the University of Edinburgh, and the exams were known as the Moray House Tests. The revenues they generated went to research on standardized testing.

Older Students Perform Better Than Younger Classmates *(Figure 1)*

The oldest students in each grade, who were born in January, consistently earn higher scores on standardized tests than their youngest classmates, who were born 11 months later, in December.



NOTE: Average performance on a standardized math test by student age and birth month.

SOURCE: Author's calculations, based on 2013 data from the Evaluacion Nacional del Logro Academico en Centros Escolares (ENLACE), the national standardized test of Mexico. Includes scores from 1.9 million students in grades 3-9 in the State of Mexico, which is the most populous state in the country.

In 1932, a Moray House Test was given to practically all Scottish children born in 1921—roughly 90,000—as part of a national intelligence measurement effort called the Scottish Mental Survey. A similar survey was conducted in 1947 among children born in 1936. Those surveys are landmarks and have allowed researchers to study the relationship between intellectual abilities and other variables that include health behaviors, socioeconomic mobility, and life satisfaction. In 1949, Thomson was knighted by King George VI for his contributions to education.

Age Allowances in Action

Thomson's tests were designed to measure aptitude and to unravel the

Relative age introduces an arbitrary bias that favors older students. And while states and school systems in the United States have mostly ignored this problem, parents consistently step in to try to correct for these differences through academic "redshirting."

tight connection between school assignments and social status. But they also revealed another sort of advantage: the boost conferred by age in tests that attempt to measure the intellectual ability of schoolchildren. In order to make appropriate comparisons across students, it was necessary to account for age differences, even if they were of just a few months. So Thomson included a formula that adds or subtracts a few points for every month of age in how each student's Intelligence Quotient is calculated. This adjustment became known as the "age allowance," and it is based on the average increase in the test score that would result from the same student taking the test at a slightly different age.

An age allowance is a simple idea. Think of the growth charts that pediatricians use to assess the height and weight of a child, which track those measures by the child's age, in months. To create that chart, someone collected information from many children and recorded their exact age. With many observations, it is possible to compute an average score for every age in months. The age allowance is simply the adjustment for the trend in scores due to age.

It was also a pioneering development. In 1959, psychologist P. E. Vernon lauded Thomson's achievement, writing, "If one were asked to name one field in which Thomson was the undisputed pioneer who led the rest of the world, it would be the standardization, and application of age-corrections to, mental tests." Thomson "perfected the technique of determining the appropriate age correction for each month-group to which the test was applicable without having to collect enormous samples of each month of children."

This scoring method was not without controversy. In 1953, the British newspaper the *Guardian* (then called the *Manchester Guardian*) reported on some of the grievances parents had with respect to the process for determining which students would attend selective secondary schools, known as "grammar schools." One specific complaint was that the age allowance gave "below-average youngsters preference over above-average older pupils who, in the considered opinion of schoolmasters, would do better at grammar school." The newspaper explained:

> To this accusation the experts blandly plead guilty—while at the same time protesting that their age-allowance (which may be as much as twelve or fourteen per cent) is scrupulously fair and accurate. [...] Where, then, lies the catch? Simply in the fact that no allowance is made for age at any other stage in the schoolchild's career.

In other words, age allowances make admissions fairer, but students who benefit from them tend to do worse than those who don't. That is not because they are worse students; rather, it's because such allowances don't follow students into the classroom. Once they are admitted, students "thereafter take all internal and external examinations at the same time, and the younger would never again get an age allowance."

That insight applies today just as it did seven decades ago. Leveling the playing field in admissions doesn't erase the differences in test scores and GPA *after* admission. On average, younger students will still perform worse than their older classmates.

In this context, then, it is crucial to clarify the purpose of using tests scores in admissions. Is it to fairly select talented students or to predict which students will perform better? If what matters is "the accuracy with which [a test] predicts performance," the *Guardian* article continued, "no allowance should be made for age and admissions will be heavily weighted in favour of children born in the right months. But so long as admission to a grammar school is regarded as a privilege to be competed for, such a criterion would be manifestly unjust."

So long as admissions exams are intended to fairly apportion opportunities to talented students, age allowances are appropriate. In Thomson's words, "The object of an age allowance is not to improve prediction, but to do justice to children born in different months of the year."

Impacts on Equity

The questions raised by the *Guardian* article make many educational authorities reluctant to adopt age adjustments in test scores. Yet the broader point is that there is unfairness in all measures of academic performance that don't take into account age differences between classmates. Age-adjusting test scores used for admission purposes is a step in the right direction. But it doesn't address by itself the handicap suffered by younger students in later tests or grading.

Still, it's better to improve fairness in admissions even if the playing field is not leveled in other indicators of academic achievement. The fact that an institution, a school district, or a country cannot fix all the distortions introduced by relative age doesn't mean it shouldn't fix some of them. Partially fixing the problem is better than not fixing any of it. Plus, there is evidence of benefits from this approach.

In 1944, a sweeping set of new rules made important changes to broaden educational opportunity throughout England and Wales. The Education Act of 1944 raised the age of compulsory schooling to 15, made secondary schools free to all, and brought church-run schools into the national system. All students were required to take a competitive admissions exam after age 11. Many schools started using the Moray House Tests, which included an age allowance.

Economists Robert Hart and Mirko Moro analyzed how the enrollment of children into grammar schools changed as a result of the reform. Before 1944, children born from January to August—the middle or end of the school year—were less likely to find a grammar-school spot than their older classmates, who were born from September to December. After the reform, students born in the middle of the year were far more likely to get a grammar-school spot, which Hart and Moro argue was due, in part, to the growth in the use of age allowances. In other words, the adoption of age allowances increased the admission rates of students who, based on their month of birth alone, would have otherwise been excluded.

The modern-day relevance of these discoveries from the past century is not hard to find. Consider a test like the ones used by school districts in Boston, Chicago, or New York City to admit students into selective public high schools. If students who are 14 years and 11 months old on the day of the test score two points higher, on average, than students who are 14 years and 10 months old on test day, their ultimate test performance should account for that age-based difference. This applies to college admissions tests, as well—not only the SAT and ACT in the U.S., but also the Gaokao in China, Vestibular in Brazil, Suneung in South Korea, Exani in Mexico, and so on. Thomson's work shows that the creators and administrators of these tests can accurately measure what correct age allowances should be, based on the unique context of the exam and the students.

Hazards Ahead

If age allowances increase fairness and are feasible—proven by Thomson a century ago—shouldn't they be more popular? Why don't we see them in more education systems? First, the belief that age differences of a few months stop mattering early on in academic contexts is as widespread as it is incorrect. But I see another culprit as well. Even if some stakeholders are aware of the effects of relative age, there is a collective action problem.

No institution or school district operates in isolation, and many use the same or similar admissions exams. So adopting age allowances unilaterally may be a bad idea. Imagine that one selective school decides to make an "in-house" age allowance in its admission process while comparable institutions don't, but they all use the same test. The institution adopting the age allowance would experience a drop in unadjusted test scores. Of course, admissions to that institution would be fairer. But the average quality of incoming students as measured by test scores would look worse relative to both past incoming classes and peer institutions.

Thomson's tests were designed to measure aptitude and unravel the tight connection between school assignments and social status. But they also revealed another sort of advantage: the boost conferred by age.

Age allowances could hurt the ranking of an institution—a high price to pay in a hypercompetitive environment in which even prestigious institutions like Claremont McKenna College and Emory University have falsely inflated the average SAT scores of their incoming freshman classes to publications like *U.S. News & World Report* to boost their public profiles. Greater numbers of relatively young students would be admitted while greater numbers of relatively old students would be rejected, bringing the average unadjusted SAT scores down. Despite a growing movement toward "test-optional" admissions, average SAT scores remain a high-profile metric for many institutions, and any school that adopted age allowances would mechanically fall in college rankings. It's unlikely that any one institution, even if interested in fairness in admissions, would want to be the first to adopt age allowances.

However, not all stakeholders in the realm of standardized testing have the same interests and concerns. To overcome our collective action problem, we can make age allowances at the source. Test creators and test administrators don't face the tradeoff between fair admissions and institutional ranking. They also observe all test takers and are well-positioned to determine how big or small the "bump" to younger students should be. They can follow Thomson's lead and account for this by design.

To gauge the potential impact of introducing age allowances, we can look at recent test scores in England on two grammar-school admissions exams. Though these "11+" exams are used for admissions at all 160 grammar schools in England, different regions and schools use different tests. Not every test used includes an age allowance, despite the longstanding precedent to do so.

I look at average student scores on two tests: one administered by the University of Durham's Centre for Evaluation and Monitoring, which includes age allowances, and one administered by The Consortium of Selective Schools in Essex, which does not (see Figure 2). In the Essex schools using the unadjusted test, the youngest students, who were born in August, score roughly 0.2 standard deviations below the oldest students, who were born the previous September. By contrast, we do not see such differences on the age-adjusted tests.

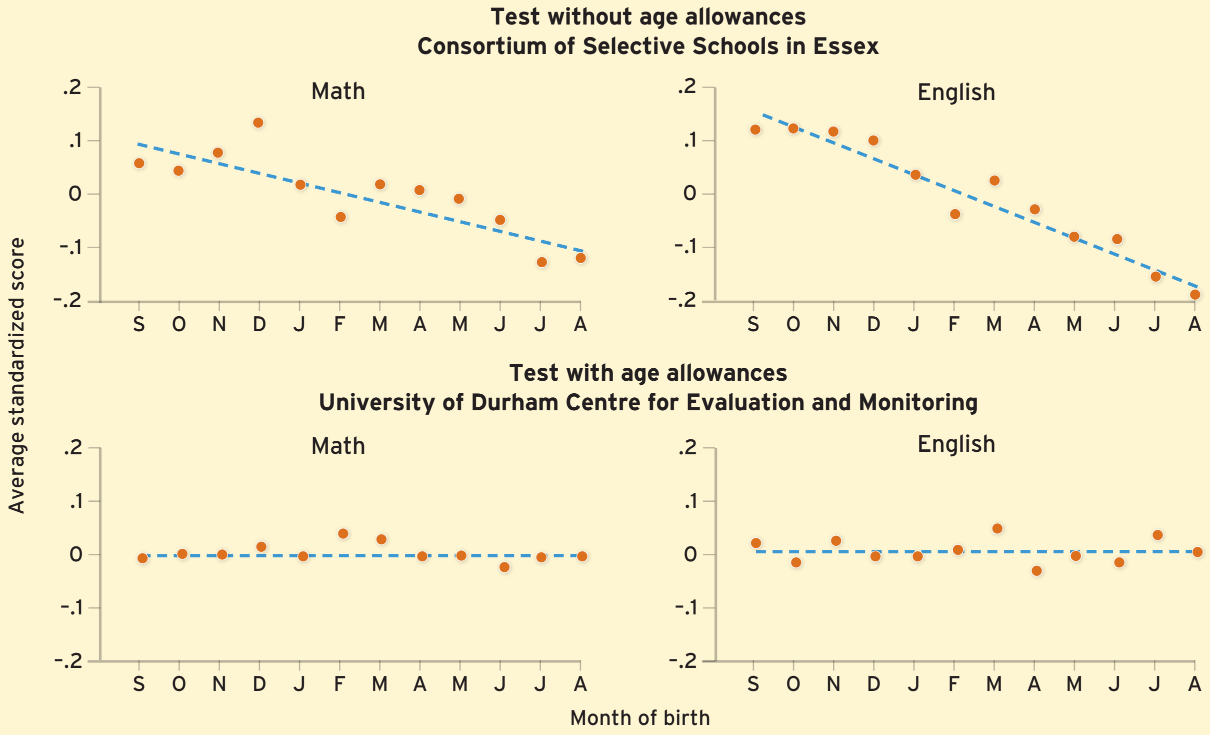
Admissions that use unadjusted scores obviously penalize students born in August relative to those born in September. But they also penalize students born in July, June, and so on, all the way to October, though to a lesser extent. Even in the same country and in the same admission process, not all schools are in the age-allowance wagon that departed Northumberland in the 1920s.

The Case for Extending Age Allowances

Age allowances have a proven track record and should be included in any test for which there is indication that age matters. There is clear evidence that age makes a difference in the measurement of intelligence until at least age 18. Just as important, there is also evidence of the effects of age on the SAT and ACT, the two most popular college admissions tests in the U.S. For example, a study by Steven Hemelt and

Age Allowances Reduce Unfairness in Admissions Tests *(Figure 2)*

Some grammar schools in England use admissions tests with age allowances while others do not. In schools using an unadjusted test, the youngest students (those born in August) on average score roughly 0.2 standard deviations below the oldest students (those born in September). Schools that use a test with age allowances do not have a relative-age effect.



NOTE: Ordinary least squares regression showing average test scores on two different grammar-school admissions tests in England by student birth month, from oldest (born in September, labeled "S") to youngest (born in August, labeled "A"). Data includes scores from 5,396 students on the CSSE 11+ test given by The Consortium of Selective Schools in Essex in 2017, which does not include age allowances, and scores from 52,879 students on the University of Durham Centre for Evaluation and Monitoring test in 2016, which includes age allowances.

SOURCE: Author's calculations

CENTRE FOR RESEARCH COLLECTIONS, UNIVERSITY OF EDINBURGH

Godfrey Thomson, a towering figure in psychology, created the Northumberland Mental Test to help assess students' math and verbal reasoning abilities.

Rachel Rosen found that 12 months of age bump scores on the ACT by as much as three percentiles. According to my preliminary analysis of the impact of age on SAT scores, students who retake the test one year after their first time gain about eight percentiles. To be sure, second-time testers may be more familiar with the SAT format and have undertaken more preparation than students sitting for the test the first time. But given the impacts of age on test scores we saw in Figures 1 and 2, the fact that they are one year older also would seem an important factor.

So long as admissions exams are intended to fairly apportion opportunities to talented students, age allowances are appropriate.

Recent moves by a growing group of prominent U.S. institutions to make standardized test scores an optional part of student applications won't make life easier for relatively young applicants. College admissions officers are focused on other signs of talent, and those signs are also biased by age. For example, one analysis of GPA among high-school seniors shows that relatively younger students are outperformed by their older classmates. To correct for this bias, age allowances could also be made in subject-specific grades as well as any academic achievement test whose score is used to award entry into competitive programs, compare performance, or give feedback to students and families.

Age allowances could also reduce academic redshirting by removing families' incentive to delay kindergarten. This isn't a minor point. At a societal level, redshirting is a wasteful practice. Essentially, it is a zero-sum game, since there will always be younger and older children in the same school class. Equally important, since redshirting is more prevalent among white children from high-income families, it contributes to the gaps in test scores observed along income and racial or ethnic lines. By making redshirting less appealing, age allowances could simultaneously save resources and help level the playing field—a rare chance to enhance efficiency and equity at the same time.

An age allowance is neither a new nor a radical idea. Allowances are as old as standardized tests themselves, and they were born with the measurement of intellectual ability of children. And, above all, including them in high-stakes measures of intellect or academics is the fair thing to do. In the words of Thomson, "Age allowances are sometimes, by those opposed to them, called a premium on youth. They are not that. When scientifically applied they are a device to compensate for the unfair premium on age."

Pablo A. Peña, who was born in January and started school a year early, is Assistant Instructional Professor at the Kenneth C. Griffin Department of Economics at the University of Chicago.