Gifted Child Quarterly
2017, Vol. 61(3) 229–238
© 2017 National Association for Gifted Children
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0016986217701834
journals.sagepub.com/home/gcq
SAGE

Should Millions of Students Take a Gap Year? Large Numbers of Students Start the School Year Above Grade Level

Scott J. Peters[1], Karen Rambo-Hernandez[2], Matthew C. Makel[3], Michael S. Matthews[4], and Jonathan A. Plucker[5]

Abstract

Few topics have garnered more attention in preservice teacher training and educational reform than student diversity and its influence on learning. However, the actual degree of cognitive diversity has yet to be considered regarding instructional implications for advanced learners. We used four data sets (three state-level and one national) from diverse contexts to evaluate how many students perform above grade level in English Language Arts and mathematics. Results revealed that among American elementary and middle school students, 20% to 49% in English Language Arts and 14% to 37% in mathematics scored 1 year or more above grade level. We address what these findings imply for K-12 schools, grouping decisions, and educational policies that strive to foster advanced abilities.

Keywords

grouping, diversity, age/developmental stage, differentiation, achievement

Student diversity, in all its forms, has long been an important consideration in both the classroom teaching and in education policy. Students differ on a number of educationally relevant dimensions including prior educational preparation, aptitude, motivation, family background, social skills, and personality, so teachers and schools work with a diverse range of needs in every classroom. Yet America's K-12 education systems heavily rely on placing students into grade levels by age, based on the implicit assumption that students of a given age have similar enough needs that all will benefit from similar learning experiences. America's schools are built almost entirely on this age-based, grade-level grouping: it undergirds standards, instruction, curriculum, and assessment. Having grade-level grouping as the basis for so many aspects of education reinforces the implicit message that getting students to grade level is the primary purpose of schooling.

At the federal level, the No Child Left Behind Act and its reauthorization, the Every Student Succeeds Act, codify the focus on grade-level standards and expectations and, in some ways, may prohibit or penalize states that attempt to move away from a focus on grade-level proficiency (e.g., Neal & Schanzenbach, 2010; Polikoff, 2016). Such an emphasis on "grade level" assumes that a large percentage of students are performing near grade level, while ignoring any possibility of individual differences. If an arbitrarily low proportion of students—say 5%—were capable of working with higher level curriculum and meeting standards for a higher grade level, then a concerted effort at differentiation and special programming likely could prove sufficient to meet the academic needs of these above-grade-level learners. However, if larger numbers of students exceeded grade-level expectations, then, the underlying assumptions about the suitability of both age-based grades and current services for advanced students would be called into question. Although some research on the variability of student learning outcomes exists (e.g., Firmender, Reis, & Sweeny, 2013), only limited attention has been paid to the actual range of learning readiness in grade-level classrooms (e.g., Herrnstein, 1971).

The purpose of this study was to answer the following question: At the beginning of the school year, how many students already perform 1 or more years *above* grade level? This foundational question has profound importance for the arguments that undergird general education, for gifted education programs, and for how we craft education policy.

[1]University of Wisconsin – Whitewater, Whitewater, WI, USA
[2]West Virginia University, Morgantown, WV, USA
[3]Duke University, Durham, NC, USA
[4]University of North Carolina at Charlotte, NC, USA
[5]Johns Hopkins University, Baltimore, MD, USA

Corresponding Author:
Scott J. Peters, University of Wisconsin, 800 W. Main Street, Whitewater, WI 53190, USA.
Email: peterss@uww.edu

Literature Review

The Context for Learning and Development

Among the most influential learning theories of the 20th and 21st centuries is Vygotsky's Zone of Proximal Development (ZPD). A key tenet of Vygotsky's ZPD is that for humans to develop (learning being part of the larger concept of development), they need to interact with content that is just beyond what they can complete independently (Vygotsky, 1997). However, academic achievement tests typically measure what students can do or perform on their own. On such tests, student performance is assessed based on the items they successfully can complete in the absence of any additional support. Vygotsky (1997) posited that instruction targeted at this level would not result in student growth or development because students already have mastered this material; in his words (as translated), "learning which is oriented toward developmental levels that have already been reached is ineffective . . . [because] it does not aim for a new stage of the developmental process but rather lags behind this process" (p. 34). Similarly, Vygotsky (1997) suggested that instruction targeted too far above a student's current performance also would fail to generate growth because the learner simply is not ready for it; it would be like trying to comprehend a television show in a foreign language one has never heard before. Instead, instruction should be targeted in a "Goldilocks" ZPD, in which the new material to be learned falls just beyond what students already are able to complete on their own.

We were interested in the degree to which schools appear faithful to Vygotsky's concept of ZPD—what data there were to show that students likely were being taught within their respective ZPDs. Engel, Claessens, and Finch (2012) sought to better understand this issue by comparing the level of math mastery for students entering kindergarten with data on what content was taught over the course of the school year. Using nationally representative Early Childhood Longitudinal Study–Kindergarten data, the authors found that 95% of students entering kindergarten already had mastered concepts such as identifying one-digit numbers, recognizing some geometric shapes, and counting up to 10. Nothing about this near-uniform level of mastery was out of the ordinary, as these skills are fairly basic. However, when teachers were asked how many days per month they spent teaching these concepts, the average was 13 days. Despite the vast majority of students having demonstrated prior mastery, over half the month was spent on teaching these tasks. Furthermore, this type of repetitive instruction of concepts already mastered was shown to be negatively associated with end-of-year mathematics test scores. Put into Vygotskian language, receiving instruction below students' ZPD prevented them from achieving further development in math. Alternatively, when these students were taught content that 25% or fewer of them had mastered (such as place value and currency or addition and subtraction), their associated learning gains in mathematics were positive and statistically significant.

Connor et al. (2009) also noted the importance of the interaction between child and instruction. The authors conducted a randomized field trial to determine whether closely aligning the amount and type of instruction to a child's particular needs resulted in greater reading gains. Instruction that was aligned to student needs resulted in uniformly greater learning gains; the authors also found that the degree of alignment was critical. The more precisely classroom instruction was targeted to the students' needs, the larger its effect. Connor et al. (2009) found that the close alignment between instruction and readiness levels had a greater impact on student learning than teacher qualifications, teacher professional development, or school characteristics. Such findings support Vygotsky's ZPD as a major requirement for effective learning and development.

The Engel et al. (2012) and Connor et al. (2009) studies and others like them carry two important and related implications. First, Vygotsky's theory seems to explain their findings. When students are taught content they already know, they do not grow in their mastery of the material. Second, if further development for all students is the goal of a K-12 education, then all students need to be taught as closely as possible to their individual Zones of Proximal Development. Although doing this should seem fairly straightforward and intuitive, its actual implementation can pose a logistical challenge. The precise targeting of instruction toward student need and readiness is clearly a good thing, but leaves us to wonder if this is possible in a typical or even ideal classroom. Before the feasibility of such instruction can be considered (much less implemented), educators first need to understand how many students have ZPDs that fall outside of the instructional range that is typically provided at each grade level.

Why Does Academic Variability Matter?

As demonstrated by Engel et al. (2012), misalignment between instruction and students' ZPD can result in no gain or even a loss of learning. The lesser the alignment between students' learning needs and the instruction provided, the greater the likelihood that these students will fail to develop to their full potential. Every classroom will have some degree of variation in students' ZPDs, necessitating what has come to be called differentiated instruction. A long-standing and related concern in gifted education circles is that many school administrators and general education teachers believe that advanced learners will be fine on their own—that being retaught grade-level content below gifted students' ZPDs is not a concern because these students somehow will learn anyway.

In an early study by Evertson, Sanford, and Emmer (1981), the authors reported that greater heterogeneity of student readiness and mastery with regard to English content significantly affected the classroom teacher's ability to adapt

instruction to meet the needs of all students. Consistent with the Connor et al. (2009) findings, this suggests that greater student heterogeneity likely makes it harder to adjust instructional levels precisely for all students. Evertson et al. (1981) also found that learner heterogeneity was associated with lower levels of class engagement. Intuitively, this makes sense because the teacher is trying to address a wider range of needs—meaning that given finite resources and teacher time, no single level of need receives as much attention. In examining individual case studies, Evertson et al. (1981) found that teachers focused most of their effort on the lowest achieving students, even though this led to average- and high-achieving students being less likely to show learning gains. Overall, these studies indicate that heterogeneity of student readiness (i.e., classrooms in which students have a broader range of prior mastery) makes effective instruction more challenging to deliver.

Findings of a more recent national survey of teacher practices (Farkas & Duffett, 2008) also indicated that struggling students received by far the greatest amount of teacher attention (63% reported), in comparison with teacher attention devoted to average students (13%) or to advanced students (7%). This was in spite of the fact that 50% of teachers expressed the belief that students at all levels should receive equal attention. What is even more telling is that when teachers were asked which students in their classrooms were most likely to receive instruction specially designed for their abilities, 51% of teachers indicated the lowest performing students, while only 10% noted advanced students. The research seems to suggest that the wider the range of student learning needs, the more a teacher must choose which students receive attention and which do not, and the more likely it becomes that the advanced students' learning needs will remain unaddressed.

Variability will always exist within classrooms, even those made up of a single age-based grade level of students. In 1971, Herrnstein noted the not-uncommon finding of the students in a third-grade class having achievement levels spanning second to eighth-grade norms in reading. In 5th-grade math, this range covered 3rd through 10th-grade norm levels. More recently, Firmender et al. (2013) also documented within-grade heterogeneity in reading comprehension and fluency across elementary classrooms, with some students performing 2 to 3 years *below* grade level and others performing 6 to 7 years *above* grade level. Despite having widely varied ZPDs in reading, these students were placed in the same classroom because they happened to be the same age. As we alluded to earlier, if such levels of heterogeneity are the exception and not the rule, then age-based instruction based solely on grade-level grouping likely is appropriate. However, the data provided by these prior studies suggest that this is not the case.

It seems important to reiterate that according to Vygotsky's ZPD, all students need to be exposed to content that is targeted based on their current performance. If this alignment is not closely tailored, further development will be hindered. It is also worth pointing out that these effects primarily occur in the academic realm; the ZPD does not consider the affective, attitudinal, or motivational side effects that may arise from a lack of challenge, which could be even more concerning. Given the great theoretical and practical influences of classroom heterogeneity on learning and development, it is essential to pose the question: How many students are not being taught in their ZPD because they already perform above the instructional level provided in their classroom?

What Is Grade Level?

An essential question that underlies this article, as well as much of K-12 educational policy, relates to how "grade-level" proficiency is determined. Although each state differs in how it sets grade-level proficiency standards, the approach used by the Smarter Balanced Consortium is illustrative. Although individual states adopt cut scores on their own, some (e.g., Wisconsin and California) adopted identical scores, as developed by the Consortium. To do this, first, an online panel engaged in bookmark standard setting to review questions and to determine each question's level of achievement. This was the first round of determining which items corresponded to "grade-level" performance. Next, Consortium members convened in-person panels of 30 people for each grade level and content area. These panelists reviewed the work of the online panels and made recommendations prior to seeing the cut scores recommended by the online panels. This suggests that cut scores started off as fully criterion based, as specific content mastery was set as the initial criterion for proficiency at each grade level, and then these proficiency expectations were modified based on observed pass rates. Cross-level teams reviewed the performance-level descriptors and cut scores before these cut scores were sent to member states for approval. More details of this process can be found in Chapter 5 of the *Smarter Balanced Technical Manual* (Smarter Balanced Assessment Consortium, 2016).

The Measures of Academic Progress (MAP) test does not use "grade-level" standards per se. Rather, Northwest Evaluation Association (NWEA) has conducted multiple alignment studies that connect their MAP assessment to results of both individual state assessments (e.g., NWEA, 2016; State of Texas Assessments of Academic Readiness [STAAR]) and Common Core assessments (e.g., Smarter Balanced; NWEA, 2015). Then, NWEA developed concordance tables that linked student-observed scores on the MAP to students' predicted scores on these other assessments. Because of this, grade-level proficiency scores on the MAP are closely aligned to grade-level proficiency scores on these state tests. Grade-level proficiency, and specifically the degree to which student performance varies around this expectation, formed the primary focus of this study.

Method

In order to understand the state of grade-level heterogeneity in American schools, we first had to operationalize at what point students were likely to be taught content that was below their ZPD, knowing that any choice we made would be imperfect. In the end, two factors guided our decision. Clearly, the actual "zone" in ZPD is ambiguous and does not represent a hard line, but still we had to operationalize such a line for the purpose of the study. First, given the emphasis on "grade-level" instruction (especially given the "grade-level" performance focus in the No Child Left Behind Act and in many state accountability systems), we felt the best approach was to look for the number of students who were by definition "above" grade level. We operationalized this as performance 1 or more years above grade-level proficiency standards on a standardized achievement test. Further influencing this decision was that the Wisconsin Badger Exam Technical Manual, for example, stated "(t)he basis for vertical scaling is that there is substantial overlap from one grade to the next of the skills taught in ELA and Mathematics. For two or more levels apart, however, the overlap is not as great" (Wisconsin Department of Public Instruction, 2015, p. 59). The same is true for data from the STAAR (Texas Education Agency, 2013). Because proximal grade-level comparisons are the most valid on vertically aligned tests, 1 year above grade-level comparisons on these vertically aligned tests are the most valid. We, therefore, used these comparisons to set the point above which students were likely to be taught outside of their respective Zones of Proximal Development.

Data Sources

We examined four different data sets, focusing primarily on measures that were criterion referenced by grade level. Because using only one data set could have led to spurious findings due to idiosyncratic factors (e.g., whether it was Common Core aligned), our goals in selecting data sources were to triangulate data sources and to provide a type of internal self-replication of our findings (Makel & Plucker, 2014). As a result, we selected data from the following assessments:

- The Common Core aligned, but nonadaptive administration, of the Smarter Balanced assessment from Wisconsin (called the Badger Exam–2015 data).
- The Common Core aligned, computer-adaptive version of the Smarter Balanced assessment from California (2015 data).
- The non-Common Core–aligned Texas STAAR Assessment (2016 data).
- The NWEA's computer-adaptive MAP results, which include data from 33 states (2013 data).

Smarter Balanced in Wisconsin. The Smarter Balanced assessment was developed by a 30-state consortium that aimed to assess student learning within traditional academic areas based on the Common Core State Standards (CCSS). Smarter Balanced produces vertically aligned scale scores across Grade Levels 3 through 8 to help inform instructional decisions and provide information about student growth from year to year. "In essence, each Badger Exam vertical scale reflects a single general underlying construct (i.e., overall mathematical ability) from Mathematics tests, grade 3 through grade 8" (Wisconsin Department of Public Instruction, 2015, p. 59). It was also designed to be computer adaptive and thus to enable evaluation of student performance across a wide range of ability levels and to, at least partially, avoid ceiling effects for high-performing students. However, not all states adopted the computer-adaptive version of Smarter Balanced, as we describe below.

In Wisconsin, the Smarter Balanced assessment was known as the Badger Exam. Wisconsin administered the test for a single school year (2014-2015) and relied on the fixed-form version rather than the computer-adaptive version in both English Language Arts (ELA) and mathematics for Grades 3 to 8. Wisconsin adopted cut scores on ELA and mathematics at four levels: *below basic*, *basic*, *proficient*, and *advanced*, and set *proficient* to indicate performance that was on grade level (Wisconsin Department of Public Instruction, n.d.).

Smarter Balanced in California. We analyzed 2014 to 2015 California data from Smarter Balanced for two reasons. First, California implemented the computer-adaptive version of Smarter Balanced (as opposed to the nonadaptive format with fixed grade-level forms used in Wisconsin). This difference has the potential to yield informative comparison data between adaptive and nonadaptive versions of the same test. Second, California's vast size, high rate of test participation, and high level of student diversity made it ideal for our study; relying solely on a state like Wisconsin could yield misleading results due to its smaller size and its lower levels of racial and ethnic diversity. Like Wisconsin, California adopted four levels of cut scores for ELA and mathematics at Grades 3 to 8, with level three representing grade-level expectations (California Department of Education, 2016).

State of Texas Assessments of Academic Readiness. Unlike the Wisconsin and California Smarter Balanced assessments, the STAAR was never based on the CCSS. Instead, the STAAR is aligned to the state-specific Texas Essential Knowledge and Skills outcomes. This was the primary reason we chose Texas as one of the data sources for this study: It used standards other than Common Core, which, therefore, potentially could result in different percentages of students performing above level. As with the Badger Exam, the STAAR was a

fixed-form/nonadaptive assessment. However, each grade-level test did include multiple items from both the grade level above and below, thereby allowing for out-of-level comparisons for proximally adjacent grades (Texas Education Agency, 2013).

One aspect unique to the Texas analysis was that it adopted grade-level proficiency cut scores, but chose to phase them in from 2015 through 2021 when the final approved cut score would enter into force. In this fashion, the proficiency score would be ratcheted up every year through 2021. For example, the Grade 5 reading cut score was set at 1,582. However, the state wanted to give schools several years to adjust curriculum before evaluating them based on these new grade-level criteria. Because of this, the 2015 to 2016 cut score for Grade 5 reading was only 1470. We made the decision to use the 2021 cut scores as these were likely to yield the most conservative results, and because they appeared to represent the skills and abilities that policy makers assigned to "grade-level" proficiency. However, this also means they may underestimate the actual number of students who currently are above grade level.

Measures of Academic Progress. The MAP test is a computer-adaptive assessment created and supported by the NWEA. Generally, the MAP is administered at least twice during the school year—once near the beginning and again at the end. The MAP is used widely, in approximately 10% of all U.S. classrooms. The MAP preadministration/postadministration helps assess potential summer loss (Rambo-Hernandez & McCoach, 2015), as well as academic year growth. Because MAP has been aligned to the Smarter Balanced Assessment, we were able to evaluate MAP scores using the Smarter Balanced criteria for grade-level proficiency (NWEA, 2015). The reading assessment is designed to cover word meaning, as well as literal, interpretive, and evaluative comprehension. The mathematics assessment is designed to cover number systems, operations, equations, geometry, measurement, problem solving, statistics and probability, and their applications (Wang, McCall, Jiao, & Harris, 2013).

Unlike the state-level population data sets used from Wisconsin, California, and Texas, the MAP data represent a broad national sample, but do not systematically reflect any single state's population. Our MAP data set consisted of approximately 45,000 Grade 5 students drawn from 33 U.S. states from the 2013 school year. We examined Grade 5 student data from 2013 because these students completed their MAP tests prior to the full implementation of CCSS, allowing for another point of comparison.

Moreover, because the MAP test is computer adaptive and has a high measurement ceiling, MAP scores have the capability to reveal higher levels of students' performance. As such, above-grade-level performance on the MAP cannot be explained away via the often-used rationale that can apply to pencil-and-paper grade-level achievement tests, in which a fifth-grade student earning a grade-equivalent score of Grade 9 is only doing as well on Grade 5 questions as a Grade 9 student would do if completing these same Grade 5-level questions. Because it is computer adaptive with a high-measurement ceiling, an eighth-grade equivalent score on the MAP test actually is equivalent to eighth-grade performance on a Grade 8 test. Thus, using the MAP data, we were able to estimate not only how many students were above grade level by our 1 year above criterion but also how many were performing *more* than 1 year above grade level.

Data Analysis

We used two different analytical approaches to address our research question—one for the Smarter Balanced and STAAR data and another for the NWEA MAP data. To determine what percentage of students scored a year or more above level on the Wisconsin Smarter Balanced, California Smarter Balanced, or Texas STAAR assessments, percentile tables were obtained for a single year of data. All three of these percentile tables (CA, WI, and TX) are available on this article's Open Science Framework webpage (https://osf.io/82f66/). These tables allowed us to determine what percentile of students at each grade level had scored at the next higher grade level's proficiency cut score. For example, in Wisconsin, the fifth-grade cut score for math was 2528. This score corresponds to the 74th percentile for fourth graders. Thus, we were able to use these data to determine that 26% of fourth-grade students scored at the fifth-grade proficiency level in math. Note that, all three tests are slightly different. Although the same assessment was used in California as was used in Wisconsin, Wisconsin used a nonadaptive version. The STAAR was different in that it was designed around standards specific to Texas, whereas the Smarter Balanced was based off the CCSS. Regardless, population-level state data represented by percentile tables received from each state's education agency were the data sources for these three of our four analyses.

Our approach to analyzing the MAP data was different. Using the concordance tables NWEA calculated for MAP and Smarter Balance assessments (NWEA, 2015), we calculated the proportion of students who, at the beginning of fifth grade, earned MAP scores indicative of end-of-year mastery or higher in Grades 5, 6, 7, and 8 (Table 1). These tables present the scores on the MAP reading or math tests that are comparable to proficiency scores on the Smarter Balanced. For our sample, we calculated the proportion of students earning above-level scores in both reading and mathematics. For Grade 5 students, the MAP reading scores accurately predicted proficiency on the fifth-grade Smarter Balanced assessment 84% of the time, and mathematics scores accurately predicted proficiency 88% of the time. Additionally, the predictions for both reading and mathematics were equally likely to be false positives as false negatives (e.g.,

Table 1. MAP Score That Corresponds to Met Standard in Smarter Balanced Assessment by Grade Level.

Grade	Reading	Mathematics
5	214	229
6	218	230
7	222	235
8	225	242

Note. MAP = Measures of Academic Progress.

8% in reading were false positives and 8% were false negatives; NWEA, 2015).

Results

Table 2 presents the percentages of Wisconsin, California, and Texas students in a given grade level who scored at or above the proficiency threshold established for *1 year above their current grade* in ELA or mathematics. Stated another way, this table presents the percentage of students who demonstrated performance 1 or more years advanced in each content area.

At the end of the 2014 to 2015 school year, between 26% and 49% of Wisconsin students scored at or above proficient for the next grade level. Similarly, between 18% and 38% of California students scored at or above the next grade level in the spring of 2015. The percentages tend to be greater for students in the higher grades, perhaps because of increasing standard deviations of scores. Between 16% and 34% of Texas students scored at or above the next grade level in the spring of 2016, with lower numbers appearing in Grade 3. Overall, the highest percentages came from Wisconsin, with a few exceptions, and the lowest percentages came from California.

NWEA MAP Reading

We used the Grade 5 Fall 2013 MAP data to estimate how many students were at least 1 year above grade level, by determining how many students at the *beginning* of Grade 5 were already achieving at *end-of-year* Grade 5 proficiency levels on MAP reading. We also were able to determine how many of these above-grade-level students achieved MAP test scores equivalent to year-end scores for Grades 6, 7, and 8 students. Figure 1 presents these findings graphically.

As shown in Figure 1, at the beginning of their Grade 5 year, approximately 34% of students had scores commensurate with end-of-year Grade 5 proficiency reading levels, roughly similar to the results from the other three assessments. Furthermore, approximately 10% of all Grade 5 students demonstrated Grade 8-level end-of-year proficiency. These students were 4 school years ahead of grade level in reading, essentially reading at high school level. We did not make this comparison beyond 1 year with the three state assessments because such distant comparisons are less valid on both Smarter Balanced and STAAR.

NWEA MAP Mathematics

Nearly 14% of Grade 5 students at the beginning of the school year were already earning MAP scores consistent with end-of-Grade 5 proficiency (Figure 1). These estimates are smaller than those from the Wisconsin, California, and Texas data, but are still quite large. About 2.4% of all Grade 5 students were achieving at levels equal to, or above, the end of Grade 8 (or high school level, 4 school years ahead of grade level) in mathematics.

As a post hoc analysis, we also conducted a one-way analysis of variance on fall Grade 5 scores with school as the grouping variable. In this analysis, between-school variability accounted for 16% of the variance in scores, whereas within-school variability accounted for 84% of the variance in scores. While 16% is not negligible, the majority of variability in scores lies within schools. This wide range of achievement within schools indicates that the observed range in achievement is not just a function of between school differences. Students within schools—all schools—are achieving at vastly different levels.

Discussion

Despite wide variability in the grade levels and data sets we assessed, our results showed consistently that a large number of students are capable of working more than a year beyond the grade-level expectations on which instruction typically is based. In math, anywhere from 16% to 37% of students scored a year or more above their current grade level. For ELA, 20% to 49% scored a year or more ahead. Considering that the state data used in this study were from some of the largest student populations in the country, these percentages represent millions of students.

Using MAP data, we estimate the 10% of Grade 5 students currently score at the Grade 9 level in ELA, with 2% scoring at similar levels in mathematics. In other words, roughly 1 out of 10 Grade 5 students demonstrates reading performance at the high school level and nearly 1 student in 50 at this age demonstrates mathematics performance at the high school level. This is an extreme level of heterogeneity for a single grade-level classroom. Any teacher, even those who are the most highly skilled at differentiation, will find it difficult to challenge all students in their respective ZPDs in such diverse, age-based classrooms.

Converting the percentages reported in Table 2 and Figure 1 to numbers of children provides a sobering picture of the number of students who are mismatched with the current grade-based educational paradigm. For example, in 2014-2015, there were 475,192 students enrolled in Grade 4 in California Public schools (http://www.cde.ca.gov/ds/sd/cb/cefenrollmentcomp.asp). With 29% demonstrating

Table 2. Percentages of Students Scoring 1 Year or More Above Grade Level.

	ELA % scoring 1+ years above			Mathematics % scoring 1+ years above		
Grade	Wisconsin	California	Texas[a]	Wisconsin	California	Texas[a]
3	34	23	20	26	19	16
4	39	29	25	26	18	29
5	44	34	30	31	22	34
6	49	34	24	36	27	32
7	47	38	30	37	28	33

[a]Texas percentages are based on the approved cut scores set for test year 2021.

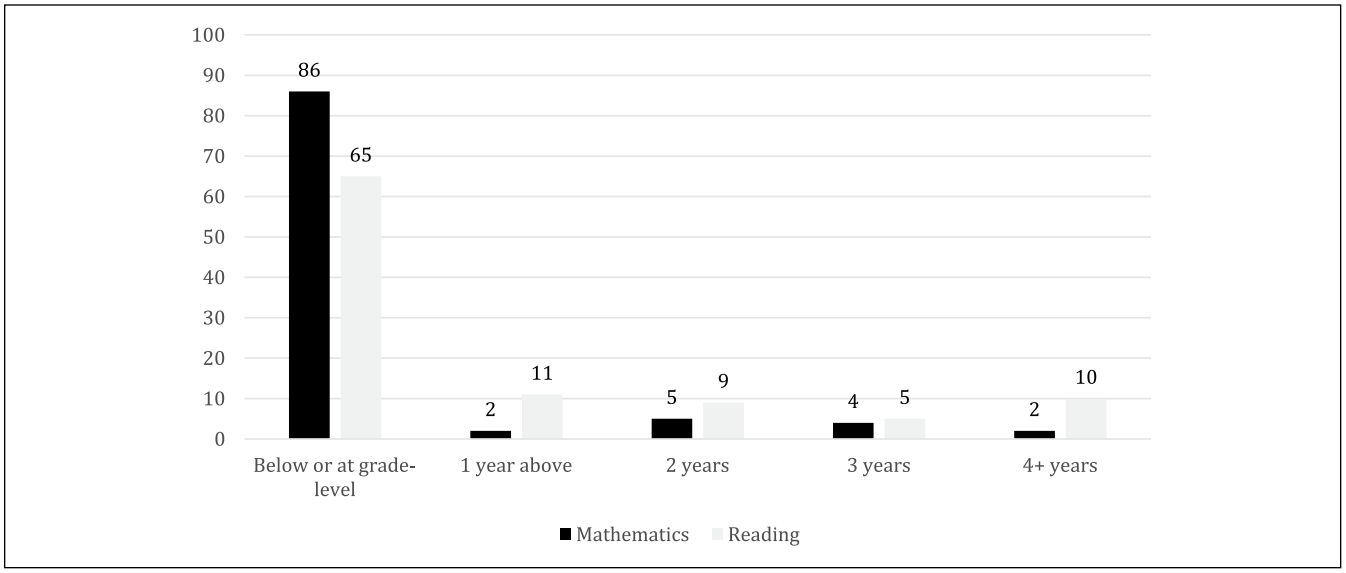

Figure 1. Percentage of fifth-grade students at and above grade level on NWEA MAP.
Note. NWEA = Northwest Evaluation Association; MAP = Measures of Academic Progress.

above-grade-level ELA performance, this means that over 137,000 Grade 4 students in California demonstrate above-grade-level performance in ELA; enough to more than fill the largest of football stadiums. With an estimated 5,858 elementary schools in the state, this means that each California elementary school, on average, has 23 students who are working above grade level in ELA; enough to fill an entire classroom. In Texas, where 397,056 students were enrolled in Grade 4 in 2015-2016 (https://rptsvr1.tea.texas.gov/adhocrpt/adste.html), 29% were above grade level in math. This means that over 115,000 students demonstrated above-grade-level performance. This is an average of nearly 29 students in each of the 4,001 elementary schools in Texas who would benefit from accelerated, above-grade-level content in mathematics. In Wisconsin, where 60,795 Grade 4 students were enrolled in public schools in 2014-2015 (http://wisedash.dpi.wi.gov/Dashboard/portalHome.jsp), over 23,700 students demonstrated above-grade-level ELA performance. Each of Wisconsin's 1,233 elementary schools, on average, would have at least 19 fourth graders who perform a year or more above grade level in ELA. In sum, there are more than 300,000 fourth-grade students demonstrating above-grade-level performance in reading or mathematics (some in both) in only these three states. This demonstrates that students performing above grade level are not rare, likely exist in every classroom in every school, and in numbers large enough to permit an accelerated classroom of these learners in every school, on average. All of this leaves us to wonder what (if any) policies or procedures may exist to make sure these students are appropriately challenged in school.

It is not surprising that the mathematics percentages, although quite high, are not as large as the reading/language arts numbers. Due to inconsistent or absent policies regarding acceleration (Assouline, Colangelo, VanTassel-Baska, & Lupkowski-Shoplik, 2015), high-performing students in Grades 5 or 6 are rarely given access to algebra, geometry, statistics, or calculus courses, and if they are, they often get placement, but not credit, when they subsequently reach high school (Plucker, Giancola, Healey, Arndt, & Wang, 2015).

The lack of acceleration in math thus provides a structural barrier to moving too far "above grade level." Achievement in reading does not face similar barriers as is it less teacher dependent than mathematics (Alexander, Entwisle, & Olson, 2001). In theory, an advanced reader could fairly easily access more challenging books through the library or online in order to continue to develop her skills, but a student who has mastered geometry might be unaware of the possibility of accessing trigonometry content without some additional guidance from a parent, teacher, or older sibling.

Additionally, it is worth noting the percentages of students above grade level were conditional on the assessment and the grade-level standards. For example, the MAP mathematics assessment results indicated 14% of students were performing above grade level in fifth grade, but the Texas STAAR results indicated 30% of students. This variation is likely due to the nature of the assessment (computer adaptive vs. traditional), but also to the discrepancy in the standards and related standard-setting process (Common Core curriculum vs. Texas Essential Knowledge and Skills). Furthermore, the content covered on each assessment is not identical, and even the content that is covered on both assessments may not be represented in the same proportion on both assessments. For example, if one assessment emphasized computation over problem solving, and the other emphasized problem solving over computations, the same student might score above grade level on the first but not the second (Rambo-Hernandez & Warne, 2015).

Also of note, MAP results generally provided the most conservative estimates of students above grade level. Most achievement tests are designed to provide the maximum information around critical cut scores, such as grade-level proficiency. As a result, the conditional reliability is the smallest around those cut scores (e.g., grade level) and is greater at more extreme scores (e.g., above grade level; Lohman & Korb, 2006; Rambo-Hernandez & Warne, 2015). Put simply, this makes scores less accurate the farther the score is from grade level. Both the Wisconsin Badger and Texas STAAR both contained items that were above grade level, but the number of items above grade level was fixed for each assessment. For example, regardless of student achievement level, all students were presented with the same number of above and below grade-level items.

This was less of an issue for the MAP and the California Smarter Balanced assessments as their computer-adaptive approach presents items to students based on their performance on previous items. Students who are working above grade level will respond to many more above-grade-level items in a computer-adaptive setting than on a traditional paper–pencil format assessment. Thus, the above-grade-level scores based on computer-adaptive assessments will have a smaller conditional reliability than above-grade-level scores based on traditional paper–pencil assessments. In other words, the above-grade scores obtained from the computer-adaptive assessments contain less error than above-grade-level scores from paper–pencil assessments.

Finally, in mathematics, the proportion of students who scored above grade level on the California Smarter Balanced taken at the end of the year was larger than the proportion of students scoring above grade level at the beginning of the year on the mathematics MAP, despite the fact that they were both adaptive. This discrepancy could simply be a manifestation of summer slide (Alexander et al., 2001). However, in reading, the proportion of students scoring above grade level at the end of the school year on the California Smarter Balanced assessment and the beginning of the school year on the MAP was quite similar. As previously noted, learning mathematics is typically more teacher-dependent than reading (Alexander et al., 2001), so students likely lost some ground in mathematics but more easily maintained their reading scores over the summer.

Conclusions

When first reviewing our results, an initial reaction was to think that the placement of grade-level cut scores might be too low, and that this might explain why there are so many students above grade level. Certainly a higher cut score in any of our example states would have the effect of decreasing the number of students who produce scores above it. But that would have no effect on the variability that exists within a given grade level. What we present in this study underscores just how much variability there is in a given grade-level classroom, and our corresponding judgment of the undesirability of continuing to focus instruction on grade-level standards. Regardless of whether the proficiency standards could be raised or lowered, the wide variance in grade-level student mastery will persist and will continue to challenge the educational system. Differentiation as a pedagogical skill is difficult in the best of cases, and only becomes more so as the range of student needs increases (Hertberg-Davis, 2009).

Based on the relatively consistent results across our four data sets, we argue there is little support for the current age-based classroom structure as the optimal organizational structure for fostering student development. The concept of ZPD posits that students need to be engaged with content that is above their current level of independent mastery, and to do so requires appropriate scaffolding from skilled educators. Classrooms where large percentages of students are above grade level, but nearly all of the teacher's focus is at or below grade level (e.g., Engel et al., 2012; Farkas & Duffett, 2008), are not going to facilitate growth or further development for students who are already working above grade level. These findings align with past concerns regarding age-based grades and further support far broader utilization of whole grade and other forms of academic acceleration (e.g., Assouline et al., 2015).

As can be seen in Table 2, the percentages of students scoring above grade level are higher in the upper grades.

Although as mentioned earlier such a finding could be an artifact of the increasing standard deviations of scores at upper grade ranges, these larger percentages could also indicate that more students may be in need of grade acceleration in the upper elementary and middle grades than in early elementary. Part of the implication of a strict, age-based grouping system is as students have advanced needs, they might need to go to completely different grade levels to have those needs met. The data from Table 2 suggest that this might be especially true in the upper grades of our analyses.

An instructional system heavily focused on grade-level content essentially ignores the learning needs of a large percentage of its students. Having established from multiple data sources that large percentage of students are achieving above grade level, educators, researchers, and policy makers need to work together to reconceptualize traditional but outdated grade-based standards, and to consider honestly where and through what mechanism these students are going to be challenged.

Authors' Note

All authors contributed equally to this article. An earlier version of the data presented in this article were published online by the Johns Hopkins Institute for Education Policy (http://education.jhu.edu/edpolicy/commentary/PerformAboveGradeLevel).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by research funds provided to the fifth author by Johns Hopkins University.

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, *23*, 171-191. doi:10.3102/01623737023002171

Assouline, S. G., Colangelo, N., VanTassel-Baska, J., & Lupkowski-Shoplik, A. (2015). *A nation empowered: Evidence trumps the excuses holding back America's brightest students*. Iowa City, IA: Connie Belin and Jacqueline N. Blank International Center for Gifted Education and Talent Development.

California Department of Education. (2016). *Smarter Balanced assessment system*. Retrieved from http://www.smarterbalanced.org/assessments/

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., . . . Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child × instruction interactions on first graders' literacy development. *Child Development*, *80*, 77-100. doi:10.1111/j.1467-8624.2008.01247.x

Engel, M., Claessens, A., & Finch, M. A. (2012). Teaching students what they already know? The (mis)alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis*, *35*, 157-178. doi:10.3102/0162373712461850

Evertson, C. M., Sanford, J. P., & Emmer, E. T. (1981). Effects of class heterogeneity in junior high school. *American Education Research Journal*, *18*, 219-232. doi:10.2307/1162383

Farkas, S., & Duffett, A. (2008). *High-achieving students in the era of NCLB: Results from a national teacher survey*. Retrieved from http://www.nagc.org/sites/default/files/key%20reports/High_Achieving_Students_in_the_Era_of_NCLB_Fordham.pdf

Firmender, J. M., Reis, S. M., & Sweeney, S. M. (2013). Reading comprehension and fluency levels ranges across diverse classrooms: The need for differentiated reading instruction and content. *Gifted Child Quarterly*, *57*, 3-14. doi:10.1177/0016986212460084

Herrnstein, R. J. (1971). *I.Q. in the meritocracy*. Boston, MA: Little, Brown.

Hertberg-Davis, H. (2009). Myth 7: Differentiation in the regular classroom is equivalent to gifted programs and is sufficient: Classroom teachers have the time, the skill, and the will to differentiate adequately. *Gifted Child Quarterly*, *53*, 251-253. doi:10.1177/0016986209346927

Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted*, *29*, 451-484. doi:10.4219/jeg-2006-245

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304-316. doi:10.3102/0013189X14545513

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, *92*, 263-283. doi:10.1162/rest.2010.12318

Northwest Evaluation Association. (2015). *Linking the Smarter Balanced Assessments to NWEA MAP Assessments*. Retrieved from https://www.nwea.org/resources/linking-the-smarter-balanced-assessments-to-nwea-map-assessments/

Northwest Evaluation Association. (2016). *Linking the Texas STAAR Assessments to NWEA MAP Tests*. Retrieved from https://www.nwea.org/content/uploads/2016/02/Texas_Linking_Study_FEB2016.pdf

Plucker, J. A., Giancola, J., Healey, G., Arndt, D., & Wang, C. (2015). *Equal talents, unequal opportunities: A report card on state support for academically talented low-income students*. Retrieved from http://www.jkcf.org/assets/1/7/JKCF_ETUO_Executive_Final.pdf

Polikoff, M. (2016, July 12). *A letter to the U.S. Department of Education* (Final signatory list). Retrieved from https://morganpolikoff.com/2016/07/12/a-letter-to-the-u-s-department-of-education/

Rambo-Hernandez, K. E., & McCoach, D. B. (2015). High-achieving and average students' reading growth: Contrasting school and summer trajectories. *Journal of Educational Research*, *108*, 112-129. doi:10.1080/00220671.2013.850398

Rambo-Hernandez, K. E., & Warne, R. T. (2015). Measuring the outliers: An introduction to out-of-level testing with

high-achieving students. *Teaching Exceptional Children*, *47*, 199-207. doi:10.1177/0040059915569359

Smarter Balanced Assessment Consortium. (2016). *Smarter Balanced Assessment Consortium: 2014-2015 Technical report*. Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf

Texas Education Agency. (2013). *State of Texas Assessments of Achievement (STAAR™): Vertical scale technical report*. Retrieved from http://tea.texas.gov/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=25769806053&libID=25769806056

Vygotsky, L. (1997). *Interaction between learning and development*. In M. Gauvain & M. Cole (Eds.), *Readings on the development of children* (pp. 29-36). London, England: Worth.

Wang, S., McCall, M., Jiao, H., & Harris, G. (2013). Construct validity and measurement invariance of computerized adaptive testing: Application to Measures of Academic Progress (MAP) using confirmatory factor analysis. *Journal of Educational and Developmental Psychology*, *3*, 88-100. doi:10.5539/jedp.v3n1p88

Wisconsin Department of Public Instruction. (n.d.). *About the data—Badger*. Retrieved from http://dpi.wi.gov/wisedash/about-data/badger#Scoring

Wisconsin Department of Public Instruction. (2015). *The Badger Exam 3-8: A Wisconsin Smarter Balanced Assessment* (Technical report). Retrieved from http://dpi.wi.gov/sites/default/files/imce/assessment/pdf/Badger%20WI%20Tech%20Manual%20Final%20Version.pdf

Author Biographies

Scott J. Peters is an associate professor of Educational Foundations at the University of Wisconsin–Whitewater where he teaches courses related to educational measurement, research methods, and gifted and talented education. His primary research area involves gifted and talented student identification and talent development with a focus on students from underrepresented populations.

Karen Rambo-Hernandez is an assistant professor of Educational Psychology in the Department of Learning Sciences and Human Development at West Virginia University. Her research interests include multilevel modeling, growth modeling, academic acceleration, gifted education, and STEM education.

Matthew C. Makel is the director of research at the Duke University Talent Identification Program. His research focuses on research methods and the nature and development of the abilities, perceptions, and environments of academically talented youth.

Michael S. Matthews is professor and director of the Academically and Intellectually Gifted graduate programs at the University of North Carolina at Charlotte. His professional interests in advanced academics and gifted education include research methods, education policy, science learning, motivation and underachievement, and parenting. His scholarship also focuses on gifted and academically advanced learners from diverse backgrounds, particularly those who are English learners.

Jonathan A. Plucker is the Julian C. Stanley professor of Talent Development at the Center for Talented Youth and School of Education at Johns Hopkins University. With a background in educational psychology and education policy, he studies excellence gaps, creativity and intelligence, and research methods.