education policy analysis archives

A peer-reviewed, independent, open access, multilingual journal

epaa aape

Arizona State University

Volume 31 Number 10 February 7, 2023 ISSN 1068-2341

Same Tests, Same Results: Multi-Year Correlations of ESSA-Mandated Standardized Tests in Texas and Nebraska

Norman P. Gibbs
Mesa Unified School District
Margarita Pivovarova
&
David C. Berliner
Arizona State University
United States

Citation: Gibbs, N. P., Pivovarova, M., & Berliner, D. C. (2023). Same tests, same results: Multi-year correlations of ESSA-mandated standardized tests in Texas and Nebraska. *Education Policy Analysis Archives, 31*(10). https://doi.org/10.14507/epaa.31.7696

Abstract: Statewide assessments in reading and math are required every year under the Every Student Succeeds Act (ESSA) of 2015, at an annual expense of billions of taxpayer dollars. Analyzing 10 years of school-level results from public schools in two states—Nebraska and Texas—we found that year-to-year correlations of schools' test scores were exceptionally high, with reading and math correlations regularly above $r = 0.9$, indicating that little new information is derived from annual testing. Furthermore, while schools experiencing the largest demographic changes had significantly lower year-to-year correlations—highlighting the sensitivity of these scores to changes in underlying student demographics—even these lower correlations remained strong. We argue that the frequency of these tests may therefore be reduced, yielding substantial savings of time and money and no loss of useful information.

Keywords: accountability; educational policy; standardized testing

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 6/30/2022
Revisions received: 10/1/2022
Accepted: 10/2/2022

Mismas pruebas, mismos resultados: Correlaciones multianuales de las pruebas estandarizadas exigidas por ESSA en Texas y Nebraska
Resumen: Se requieren evaluaciones estatales de lectura y matemáticas todos los años según la Ley Every Student Succeeds (ESSA) de 2015, con un gasto anual de miles de millones de dólares de los contribuyentes. Al analizar 10 años de resultados a nivel escolar de escuelas públicas en dos estados, Nebraska y Texas, encontramos que las correlaciones de año a año de los puntajes de las pruebas de las escuelas fueron excepcionalmente altas, con correlaciones de lectura y matemáticas regularmente por encima de $r = 0.9$, lo que indica que poca información nueva se deriva de las pruebas anuales. Además, aunque las escuelas que experimentaron los cambios demográficos más grandes tuvieron correlaciones significativamente más bajas de año a año, lo que resalta la sensibilidad de estos puntajes a los cambios en la demografía subyacente de los estudiantes, incluso estas correlaciones más bajas se mantuvieron sólidas. Argumentamos que, por lo tanto, la frecuencia de estas pruebas puede reducirse, con ahorros sustanciales de tiempo y dinero y sin pérdida de información útil.
Palabras clave: rendición de cuentas; política educativa; pruebas estandarizadas

Mesmos testes, mesmos resultados: Correlações de vários anos de testes padronizados exigidos pela ESSA no Texas e Nebraska
Resumo: Avaliações estaduais em leitura e matemática são exigidas todos os anos sob o Every Student Succeeds Act (ESSA) de 2015, com um gasto anual de bilhões de dólares dos contribuintes. Analisando 10 anos de resultados de nível escolar de escolas públicas em dois estados—Nebraska e Texas—descobrimos que as correlações ano a ano dos resultados dos testes das escolas eram excepcionalmente altas, com correlações de leitura e matemática regularmente acima de $r = 0,9$, indicando que pouca informação nova é derivada de testes anuais. Além disso, enquanto as escolas que passavam pelas maiores mudanças demográficas apresentavam correlações ano a ano significativamente mais baixas, destacando a sensibilidade dessas pontuações às mudanças na demografia subjacente dos alunos, mesmo essas correlações mais baixas permaneceram fortes. Argumentamos que a frequência desses testes pode, portanto, ser reduzida, com economias substanciais de tempo e dinheiro e sem perda de informações úteis.
Palavras-chave: prestação de contas; política educacional; testes padronizados

Same Tests, Same Results: Multi-Year Correlations of ESSA-Mandated Standardized Tests in Texas and Nebraska

Researchers, teachers, and school administrators, as well as parents, have frequently expressed concerns about the value and costs of mandated standardized achievement testing. Nevertheless, by federal law and with state support, such tests are an annual ritual in nearly all of the nation's 13,000 school districts. These tests are costly to purchase, time consuming to administer and analyze, and frequently produce anxiety for educators, parents, and students (Heissel et al., 2017, 2021; Segool et al., 2013; von der Embse et al., 2017). Further, in many states, the information obtained from such tests is of limited use to teachers because the data on individual students are typically not available until well after students have moved on to the next grade. These annual tests do, however, supply data of interest, often presented as part of a school or district "letter grade" (GGOSA, 2021; MDE, 2021; TEA, 2021a). These data are of interest to parents, educational

administrators, journalists, and politicians at district, state, and federal levels (Jacobsen et al., 2014; Murray & Howe, 2017).

Furthermore, the data from annual assessments of student achievement at the school level are frequently used to target federal, state and district funding, personnel assignment, and programs. While these tests do provide important data for long-term program evaluation, the informational value of the federal government's every-child, every-year approach merits further examination (Kahl, 2021). We present evidence here indicating that the frequency of these tests may be reduced, with substantial savings of time and money and no loss of useful information.

Context

The accountability protocols formalized in the Every Student Succeeds Act of 2015 (ESSA) require annual testing in English and mathematics for all students in third through eighth grade and once in high school (ESSA, 1111(b)(2)). While ESSA was designed to devolve more control to states and districts (Black, 2017; Edgerton, 2019; Egalite et al., 2017), the legislation nevertheless maintained the test mandate established under the No Child Left Behind Act of 2001 (NCLB), which required that states submit accountability data in order to help the federal government ascertain that Title I funds create more equitable opportunities for less advantaged students. The tests also serve instrumental purposes (Sutherland, 2022), allowing the federal government to pressure schools toward school improvement on an annual basis in lieu of any federal capacity to direct the work of individual LEAs. Thus, for the federal government, the tests serve both as accountability mechanisms for Title I spending and as instruments for triggering a culture of continuous improvement.

Yet it is not only—nor even primarily—federal institutions that derive value from standardized tests. A range of audiences, including parents, teachers, school administrators, district administrators, and state administrators, may use the findings of these assessments for decision making. For parents, the findings of standardized tests, summarized annually at the school level in ESSA-mandated school report cards (ESSA, 1111(h)(2)), help them make decisions about which schools to attend—and, as members of the general public, which schools to praise and which schools to pressure for reform (Sutherland, 2022). Likewise, the state assessments provide parents with information on the educational development of their own children—albeit in addition to regular classroom grades and other school or district assessments that may be administered.

For teachers, the use of standardized test data is more complicated. In most states, the results of these assessments rarely arrive before the end of the school year, affording little formative value for the students they assess. Instead, educators regularly rely on LEA-selected formative assessments rather than state-level assessments to make decisions about at-risk students. For example, Sun et al. (2016) found in a review of the relevant literature that teachers give preference to formative (student portfolios, group projects, journaling, enrichment games, quick group activities, or other school or district formative assessments) over summative accountability tests. Indeed, teachers have been shown to make accurate judgements of their students' abilities even before they know the results of accountability tests (see Fryer & Levitt, 2004).

Even if standardized test data was received on a more timely basis, for teachers to use that data effectively and have a comprehensive understanding of their students, they would need to tap into other student data, such as attendance, socio-economic conditions, and socio-emotional learning, among other indicators (Datnow & Park, 2018; Mandinach et al., 2019). Yet, as pointed out in Mandinach and Schildkamp (2021), teachers often face challenges when attempting to integrate these diverse data sources for instructional improvement—challenges including teachers' inability to

access this data themselves (Schildkamp & Kuiper, 2010) or, for those who are able to access these data, a limited understanding of the analytic approaches best suited to integrating data for instructional improvement (Hoover & Abrams, 2013).

In addition, some have argued that, among educators, data use for accountability purposes has narrowed the focus of how teachers use the data from standardized assessments. Often these data are used "superficially, supporting the deficit mindset" (Mandinach & Schildkamp, 2021, p. 4; see also Ingram et al., 2004 and Sun et al., 2016) and with a focus on test achievement rather than on deeper learning (Au, 2007; Berliner, 2011; Datnow et al., 2018). Given this focus, educators may employ a superficial use of data for triage purposes (Booher-Jennings, 2005; Garner et al., 2017; Lai & Schildkamp, 2016) or resort to test preparation without due attention to underlying teaching practice (Garner et al., 2017).

This reactive pedagogy is often tied to the strong correlation between socio-economic status (SES) and students' academic performance. While student SES lies beyond the control of educators, prior research has repeatedly demonstrated that this correlation nevertheless accounts for much of the variation that does exist in accountability data, resulting in increased psychological and pedagogical pressure on teachers (Coleman, 1996; Klein et al., 2000; Perry & McConney, 2010; Reardon, 2011; Sirin, 2005; White, 1982). Schools with the highest proportion of low-income students face greater difficulties in meeting state proficiency expectations and thus attract increased scrutiny from state accountability systems (Cunningham & Sanzo, 2002). It is little surprise, then, that school participation in these high-stakes state accountability activities has been associated with counterproductive teaching practices and poorer school climate, with negative effects on teachers' mental health and increased stress and anxiety (Nathaniel et al., 2016; von der Embse et al., 2017). Furthermore, this scrutiny strengthens the incentives for educators in low-scoring schools to adopt short-term strategies in an attempt to inflate test scores (Mintrop & Sunderman, 2009)—strategies such as targeting students at cutoff points (so-called "bubble kids"), reducing instructional differentiation, and narrowing the curriculum through teaching to the test and standard-specific remediation, to name a few (Amrein-Beardsley, 2022). Given these findings of disproportionately negative effects on schools with higher shares of low income and minority students, researchers have repeatedly raised concerns about the unintended consequences of test-based accountability policies (Baker & Johnston, 2010; Cahill, 2019; Cunningham & Sanzo, 2002; Herman & Golan, 1993; Mintrop & Sunderman, 2009; Polesel et al., 2012; Polesel et al., 2014).

It would appear, then, that these instrumental federal mandates do indeed have systemic effects—for better or worse. As discussed above, knowing that accountability tests are coming, educators introduce a variety of instructional changes tailored to the test and implement formative assessments to prepare their students for federal testing as well as minimizing the likelihood of receiving disappointing end-of-year results. Yet it is unclear to what extent an *annual* federal assessment would be necessary to trigger this response and whether it is likely that schools would abandon their existing school improvement efforts were the test to happen every second or third year. This question has fiscal implications, as the multi-billion-dollar investments states make to comply with ESSA's annual federal mandate fall on top of schools' investments in formative assessments. In the two states featured in the present study, Nebraska paid $29,000,000 to the Northwest Educational Association (NWEA) in 2017 for a five-year contract for its 410 schools (a mean of $14,000 per school, per year; Cavanaugh, 2017) while Texas paid Cambrium and Pearson $388,000,000 in 2021 for a four-year contract for its 7,510 schools (a mean of $13,000 per school, per year; Swaby, 2021). Thus, given the costs associated with federally-mandated accountability assessments—costs that are both pedagogical and fiscal —the question remains as to what value

those annual assessments bring that would not be served from testing less frequently. We explore this question below.

Data

Data from this study is composed of school-level results on end-of-year mandated assessments in reading and math for 7,920 elementary and middle schools in the states of Texas (n = 7510) and Nebraska (n = 410), covering the period 2010–2019. For both states, the data are publicly available records from websites at both states' departments of education (NDOE, 2021; TEA, 2021b). The sample for the final analysis included a balanced panel of public elementary and middle schools in each state for which we had an uninterrupted time series of average school-level scores for reading (or ELA) and mathematics.

The variables of interest in this study were the schools' average scores for reading and math on their states' standardized tests as required every year under ESSA for students in grades 3–8. While both Texas and Nebraska report separate scores for mathematics, and Texas reports a separate score for reading, Nebraska's testing shifted in the 2017–2018 academic year from a single reading score to a combined English language arts (ELA) score, following the state's adoption of a new computer-adaptive test administered by NWEA (Jespersen, 2018). We follow Nebraska's practice of treating the new combined ELA score as comparable to the prior test's reading score and treat it as such for comparison with Texas.

Methods

In our analysis, we sought to understand the predictive power of a given year's test results in forecasting test scores in the following years. To answer this research question, we conducted bivariate correlation analysis between schools' test scores in a given year and the test scores for up to five (in our main analysis) and nine (for some analyses) years following.

In addition to our main analysis, we wanted to understand to what extent changes in annual accountability data are a measure of socioeconomic shifts rather than indicators of substantive instructional changes. To do that, we created two indicators to measure change in socio-economic composition: standard deviation in the share of free- and reduced-lunch (FRL) eligible students over time, and the difference in the largest and smallest share in FRL-eligible students over the same period. Using these indicators, we applied a cut-off rule, selecting the top 10% of schools in that distribution of standard deviations and min-max difference to identify schools that experienced substantial changes in the socio-economic composition of their students. We hypothesized that, if the demographic composition of a school did not change dramatically over the years, we should expect more stability in such schools' average school test scores. We tested this hypothesis by statistically comparing year-to-year average school test score correlations between those schools that experienced greater changes in their students' socio-demographic composition and schools with a more stable composition. To test for the difference between these correlations, we applied a Fisher z-transformation and conducted a series of z-tests, the results of which are reported below.

Results

In our main analysis, we found a marked and consistent year-to-year correlation within both states, with the correlation of each year's tests with those administered in the years immediately following falling between 0.87 and 0.98 for the period 2014–2019 (Table 1, below). Translated into predictive power as r-squared values (Cohen et al., 2003), these correlations indicate that on average,

from about 76% to 96% of the variability in scores between schools is explained by their performance in the previous year(s). Even with major changes in test administration, as in the case of Nebraska's adoption of NWEA's computer-adaptive test in the 2017–2018 academic year, scores remain remarkably consistent. All of Nebraska's scores on the new test remained highly correlated ($r \geq 0.90$) with the prior year's test. Furthermore, as shown in Table 1, strong correlations persist in both states beyond the one-year correlation and into the second, third, fourth, and fifth years.

Table 1

Correlations of Test Scores with Tests in Following Years, Nebraska and Texas, 2014–2018

	2014		2015		2016		2017		2018	
Years Later	**NE**	**TX**	**NE**	**TX**	**NE**	**TX**	**NE**	**TX**	**NE**	**TX**
Mathematics, Elementary Schools										
1 Year Later	0.93	0.88	0.95	0.91	0.94	0.91	0.90	0.92	0.95	0.91
2 Years Later	0.91	0.84	0.91	0.85	0.88	0.86	0.87	0.86		
3 Years Later	0.87	0.81	0.86	0.81	0.84	0.81				
4 Years Later	0.85	0.77	0.82	0.78						
5 Years Later	0.81	0.70								
Mathematics, Middle Schools										
1 Year Later	0.98	0.87	0.97	0.90	0.98	0.90	0.97	0.90	0.98	0.91
2 Years Later	0.96	0.84	0.96	0.86	0.96	0.84	0.97	0.86		
3 Years Later	0.94	0.81	0.95	0.80	0.96	0.82				
4 Years Later	0.92	0.76	0.95	0.79						
5 Years Later	0.92	0.76								
Reading, Elementary Schools										
1 Year Later	0.95	0.93	0.96	0.94	0.96	0.93	0.91	0.94	0.96	0.93
2 Years Later	0.93	0.90	0.93	0.90	0.91	0.90	0.89	0.89		
3 Years Later	0.92	0.87	0.89	0.87	0.90	0.86				
4 Years Later	0.91	0.85	0.87	0.84						
5 Years Later	0.89	0.81								
Reading, Middle Schools										
1 Year Later	0.98	0.92	0.93	0.91	0.94	0.91	0.97	0.93	0.97	0.93
2 Years Later	0.92	0.89	0.95	0.89	0.91	0.89	0.97	0.90		
3 Years Later	0.94	0.89	0.95	0.87	0.92	0.86				
4 Years Later	0.93	0.86	0.96	0.85						
5 Years Later	0.93	0.84								

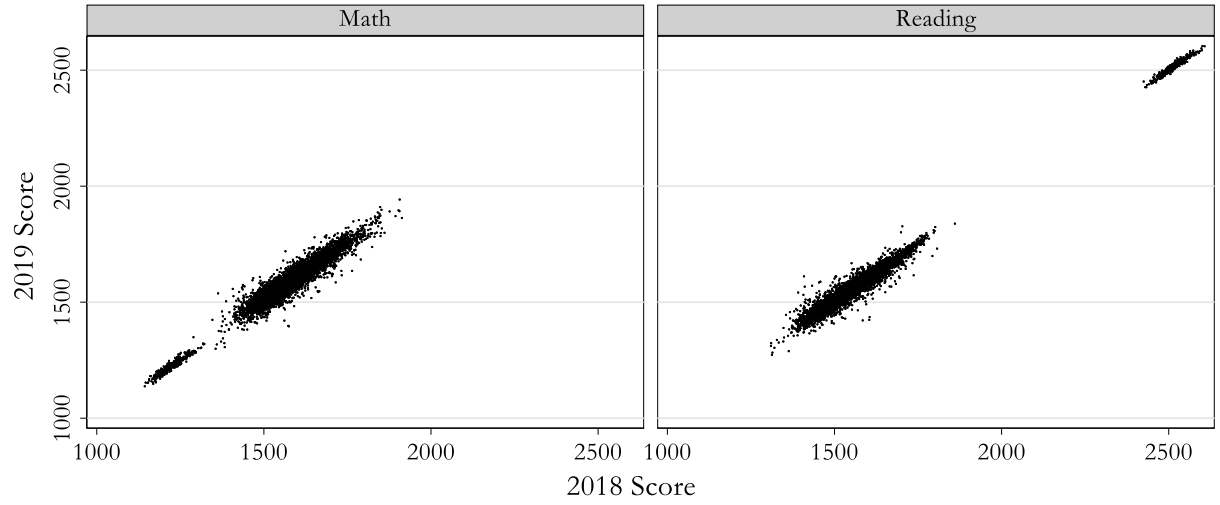
Note: All values are significant at $p < .001$. "NE": Nebraska; "TX": Texas.

These strong correlations are readily seen in graphical representations. Since all pairs of consecutive-year correlations are similar to each other in absolute values, we plotted all reading and mathematics average school scores for one representative pair of years in a scatter diagram (Figure 1, below). The graphs in Figure 1 demonstrate a strong linear relationship between average school scores in reading and mathematics between 2018 and 2019. Almost all data points are situated along the straight 45-degree line. These graphs also show the identical relationship between average test

scores in the two states, as, due to differences between scales of measurement, one cluster represents Texas while a second, smaller, cluster represents Nebraska.

Figure 1

Correlation of Elementary/Middle School Average Test Scores, 2018–2019, Nebraska and Texas

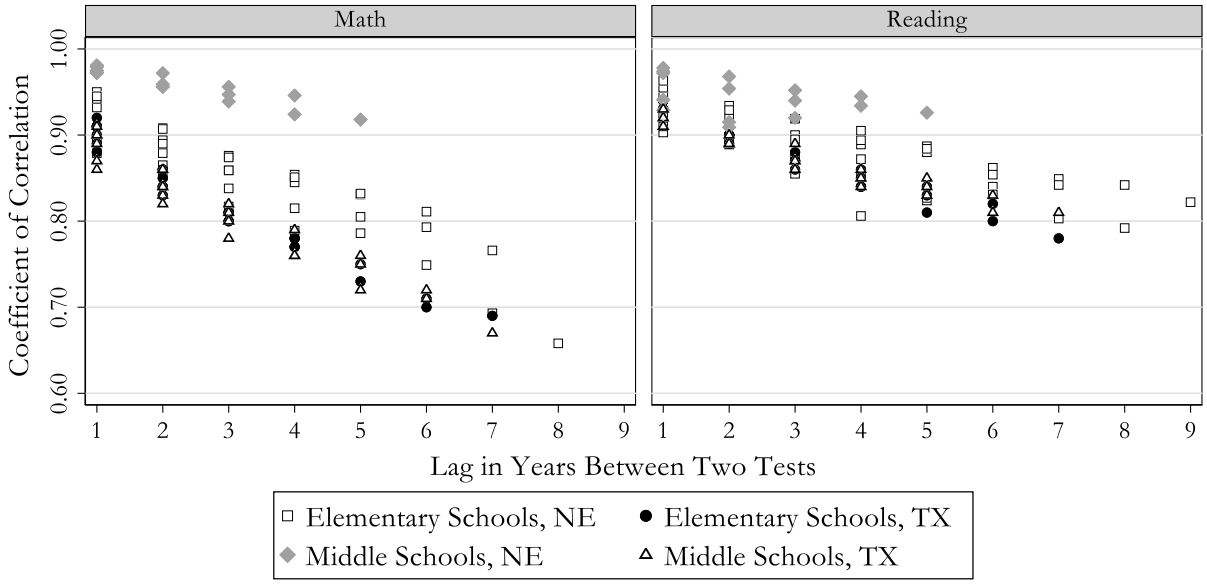


Note: N = 5,012 schools. Due to differences in test scales between subjects in Nebraska, the lower cluster in math and upper cluster in reading represents the average scores for individual schools in Nebraska (*N* = 268), with means ranging from 1,137 to 1,322 in math and 2,425 to 2,607 in reading. The larger cluster in both reading and math represents Texas (*N* = 4,744), with means ranging from 1,202 to 1,942 in math and from 1,186 to 1,860 in reading.

We also plotted the correlation of all reading and math scores for 2010–2019 for up to nine years' lag (Figure 2, below). The charts in Figure 2 visualize the high stability and slow decline in the bivariate correlations as the lag between the years of test scores increases. As we mentioned above, correlations for all mathematics and reading scores with a one-year lag (i.e., between two consecutive years) range from 0.87 to 0.98, with reading score correlations having a tighter distribution. As the lag increases, we observe both a decline in the absolute value of the correlations and also more dispersion. However, as can be seen in Figure 2, this fading out of the predictive power from more than one year back is slow. For mathematics, the correlations measured for average scores seven years apart remain very strong, ranging from 0.67 to 0.77. For reading, these correlations are even stronger, ranging from 0.78 to 0.85. This means that average scores in the same school on a standardized test taken by a completely different cohort of students *seven years prior* can still explain (and predict) from 45% to 72% of the test scores' variability in the current year. This remarkable stability of the predictive power of prior years' scores is especially evident in the reading correlations plotted in Figure 2.

Figure 2

Correlation of Elementary/Middle School Test Scores, 2010–2019, Nebraska and Texas



Note: N = 106. Lag indicates the number of years between the first and second test administrations, represented by a correlation. As the lag increases, the number of observations for that lag decreases as it would require a wider range for the data. As a result, we observe more correlations for shorter lags (correlations of one to three years) compared to greater lags.

The results above indicate that, on average, from one year to the next, very little new information is obtained from these scores. To understand if this consistency is common to all schools or if there is variation depending on changes in school demographics (considering, as mentioned above, the repeated evidence in prior research on the strong correlations between students' SES and their academic achievement), we examined whether the same phenomenon held true among schools that had experienced greater demographic shifts over the same time periods. For this analysis, we only used the data from Texas to allow for a sufficient number of observations in order to compare two groups of schools: (1) the top 10% of schools with the largest changes in SES (that is, those schools that were at or above the 90th percentile on the distribution of standard deviation of changes in schools' mean FRL status); and (2) all other schools. We also restricted the sample to schools for which we had data on all of the years included in the analysis, from 2014–2019 (average N = 239 for schools with the largest changes in SES over the observed period; average N = 2,114 for all other schools). We contrasted 60 pairs of correlations with test score lags from one to five years between the tests, in both reading and mathematics for elementary and middle schools. As described in our methods section above, we carried out a series of z-tests to assess whether there were significant differences in correlations between schools that had greater shifts in demographics and those that did not. In addition to comparing schools in the top 10% with all other schools, we compared them with the schools in the bottom 10%, namely, with those schools which experienced almost no changes in their student demographics over five years. We found that, in those schools which had undergone large changes in the socio-economic composition of their students, the year-to-year correlations were statistically significantly lower (Table 2, below). In other words, the

predictive value of the average test score was, indeed, sensitive to changes in school demographics. Nevertheless, even these correlations remained strong, with correlations up to two years later ranging between 0.68 and 0.79. Thus, while changes in school socioeconomic composition are significantly associated with average school performance on standardized assessments—and may explain much of what little variation exists between scores from year to year—even those schools which are experiencing large shifts in SES may be expected to yield highly correlated results over the course of two or three years.

Table 2

Correlations of Test Scores for Schools with the Largest Changes in SES vs. All Other Schools

		2014	2015	2016	2017	2018
Mathematics, Elementary Schools						
1 Year Later	Largest changes in SES	0.73	0.76	0.86	0.81	0.78
	All other schools	0.89	0.92	0.92	0.93	0.92
	z-statistic	7.10***	8.74***	4.11***	7.25***	7.38***
2 Years Later	Largest changes in SES	0.70	0.70	0.70	0.73	
	All other schools	0.85	0.86	0.87	0.87	
	z-statistic	5.60***	6.08***	6.44***	5.54***	
3 Years Later	Largest changes in SES	0.67	0.62	0.63		
	All other schools	0.82	0.82	0.82		
	z-statistic	4.78***	6.13***	6.30***		
4 Years Later	Largest changes in SES	0.56	0.56			
	All other schools	0.78	0.79			
	z-statistic	5.85***	6.46***			
5 Years Later	Largest changes in SES	0.54				
	All other schools	0.75				
	z-statistic	5.27***				
Mathematics, Middle Schools						
1 Year Later	Largest changes in SES	0.72	0.80	0.80	0.82	0.85
	All other schools	0.89	0.92	0.92	0.92	0.92
	z-statistic	8.09***	7.30***	6.95***	6.25***	5.32***
2 Years Later	Largest changes in SES	0.68	0.72	0.75	0.79	
	All other schools	0.86	0.87	0.86	0.87	
	z-statistic	6.89***	6.42**	4.88***	4.28***	
3 Years Later	Largest changes in SES	0.62	0.68	0.75		
	All other schools	0.83	0.82	0.83		
	z-statistic	7.08***	4.87***	3.08**		
4 Years Later	Largest changes in SES	0.56	0.68			
	All other schools	0.79	0.80			
	z-statistic	6.50***	3.80***			
5 Years Later	Largest changes in SES	0.59				
	All other schools	0.78				
	z-statistic	5.31***				

		2014	2015	2016	2017	2018
Reading, Elementary Schools						
1 Year Later	Largest changes in SES	0.71	0.82	0.84	0.84	0.72
	All other schools	0.94	0.94	0.94	0.94	0.94
	z-statistic	12.59***	8.52***	7.14***	7.77***	11.64***
2 Years Later	Largest changes in SES	0.70	0.72	0.72	0.74	
	All other schools	0.91	0.91	0.91	0.90	
	z-statistic	9.63***	8.55***	8.66***	7.43***	
3 Years Later	Largest changes in SES	0.64	0.63	0.64		
	All other schools	0.88	0.88	0.86		
	z-statistic	8.97***	9.50***	8.01***		
4 Years Later	Largest changes in SES	0.65	0.64			
	All other schools	0.87	0.85			
	z-statistic	7.85***	7.12***			
5 Years Later	Largest changes in SES	0.54				
	All other schools	0.83				
	z-statistic	8.28***				
Reading, Middle Schools						
1 Year Later	Largest changes in SES	0.82	0.80	0.80	0.85	0.84
	All other schools	0.94	0.94	0.94	0.94	0.95
	z-statistic	8.89***	9.62***	9.08***	7.75***	8.90***
2 Years Later	Largest changes in SES	0.70	0.74	0.76	0.78	
	All other schools	0.92	0.92	0.91	0.93	
	z-statistic	10.52***	9.38***	7.90***	8.84***	
3 Years Later	Largest changes in SES	0.67	0.69	0.70		
	All other schools	0.91	0.89	0.89		
	z-statistic	10.74***	8.61***	8.22***		
4 Years Later	Largest changes in SES	0.66	0.64			
	All other schools	0.89	0.88			
	z-statistic	9.85***	9.06***			
5 Years Later	Largest changes in SES	0.59				
	All other schools	0.87				
	z-statistic	9.88***				

Note: "Largest changes in SES" refers to the 10% of schools which experienced the largest degree of change in the socioeconomic status of their student body. These schools are at or above the 90th percentile on the distribution of standard deviation of changes in schools' mean FRL status. Only those schools from Texas that had data available for all years, from 2014 through 2019, are included. Sample sizes: 1) for elementary school mathematics: N (90th percentile) = 223, N (Other schools) = 4181; 2) for middle school mathematics: N (90th percentile) = 256, N (Other schools) = 2112; 3) for elementary school reading: N (90th percentile) = 223, N (Other schools) = 4182; 4) for middle school reading: N (90th percentile) = 254, N (Other schools) = 2116. Z-statistics are from the tests for the difference between correlations in two samples. *** $p < 0.001$. ** $p < 0.01$.

Conclusions

In this study, drawing upon data from two separate states, we find little informational gain between consecutive years of test results at a school level and, by generalization, at the district or state level. We find that year-to-year correlation coefficients for schools' standardized tests scores are consistently very high ($r \geq 0.87$). Furthermore, while schools experiencing demographic changes had significantly lower year-to-year correlations, even these correlations remained strong. The implication is that statewide mandates tell us what we already know: that, even in the midst of significant changes in factors such as SES—or in correlates such as funding, teacher retention, or student demographics such as first-language status—last year's score will be quite close to this year's score, again calling into question the informational value of the annual federal mandate. Further, as discussed above, schools are already investing in other formative measures since accountability tests—due to their end-of-year timing—fail to perform the role of a diagnostic tool for individual students (Kahl, 2021).

All of this then raises the question as to whether an *annual* accountability measure is truly necessary. Our results above suggest that ESSA's annual testing requirement yields little new information on a year-to-year basis. And, given the similar patterns in year-to-year correlations in two states as different in size, demographics, and assessments as Texas and Nebraska, we expect to find the same pattern of school-level scores on federally mandated achievement tests in other states as well. Reducing the frequency of these tests to every second or third year would afford little loss of data for policy making while freeing up fiscal resources to be more productively spent elsewhere. Such a reduction in the frequency of testing would simultaneously reduce teacher anxiety and provide additional time for improving teaching and learning—the very process about which accountability measures are most concerned.

Acknowledgements

The authors thank Bert Peterson, a concerned Nebraska citizen, for alerting us to the issues we address in this paper.

References

Amrein-Beardsley, A. (2022). Using standardized tests for educational accountability: The bad idea that keeps on giving nothing in return. *Journal of Policy Analysis and Management.* https://doi.org/10.1002/pam.22426

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Baker, M., & Johnston, P. (2010). The impact of socioeconomic status on high stakes testing reexamined. *Journal of Instructional Psychology*, *37*(3), 193. https://eric.ed.gov/?id=EJ952120

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*(3), 287–302. https://doi.org/10.1080/0305764X.2011.607151

Black, D. W. (2017). Abandoning the federal role in education: The Every Student Succeeds Act. *California Law Review, 105*(5), 1309–1374. https://doi.org/10.15779/Z38Z31NN9K

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Cahill, P. (2019). *How do school educators experience high-stakes testing? A case study of the Australian New South Wales Higher School Certificate in two independent schools* [Ed.D. dissertation, Australian Catholic University]. https://doi.org/10.26199/ACU.8VYQ9

Cavanagh, S. (2017, May 31). NWEA poised to capture statewide testing contract in Nebraska. *EDWeek Market Brief*. https://marketbrief.edweek.org/marketplace-k-12/nwea-poised-win-statewide-testing-contract-nebraska/

Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003*). Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum.

Cunningham, W. G., & Sanzo, T. D. (2002). Is high-stakes testing harming lower socioeconomic status schools? *NASSP Bulletin, 86*(631), 62–75. https://doi.org/10.1177/019263650208663106

Datnow, A., & Park, V. (2018). Opening or closing doors for students? Equity and data use in schools. *Journal of Educational Change, 19*(2), 131–152. https://doi.org/10.1007/s10833-018-9323-6

Datnow, A., Park, V., & Choi, B. (2018). "Everyone's Responsibility 1": Effective team collaboration and data use. In N. Barnes & H. Fives (Eds.), *Cases of teachers' data use* (pp. 145–161). Routledge. https://doi.org/10.4324/9781315165370-10

Edgerton, A. K. (2019). The essence of ESSA: More control at the district level? *Phi Delta Kappan, 101*(2), 14–17. https://doi.org/10.1177/0031721719879148

Egalite, A. J., Fuseli, L. D., & Fusarelli, B. C. (2017). Will decentralization affect educational inequity? The Every Student Succeeds Act. *Educational Administration Quarterly, 53*(5), 757–781. https://doi.org/10.1177/0013161X17735869

Embse, N. P. von der, Schoemann, A. M., Kilgus, S. P., Wicoff, M., & Bowler, M. (2017). The influence of test-based accountability policies on teacher stress and instructional practices: A moderated mediation model. *Educational Psychology*, *37*(3), 312-331. https://doi.org/10.1080/01443410.2016.1183766

Every Student Succeeds Act, P.L. 119-95, 20 U.S.C. § 6301 (2015). https://www.congress.gov/bill/114th-congress/senate-bill/1177

Fryer, R. G., Jr., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *The Review of Economics and Statistics, 86*(2), 447–464. https://doi.org/10.1162/003465304323031049

Garner, B., Thorne, J. K., & Horn, I. S. (2017). Teachers interpreting data for instructional decisions: Where does equity come in? *Journal of Educational Administration, 55*(4), 407–426. https://doi.org/10.1108/JEA-09-2016-0106

Georgia Governor's Office of Student Achievement. (2021). *Georgia school grades report (2018-2019)*. https://schoolgrades.georgia.gov/school-search

Heissel, J. A., Levy, D. J., & Adam, E. K. (2017). Stress, sleep, and performance on standardized tests: Understudied pathways to the achievement gap. *AERA Open, 3*(3), 2332858417713488. https://doi.org/10.1177/2332858417713488

Heissel, J. A., Adam, E. K., Doleac, J. L., Figlio, D. N., & Meer, J. (2021). Testing, stress, and performance: How students respond physiologically to high-stakes testing. *Education Finance and Policy*, *16*(2), 183–208. https://doi.org/10.1162/edfp_a_00306

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice*, *12*(4), 20-25. http://doi.org/10.1111/j.1745-3992.1993.tb00550.x

Hoover, N. R., & Abrams, L. M. (2013). Teachers' instructional use of summative student assessment data. *Applied Measurement in Education, 26*(3), 219–231. https://doi.org/10.1080/08957347.2013.793187

Ingram, D., Louis, K. S., & Schroeder, R. G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record, 106*(6), 1258–1287. https://doi.org/10.1111/j.1467-9620.2004.00379.x

Jacobsen, R., Snyder, J. W., & Saultz, A. (2014). Informing or shaping public opinion? The influence of school accountability data format on public perceptions of school quality. *American Journal of Education*, *121*(1), 1-27. https://doi.org/10.1086/678136

Jespersen, D. (2018, May 21). First year of NSCAS summative assessment successfully concludes. *Nebraska Department of Education*. https://www.nwea.org/content/uploads/2018/05/NSCAS-Summative-First-Year-Concludes-Successfully_5_21_2018.pdf

Kahl, S. (2021, May 5). What federally mandated state tests are good for. *Education Week*. https://www.edweek.org/teaching-learning/opinion-what-federally-mandated-state-tests-are-good-for-and-what-they-arent/2021/05

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives*, *8*(49), 1–22. https://doi.org/10.14507/epaa.v8n49.2000

Lai, M. K., & Schildkamp, K. (2016). In-service teacher professional learning: Use of assessment in data-based decision-making. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 77–94). Routledge. https://doi.org/10.4324/9781315749136

Mandinach, E. B., & Schildkamp, K. (2021). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in Educational Evaluation, 69*, 100842. https://doi.org/10.1016/j.stueduc.2020.100842

Mandinach, E. B., Warner, S., & Mundry, S. E. (2019, November 6). *Using data to promote culturally responsive teaching* [Webinar]. Regional Educational Laboratory Northeast and Islands, Education Development Center. https://ies.ed.gov/ncee/edlabs/regions/northeast/Docs/Events/CRDL_Workshop_Sep_30_2019_508c.pdf

Mintrop, H., & Sunderman, G. L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement—and why we may retain it anyway. *Educational Researcher*, *38*(5), 353-364. https://doi.org/10.3102/0013189X09339055

Mississippi Department of Education. (2021.) *Mississippi Succeeds report card.* https://msrc.mdek12.org/Index

Murray, K., & Howe, K. R. (2017). Neglecting democracy in education policy: A-F school report card accountability systems. *Education Policy Analysis Archives*, *25*(109). http://doi.org/10.14507/epaa.25.3017

Nathaniel, P., Pendergast, L. L., Segool, N., Saeki, E., & Ryan, S. (2016). The influence of test-based accountability policies on school climate and teacher stress across four states. *Teaching and Teacher Education*, *59*, 492-502. https://doi.org/10.1016/j.tate.2016.07.013

Nebraska Department of Education. (2021). *NEP data downloads.* https://nep.education.ne.gov/Links

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002). https://www.congress.gov/bill/107th-congress/house-bill/1

Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record, 112*(4), 1137–1162. https://doi.org/10.1177/016146811011200401

Polesel, J., Dulfer, N., & Turnbull, M. J. (2012). *The experience of education: The impacts of high stakes testing on school students and their families* [Literature review]. The Whitlam Institute, Melbourne Graduate School of Education, University of Western Sydney. https://www.whitlam.org/publications/2017/10/17/the-experience-of-education-the-impacts-of-high-stakes-testing-on-school-students-and-their-families

Polesel, J., Rice, S., & Dulfer, N. (2014). The impact of high-stakes testing on curriculum and pedagogy: A teacher perspective from Australia. *Journal of Education Policy, 29*(5), 640–657. https://doi.org/10.1080/02680939.2013.865082

Reardon. S. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. Duncan & R. Murnane (Eds.) *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 91-116). Russell Sage Foundation. https://www.russellsage.org/publications/whither-opportunity

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education, 26*(3), 482–496. https://doi.org/10.1016/j.tate.2009.06.007

Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools*, *50*(5), 489-499. https://doi.org/10.1002/pits.21689

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Sun, J., Przybylski, R., & Johnson, B. J. (2016). A review of research on teachers' use of student data: From the perspective of school leadership. *Educational Assessment, Evaluation and Accountability*, *28*(1), 5-33. https://doi.org/10.1007/s11092-016-9238-9

Sutherland, D. H. (2022). School board sensemaking of federal and state accountability policies. *Educational Policy, 36*(5), 981–1010. https://doi.org/10.1177/0895904820925816

Swaby, A. (2021, January 4). Texas hires two companies to run STAAR, moving toward statewide online testing. *The Texas Tribune*. https://www.texastribune.org/2021/01/04/texas-staar-tests

Texas Education Agency. (2021a). *Texas school finder*. https://txschools.gov

Texas Education Agency. (2021b). *STAAR aggregate data*. https://tea.texas.gov/student-assessment/testing/staar/staar-aggregate-data

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461–481. https://doi.org/10.1037/0033-2909.91.3.461

About the Authors

Norman P. Gibbs
Mesa Unified School District
npgibbs@mpsaz.org
https://orcid.org/0000-0002-5064-2816
Norman P. Gibbs is a program evaluator for Mesa Unified School District, Mesa, Arizona. His research focuses on assessment and accountability, comparative and international education, and inclusive and participatory decision-making.

Margarita Pivovarova
Arizona State University
mpivovar@asu.edu
https://orcid.org/0000-0002-2965-7423
Margarita Pivovarova is an associate professor in the Mary Lou Fulton Teachers College at Arizona State University. Her research focuses on the relationship between student achievement, teacher quality, and school contextual factors.

David C. Berliner
Arizona State University
berliner@asu.edu
https://orcid.org/0000-0003-3930-7280
David C. Berliner is Regents' Professor Emeritus at the Mary Lou Fulton College of Education. He has published in educational psychology, teacher education and educational policy. He is a past president of the American Educational Research Association and a member of both the National and International Academies of Education.

education policy analysis archives

Volume 31 Number 10 February 7, 2023 ISSN 1068-2341