The A-F Accountability Mistake

John Tanner, Test Sense

November 2016

In fall 2017, Texas will join 16 other states in implementing a public school rating system that assigns letter grades to schools and districts. By December 1, 2016, the Texas Education Agency (TEA) must adopt indicators showing how the A-F ratings will be determined, and by January 1, 2017, TEA must submit a report to the Texas House and Senate Education Committees showing the ratings that schools and districts would have been given if the system had been in place for the 2015–16 school year.

As we begin this important rule-making period, and as another Texas Legislature with the authority to change the law that established Texas' A-F system prepares to meet, it is imperative that stakeholders know that the research is clear: A-F school rating systems fail as an indicator of school quality, but there is evidence that supports more meaningful kinds of accountability systems.

This essay provides an overview of A-F systems and their failures. In addition, to question A-F systems is to question test-based accountability, and criticisms of controversial topics are most likely to be heard when solutions accompany the critique, so this essay is the first of three in the Texas Accountability Series, published by the Texas Association of School Administrators. The other essays in the series cover: why, to be meaningful, school accountability must be community-based and not solely focused on compliance with state testing mandates (see "Creating a Meaningful Community-Based Accountability System"); and the misfit of state testing programs with school accountability (see "The Misfit Between Testing and Accountability"). Each of these essays was written by John Tanner, executive director of Test Sense and author of The Pitfalls of Reform.

As additional issues related to school accountability arise, the series will be continued to ensure that Texas educators have the information they need to work with policymakers and the public in a meaningful way.

Executive Summary

The argument: The reduction of school quality to a single mark is the purpose of A-F school rating systems. The argument is that a grade will signal a level of quality and make it difficult for low-rated schools to escape scrutiny. Advocates of such rating systems use terms such as "simple," "clear," and "transparent" to describe them, and frequently cite competition and subsequent improvement as key outcomes. Former Gov. Jeb Bush of Florida and a number of organizations he supports are the most vocal proponents of such systems. Florida adopted its system in 1999 and sixteen other states have since followed. Texas is scheduled to implement its A-F rating system for the 2017-18 school year.

Research on such systems is surprisingly inadequate given the prevalence of A-F as a policy tool. What does exist is almost universally negative. Florida cites significant gains in the first few years of its program, a fact that is a primary argument in support of such systems. Nevertheless, by Florida's own admission, the majority of the "gains" were due to changes in the rules, a fact not shared with the Texas Commission on Next-Generation Assessments and Accountability when the Bush-supported organizations offered testimony on this topic in 2016.

Most states with A-F rating systems have adjusted the rules to their systems following implementation so the results more closely match the public and policymakers' expectations for the distribution of grades. These adjustments call into question the logic behind such systems: It appears they are only declared successful once they reflect a preconceived notion of expectations, not an objective reality.

Preferred Citation: Tanner, J., (2016). The A-F Accountability Mistake. The Texas Accountability Series. Austin, TX: The Texas Association of School Administrators.

The few basic rules behind A-F appear simple on the surface but generate an inordinate number of behind-the-scenes calculations and numerous additional rules that render the results unusable for informing change. In many cases schools that perform in a statistically similar manner are awarded vastly different grades, while schools that perform quite differently are awarded similar grades. The reduction to a single grade tends to downplay achievement gaps. In a study of the Oklahoma system, gaps were shown to be wider in higher graded schools than in lower graded schools, and lower graded schools were shown to be performing better with subgroups than higher graded schools.

Based heavily on standardized test scores, A-F school rating systems tend to assign grades in which the socioeconomic status of the school is the single best predictor of the grade, ignoring the efforts being made in some of the most challenged educational environments.

The reduction of a school to a single grade has the tendency to color the judgments and subsequent actions of the entire school, even though each school is a diverse place with the need to serve all students. Reducing a school to a single grade has the predictable effect of telling a school with a good grade that all is well and telling a school with a bad grade that all must change, even though neither can ever be accurate.

Conclusion: Rating schools and districts with A-F letter grades is a policy idea that fails every criterion put forth as a reason for having it. It is neither simple nor transparent. It misrepresents a large proportion of what happens in schools by reducing an entire school to a single mark that can only be partially appropriate given the complexity of schooling. In the end, A-F school ratings do more harm than good. They create confusion among educators, and fail to offer the public useful or accurate information about their schools.

Florida and the Tweaking of School Grades

School grading appeals to policymakers and the public as a sort of consumer's guide to schools, with both groups believing that such grades make clear what is otherwise vague and not easily understood.[1] The idea is to provide an objective indicator of quality. For schools that receive low grades, the idea is to insist upon change. For schools with high grades, the idea is to reward excellence.

Florida started the test-based rating trend in 1999 when Gov. Jeb Bush pushed for an A-F system for schools. Since then, sixteen states, including Texas starting with the 2017-18 school year, have adopted some form of the idea (some have also since moved away from it). One of the educational organizations founded by Gov. Bush after he left office, the Foundation for Excellence in Education, advocates strongly for the model, declaring it "simple and transparent."[2] In some states, test scores are the sole contributor to the letter grades, while in others an additional indicator such as attendance or a calculated graduation rate is included. In virtually all cases, the law is itself reasonably simple, but the rules for how to arrive at the grades tend to be lengthy and complex.[3]

The increase in the percentage of A's being earned during the first few years of the Florida A-F system is frequently cited as evidence that the policy is an effective one.[4] And the change during that time period was dramatic: The percentage of schools receiving A's rose from twelve to fifty-three percent. In 1999 when the program was first implemented, schools had to show that only a minimum number of students were performing at the lowest possible levels or the A would not be available to them. This was true across subgroups, meaning that any poor subgroup performance prevented a school from earning an A. In 2000, claiming that the rules were too harsh, the state changed this rule so that schools had to show only that the number of students at the lowest level was decreasing, and it removed the subgroup requirement. The result was that during a single year the number of schools receiving A's nearly tripled.

"In virtually all cases, the law is itself reasonably simple, but the rules for how to arrive at the grades tend to be lengthy and complex."

In 2002 the rules changed again, placing additional emphasis on growth, leading to another significant jump in schools receiving A's. This time, however, an analysis performed by Mathew DiCarlo at the Albert Shanker Institute showed, in a simulation, that if the 2001 system had been continued the 2002 grades would not have differed significantly. Nevertheless, by that point fifty-three percent of schools were receiving A's,[5] considerably more than the initial twelve percent. To call that significant growth brought about by A-F school grades defies logic: The majority of the difference was created through a change in the rules, not a change in the educational system. As a result, the primary claim that grades would create incentives for improvement was true only if the vehicle for improvement was a rules change by the state.

The Florida example shows how easily manipulated such systems are. While that represents the norm for rule-driven systems, it is also the reason that such changes must be made with a high degree of discipline and only after very careful consideration. Otherwise, the underlying reality risks being misrepresented.

The rules-based system for financial accounting is one such system that operates under those exact constraints and offers a valuable lesson. Financial accountants regularly adjust the business rules for financial statements so that they more accurately reflect the fiscal status of an organization. Each proposed change is examined in a disciplined way for its ability to increase the capacity of the statements to accurately represent the underlying fiscal status. Each change is accepted only when it can be shown to increase the accuracy in the reports.

Now imagine that those same accountants applied the Florida model of tweaking to produce a desired or an acceptable result. Consider the implications if a CEO, unhappy with the quarterly results, asks an accountant to tweak the rules by which they are calculated so that the numbers look like what the shareholders expect. The numbers would then match a feeling or a bias and for a brief, delusional moment the shareholders will feel good, but the simple fact is that no one will have any sense of the underlying reality, and the CEO and the accountant would likely land in jail.

"…Florida isn't the only state to adjust its rules so that the results more closely align with a perceived bias."

Florida isn't the only state to adjust its rules so that the results more closely align with a perceived bias. A-F states frequently tweak the rules behind their systems to produce results that better match some preconceived or politically acceptable expectation.[6] But what then is the underlying reality that the tweaking is attempting to more accurately present? The only answer is that there is no reality; just a bias as to what things should look like. It just so happened that the grades didn't "look right" or "feel fair" and so adjustments were made, not in the direction of a more accurate reflection of reality,[7] but in the direction of someone's opinion for what the system should show.

Another example of this is from 2012, when Indiana state superintendent Tony Bennett noticed that a charter school founded by one of his campaign donors was about to be awarded a C, while Bennett, knowing the school personally, believed that was an inappropriate grade. Bennett instructed his staff to tweak the system to produce a result he deemed acceptable and they did, changing the system to fit the superintendent's bias. The following year Bennett became the state superintendent in Florida. A week into the job news broke that he had directed his staff back in Indiana to tweak the system for the benefit of one school, forcing him to resign.[8]

Bennett's mistake was that his bias was too narrow: Had he recast the rules to match a system-wide bias he would have been able to claim he was simply doing what Florida and so many others had done before.

In none of the above situations can the argument be made that the changes resulted in a more accurate picture because there is no way of adjudicating what accuracy is.[9] The validity of A-F is determined via perception: A good school is one already perceived to be a good school and the system will be deemed

"…nobody understands what the assignment of a school grade actually means."

valid if it is awarded a good grade. The opposite is true with bad schools. In fact, one can argue that the need among certain stakeholders for the system to generate a pattern in line with preconceived biases on its own threatens the integrity of the system from the beginning. Political expediency cannot help but be a driving force in such situations, which in turn admits that nobody understands what the assignment of a school grade actually means.

The Type A Effect

A team researching the Oklahoma A-F system[10] examined a particularly important aspect of accountability systems called the "Type A Effect." The Type A Effect consists of the effects of education attributable to the school, as opposed to those attributable to external forces that the school does not control.[11]

One can know if an accountability program is addressing the Type A Effect or something else in the following way: "[D]ifferences in student test performance should be substantial and persist after controlling for factors unrelated to teaching effectiveness or school practices."[12] If you can demonstrate that a particular student would likely perform differently in a "typical" school than the one he or she is in, you will have identified a Type A Effect. If the student would likely do better in the "typical" school, then the sending school needs to accept that it should do better. If the student would likely do worse, then the sending school should be commended for having had a meaningful impact on the student.

"Failure to limit school accountability to Type A Effects means that the school risks being held accountable for things it does not control."

Failure to limit school accountability to Type A Effects means that the school risks being held accountable for things it does not control. For example, the acquisition of skills such as numeracy and literacy is determined largely by the amount and quality of practice lifetime to date, which is determined by both schooling and non-schooling effects. The simplest example of this is when a student comes to school knowing his or her ABCs: This is clearly a step toward literacy that occurred outside of school. During each year of schooling as literacy and numeracy skills increase, some of that will be due to schooling, and some of that will be due to factors outside of schooling.

Poverty, for example, has a significant impact on the access to such practices outside of regular schooling, which leads to the predictable result that schools in poorer neighborhoods are likely to compare unfavorably to schools in wealthier neighborhoods on any numeracy or literacy indicator as of a particular date in time. It would be reasonable to think that, if we could adjust those results and compare only those students with a similar history of practice, the results would be comparable. In other words, the difference between the two schools as of a particular date is that the students in one have had better and more practice than in the other. That cannot on its own make either a good school or bad school.

"…the result is that educators in wealthy schools are likely to be rewarded for teaching in wealthier schools, while educators in poor schools are likely to be punished for teaching in poorer schools, regardless of the quality of the decisions being made..."

When accountability—based as it is largely on annual measures of numeracy and literacy—fails to limit itself to Type A Effects, the result is that educators in wealthy schools are likely to be rewarded for teaching in wealthier schools, while educators in poor schools are likely to be punished for teaching in poorer schools, regardless of the quality of the decisions being made, or their execution. A-F school ratings or labels will be based largely on factors those working in a school do not control, thereby ignoring those factors they do control.

The way to see this is again to test for the Type A Effect: Would a student from a lower graded school perform better in a higher graded school? If yes, then the grade is a signal of quality and the

higher graded school can be said to be "better" than the lower graded school. If no, then the grade is based not on the quality of the school or what it controls, but on something else.

The Oklahoma team tested their system to see if the Type A Effect was the basis for their assigned grades or if it was something else. They were clear in their findings: "We cannot conclude from our evidence … that the average student in a C, D, or F school would have performed any better in math and reading than in an A or B school."[13] If the grades were indeed identifying a quality component of schooling, one would expect that A or B schools would be better for an average student from C or D schools, but that was not the case: There was no difference.

"They were clear in their findings: 'We cannot conclude from our evidence … that the average student in a C, D, or F school would have performed any better in math and reading than in an A or B school.'"

In other words, the Oklahoma system can be shown empirically to fail in holding schools accountable for the quality of their decisions and the things they control. Instead, it declares that A or B schools are better for students than C or D schools when the facts don't support that conclusion. School grades in Oklahoma are awarded, then, based in large part on factors that have nothing to do with the actual efforts in a school or the quality of the decisions being made. The fact that the Texas system has a good bit in common with the Oklahoma system should be seen as disconcerting for this very reason.[14]

The Lack of Usefulness

A primary argument made for the use of A-F is that the system will add clarity and objectivity useful to the public, policymakers, and educators alike.[15] Low-performing schools will be forced to own that performance and make changes, while high-performing schools will receive the accolades they deserve.

The Type A Effect notwithstanding, that clarity is a myth. Consider that the Florida system, the most mature of the A-F programs, has a set of rules that spans thirty-one pages, and includes dozens of computations and page after page of requirements and explanations for how a grade is to be determined. One of the more interesting is the description of learning gains, which are "based on the percentage of students who met one or more of the following," included here for illustrative purposes:

- Students who increase at least one (1) achievement level on the statewide standardized assessment in the same subject area.
- Students who scored below Achievement Level 3 on the statewide standardized assessment in the prior year and who advance from one subcategory within Achievement Level 1 or 2 in the prior year to a higher subcategory in the current year in same subject area. See the tables below [not included here] for the scores that comprise each subcategory. Achievement Level 1 is comprised of three (3) subcategories, and Achievement Level 2 is comprised of two (2) subcategories; subcategories are determined by dividing the scale of Achievement Level 1 into three (3) equal parts and dividing the scale of Achievement Level 2 into two (2) equal parts. If the scale range cannot be evenly divided into three (3) equal parts for Achievement Level 1 or into two (2) equal parts for Achievement Level 2, no subcategory may be more than one (1) scale score point larger than the other subcategories; the highest subcategories shall be the smallest.
- Students whose score remained at Achievement Level 3 or 4 on the statewide standardized assessment in the current year and whose scale score is greater in the current year than the prior year in the same subject area. This does not apply to students who scored in a different achievement level in the prior year in the same subject area.
- Students who take a FSA EOC assessment and remained at Achievement Level 3 or Achievement Level 4.

- Students who scored at Achievement Level 5 in the prior year on the statewide standardized assessment and who score in Achievement Level 5 in the current year in the same subject area.[16]

The description above applies to four of the components that make up a district grade. These are combined with seven others and summed to produce the final tally. For each of the other seven there are additional rules for how to perform the calculations, adding layer after layer of complexity.

Alabama is the latest to go through the process of an A-F adoption, with their first school grades to be released as a phase-in to a larger report card system in December 2016. The simple version of the requirement to be carried out by the state superintendent is this:

> In order to create … a total profile of the school or school system, or both, a school's grade, at a minimum shall be based on a combination of student achievement scores, achievement gap, college and career readiness, learning gains, and other indicators as determined by the State Superintendent of Education to impact learning and success.[17]

The Alabama State Department of Education is still working through the process for how to do this, but consider just one set of rules for calculating achievement presented in a draft document (each of the areas to be combined has a similar set of rules as well, and a set of rules for how they will be combined into a grade):

1. Identify the number of students that made Average and High Growth.
2. Multiply the total for each growth category by the appropriate multiplier.
3. Add the total for each multiplier together to obtain the weighted sum.
4. Divide the sum by the number of total students with a growth category (L,A,H) = (percent).
5. Multiply by points possible to get the number of Points Earned.
6. Complete the above steps for Reading and Math Weighted Learning Gains Calculations.
7. Add the Reading and Math points earned together to get the total points earned.[18]

"…the argument that the results will produce easily understood school grades that are clear, objective, and useful for directing change has to be viewed as wrong."

Note that much more remains to be done to finish out the system, as Alabama, like Florida, will have to determine numerators and denominators and consider many more details and formulas and calculations to reduce all of that complexity to a grade. As a result, the argument that the results will produce easily understood school grades that are clear, objective, and useful for directing change has to be viewed as wrong. Collapsing the thousands of computations it will take to produce one of five grades offers virtually no capacity for a school administrator to know how to affect the types of changes needed. It offers a potentially flawed view of a school to the public, and if the bias of policymakers enters the fray it just confuses the issue more.

"…the argument that the results will produce easily understood school grades that are clear, objective, and useful for directing change has to be viewed as wrong."

In a report out of the National Education Policy Center at the University of Colorado, the authors thoroughly debunk all claims to the validity of such A-F systems based largely on the level of complexity in performing the reductions. They argue what is being argued here, that while test-based A-F school rating systems appear to be highly intuitive, even a cursory investigation shows that to be

"The Colorado report focused on A-F systems as a tool for identifying school quality, their validity as a policy instrument, as well as their validity in supporting schools as they prepare students to participate in our democracy. …Their conclusion is that for each of these purposes A-F systems are thoroughly invalid."

otherwise. The Colorado report focused on A-F systems as a tool for identifying school quality, their validity as a policy instrument, as well as their validity in supporting schools as they prepare students to participate in our democracy (often referred to as college and workplace readiness). Their conclusion is that for each of these purposes A-F systems are thoroughly invalid.[19]

Whitewashing Meaningful Variances

The variance that exists within a school can be enormous. The full range of that variance can and must be identified, and each piece managed appropriately. Within any school are successes as well as areas for improvement covering every aspect of schooling. It is imperative that the variance is a visible part of an accountability system or changes risk being both generic and counterproductive.

What tends to happen when a school is reduced to a grade is that the variance in the system is treated as if it can be summed up in that one grade. The simple version of this would be for an A-graded school to declare success and repeat the work of the previous year, while a D-graded school might be declared as near failing and decide to change everything. Both sets of decisions risk being dead wrong. No matter a grade, every school needs to make a range of decisions that address the variances. Pretending there is no variance marginalizes the good decisions being made in the "D" school, while excusing the poor ones being made in the "A" school.

Consider just one of the many examples of this phenomena identified, again in the Oklahoma system:

> "[A]chievement gaps in reading and math were larger in A and B schools than C, D, and F schools. Also, reading and math achievement of minority and FRL [Free and Reduced Price Lunch] students remained higher in the lowest ranked schools (D and F) compared to the highest ranked schools (A and B). D and F schools were on average more effective with FRL and minority students."[20]

What should be obvious is that the assigned grades sent the wrong messages: The higher grades signaled success in schools actually underserving their at-risk populations, while the lower grades signaled failure in schools that are at the very least more effective with those populations than those receiving A's and B's. Should the schools act on their grades, the gaps are likely to continue in the higher graded schools, and the lower graded schools risk changing the parts of their systems that are working.

"…the reduction of the quality of a school to a single indicator will always be misleading… A-F systems put at risk the ability of schools to meet actual student needs."

The variance problem also happens in reverse: Grades that are supposed to signal meaningful variances frequently don't. The Oklahoma study found that in many instances schools receiving the grades of B, C, and D had actual student performances that were "nearly identical."[21] Whatever intended meaning was assigned to each grade, then, is both statistically and intellectually meaningless. School grades, which are supposed to signal meaningful differences, frequently fail even to do that.[22]

In short, the reduction of the quality of a school to a single indicator will always be misleading.[23] A-F systems both wash out the variance it is the school leader's responsibility to address in whatever nuanced educational environment they operate in, and frequently fail to identify actual differences. Whether pretending that no variance exists, or creating the illusion of a variance where there is none, A-F systems put at risk the ability of schools to meet actual student needs.[24]

The Risk of Combining Unlike Things

One purpose of A-F systems is to condense a variety of educational indicators into a single grade that can quickly and easily signal educational quality. The desire for a single metric is understandable: Education is an extraordinarily complicated endeavor and trying to ascertain quality requires an

expertise few people have. The idea behind a grade is that the necessary expertise can be brought to bear across multiple indicators via a set of calculations and algorithms such that the grade makes clear what would otherwise be impossible to understand.

In reality, the idea that a grade will be meaningful quickly falls apart for a fairly simple reason: It has to reduce each indicator to a blunt tool in order to combine them, removing the ability of an indicator to do whatever it was designed to do. Height and college completion rates, for example, are both indicators that on their own have meaning, but if combined result in a meaningless number. The fact that doing so is mathematically possible does not make the results useful: The results would be impossible to explain.[25]

"…the idea that a grade will be meaningful quickly falls apart for a fairly simple reason: It has to reduce each indicator to a blunt tool in order to combine them, removing the ability of an indicator to do whatever it was designed to do."

Even metrics that are seemingly related will experience the same thing in any attempt to combine them. Consider what could happen if a system tried to combine graduation rates and achievement into a single metric. Mathematically it is possible to assign weights to both metrics, combine them to produce a number for each school, and then rank schools from the highest to the lowest. The highest could be awarded A's, the lowest F's, and a few additional lines could be drawn to delineate the remaining grades.

But what would the assigned grade mean? Achievement and graduation rates often show very different things when viewed against each other. As graduation rates in a school rise, the population making up that rise tends to come from a demographic that historically struggles on achievement measures. As more and more of those students are kept in school and supported through graduation—clearly a good thing—their inclusion in the achievement measure will tend to depress average scores, even if achievement for all the other students remains stable or even improves somewhat.

If, on the other hand, graduation rates fall in another school, it is possible that averages for achievement measures will rise, given that a few of the at risk-students will no longer be there to pull scores down. In that case the rising test scores are the result of having lost some students, and rewarding that would be wrong: In fact, the school should need to answer for the loss of those students.

Combining the metrics blunts the ability to offer accurate interpretations: For example, it would be accurate to say that the first school succeeded in serving a group of at-risk students while the second did not. In light of that fact, the rise and fall in test scores is not something that should be judged: The fall in the average in the first school does not signal a negative, and the rise in the second does not signal a positive.

Reduced to a single grade, that important subtlety disappears: Each metric would contribute to the grade independent of the other, resulting in both schools appearing to have one success and one failure. Should that result in similar grades, those grades would seriously misrepresent the quality of the efforts in both schools, which a thoughtful examination of both metrics makes obvious.

While an argument can be made that perhaps such nuance could be inserted into a system via algorithms (if/then statements that would apply this sort of logic when certain conditions are met), the universe of possibilities is practically endless: As a result, any attempt at a complete set of rules would necessarily fail. Consider a relatively simple example regarding achievement and growth: Imagine a school in which average achievement remains stable over several years of a demographic shift in which the percentage of students receiving free and reduced price lunch increases each year. That would likely mean that the school is adding real value to those students' lives from the moment they enroll in the school, given that the average remained stable even as more and more at risk students were included in the results. A blunt interpretation of achievement that failed to take into account the demographic shift risks signaling otherwise. If the demographic movement is toward a somewhat

wealthier population, both achievement and growth may rise, but any sort of reward or accolade may be unwarranted given that the effort of the school had little to do with it.

Absent a nuanced interpretation of each metric in light of other information, the risk is that the first school will be punished and the second rewarded even though the reward should go to the first school and the second should be focused on improvement. Pushing the two metrics into a single grade warps reality: In a system that naively rewarded high or rising test scores absent other information, the first school would be judged as the one that needs to improve, and the second would be rewarded for its demographic change. It would be unfair to award a higher grade to the second school, and yet that would likely be the case absent the information required to make a proper interpretation.

"The practical result is that educators will be unable to understand or explain why a grade was assigned, or know what sorts of changes might lead to a higher grade. That will have the further effect of making educators look bad at not being able to explain what the public will consider a simple measure of school quality."

These examples each suffer from the fact that they are far simpler than what actually occurs in reality. Understanding and interpreting growth, absolute measures, graduation rates, attendance, or any other indicator, requires a great deal of nuanced information that is not a part of the measures in order to be properly understood and interpreted. The combination of multiple unlike things into a single indicator removes those nuances. The outcome can only be results that are at best inaccurate, and at worst dead wrong. To assign such an indicator the job of ascertaining quality risks communicating to schools that they are doing a good or a bad job, or being effective or ineffective, without knowing whether or not that is actually true.[26]

Consider as well that the grade of a school combines the types of results described above across multiple grades and subjects, further exacerbating the inability of the grade to mean anything. The practical result is that educators will be unable to understand or explain why a grade was assigned, or know what sorts of changes might lead to a higher grade. That will have the further effect of making educators look bad at not being able to explain what the public will consider a simple measure of school quality.

As with so much of the accountability conversation, the burden for such explanations will fall primarily on those serving at-risk populations, who when they struggle to explain the grades assigned their schools will confirm in the public's eye the reasons for the struggle: Those in the school don't know what is going on. Wealthier schools will be just as helpless to explain the causes for their assigned grades, only the odds are they won't have to.

The Texas System

The Texas A-F campus and district rating system, as signed into law, builds upon those that have gone before it. It includes the requirement for domain grades, as well as overall campus and district grades. For the campus and district grades, state standardized test scores represent the majority component (both as an absolute and a growth measure) and secondary components such as attendance or graduation rates also are figured in. The law also states that while fifty-five percent of each campus and district rating will come from test scores, several rules add significant weight to low test scores, given that any campus performing at the D or F levels at either a domain or overall will prevent a school or district from obtaining an A grade in either the domain or overall.

In addition, a statement in the law suggests that rules need to be adopted that prevent repeat unacceptable performances from being glossed over by strong performances elsewhere,[27] again suggesting that while the initial weighting in the formulas will seem reasonable, low test scores will still continue to be a source of negative consequences, giving them a practical weight in excess of the initial fifty-five percent assigned formally in the law.

The Texas A-F system carries with it all the risks inherent in such systems: It stands to ignore the variance in the majority of school systems; it stands to create confusion rather than add clarity by combining absolute performance, growth, and additional measures; and it risks violating the Type A rule, whereby schools will be rewarded or punished for things beyond their control, generating inefficiency and confusion as a result.

Supporters of the Texas legislation that created an A-F system have described it as an opportunity for parents to have more information and to make it so a low-rated school cannot hide behind a rating. However worthy a sentiment, the practical reality for those desiring clarity is they selected a tool that cannot provide it. What will be hidden is the underlying reality regarding schools. If history repeats itself, the most negative impact will again be on those serving our most disadvantaged students, and in turn, the students themselves.

Conclusion

"Simple is good, unless it is wrong."

Advocates present A-F school rating systems as a simple, accurate, easy-to-understand way to inform the public as to school quality. This is where a saying of a colleague is appropriate to add: Simple is good, unless it is wrong. And, in the case of A-F school rating systems, it is wrong on all counts. A-F systems are anything but simple, prone to inaccuracies, and impossible to understand. They represent one more way to punish poor schools for being poor and reward rich schools for being rich, rather than look at the unique needs of each to determine the degree to which of those needs are being met.

The issue is not one that can be remedied by a retooling of the process, or a rethinking of the rules. The reduction of everything in a school to a single grade, even if performed in a statistically robust manner with rigorous rules that do not change based on political whims, ignores the fact that most of what happens in that school is well outside that letter grade.

Reducing a school to a crude representation of average hides the fact that somewhere in that school exist students and issues and needs that cover the entire range of possibility. A school that gets an A, however, will be presumed to be doing a better job of meeting all of those needs, for all its students, just as a school that gets a D will be presumed to be doing a worse job of meeting all of those needs, for all its students. Yet the A school will have students it is failing, and the D school will have students who are achieving success. No amount of retooling can overcome that bluntness. Leaving it in place risks the grade badly misrepresenting what happens in schools, completely failing any test of a rational accountability system.

Notes

[1] Howe, K.R. & Murray, K. (2015). Why School Report Cards Merit a Failing Grade. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/why-school-report-cards-fail

[2] See http://www.excelinedinaction.org/policies/.

[3] For example, see technical description of the Utah system here: http://schools.utah.gov/assessment/Accountability/TechnicalManual.aspx, and a PowerPoint of the new Indiana system here: http://www.doe.in.gov/sites/default/files/accountability/accountability-presentationadvanced.pdf.

[4] One of the organizations deeply supportive of Florida's A-F program is the Foundation for Florida's Future, of which Governor Bush is a trustee. Their website declares: "our students have shown continuing achievement gains since the first year of school grading in 1999" (see http://www.afloridapromise.org/Pages/Florida_Formula/Facts_on_the_FCAT_and_Floridas_Path_to_Success/School_Grades_Q_and_A.aspx). Another foundation called Excel in Ed in Action lists the former governor as a board member and advocates repeatedly for the installation of A-F school grading policies (see http://www.excelinedinaction.org/policies/). In no place on either website are claims that A-F systems work substantiated, nor is any research on the matter included or referenced. Nevertheless, the Florida model remains

that primary model for such systems and continues to replicate itself based on the notion that it was effective at motivating change in schools.

[5] DiCarlo, M., (2013, March 5). Why Did Florida Schools' Grades Improve Dramatically Between 1999 and 2005? Washington, DC: Washington Post. Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2013/03/05/jeb-bushs-a-f-school-grading-scheme-the-facts/

[6] Ohio recently released their A-F grades and the negative reaction was swift, given this exact issue with rule changes. See http://www.cincinnati.com/story/news/2016/09/15/school-report-cards-can-we-trust-grades/90402648/ for just one of a rash of articles as one example of the confusion being experienced by schools.

[7] During the writing of this essay Florida released a new set of school grades against a "simplified" set of rules designed to make the system clearer and easier to understand. The results showed a significantly different picture than the year prior, with improvements and declines occurring against no clear pattern. In addition, the actual differences in test scores were negligible.

[8] Layton, L., (2013, August 3). A-to-F systems for grading public schools get new scrutiny. Washington DC: Washington Post. Retrieved from https://www.washingtonpost.com/local/education/a-to-f-grading-systems-for-public-schools-get-new-scrutiny/2013/08/03/03533aa2-fbab-11e2-a369-d1954abcb7e3_story.html.

[9] Dicarlo, M., (2012, January 25). A Dark Day For Educational Measurement In The Sunshine State. Web log content. Retrieved from http://www.shankerinstitute.org/blog/dark-day-educational-measurement-sunshine-state.

[10] The Oklahoma research is applicable to other systems that combine a static measure and a growth measure. While the nuances of each system do differ—sometimes a great deal—the bluntness of combining growth and static measures tends to result in a similar enough effect that there are lessons to be learned. What is definitely needed is a larger research agenda on the topic, one that has published as much and with as much care as the Oklahoma team.

[11] Raudenbush, S., (2004). Schooling, Statistics, And Poverty: Can We Measure School Improvement? Princeton, NJ: Educational Testing Service, Policy Evaluation and Research Center. Retrieved from: https://www.ets.org/Media/Research/pdf/PICANG9.pdf

[12] Adams, et.al., (2016)

[13] Adams, et.al., (2016)

[14] The three components the Oklahoma system and the Texas system share: absolute measures, growth measures, and additional consequences for low performing schools and districts.

[15] Howe, K.R. & Murray, K. (2015).

[16] Florida Department of Education, 2016. 2015-16 Guide to Calculating School and District Grades. Retrieved http://schoolgrades.fldoe.org/pdf/1516/SchoolGradesCalcGuide16.pdf. Note as well that Florida also uses "comprise" both correctly and incorrectly in one of the rules, further confusing the issue.

[17] Alabama State Department of Education, (2016). Alabama's A-F Report Cards: Update on ESSA Accountability. Retrieved from https://www.alsde.edu/sec/acct/Resources%20Tabbed/AASB%202016%20-%20%20A-F%20Report%20Card.pdf

[18] Alabama State Department of Education, (2016). A-F Report Card: Review of Data Input. Downloaded from https://ebssp.eboardsolutions.com/sites/aasb/Documents/Advocacy/A-F%20Report%20Card%20Review%20of%20Data%20Input.pdf

[19] Howe, K.R. & Murray, K. (2015). Why School Report Cards Merit a Failing Grade. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/why-school-report-cards-fail

[20] Adams, et.al., (2016).

[21] Adams, C.M., Forsyth, P.B., Ford, T.G., Ware, J.K., Barnes, L.B., Khojasteh, J., Mwavita, M., Olsen, J.J., & Lepine, J.A. (2015). Next Generation School Accountability. A Report Commissioned by the Oklahoma State Department of Education. Oklahoma Center for Education Policy (The University of Oklahoma) and The Center for Educational Research and Evaluation (Oklahoma State University).

[22] The lessons from Oklahoma apply to any system that intends to combine absolute measures and growth measures, and then will assign additional weight to schools that receive lower grades by preventing the achievement of a higher grade, very similar to how test scores will be treated in the Texas system.

[23] Adams, et.al., (2015), referencing an unattributed quote by Linda Darling Hammond.

[24] Adams, et.al., (2016)

[25] Howe, K.R. & Murray, K. (2015

[26] See, for example, http://www.cincinnati.com/story/news/2016/09/15/school-report-cards-can-we-trust-grades/90402648/, and https://nondoc.com/2016/10/29/ridiculous-school-grades/, as but two examples among

hundreds published each year in A-F states by those trying to understand what the grades actually mean. The near consensus position of articles discussing the topic suggests that doing so is generally futile.

[26] Texas House of Representatives (2015).

[27] Texas House of Representatives (2015).