Modeling Confidence Intervals on End-of-Course Assessments

Technical Report

Clint Sattler
Supervisor of Research and Evaluation
Knox County Schools
Department of Research, Evaluation, and Assessment

April 2020

Overview

All measurements are impacted by uncertainty, including results from the Tennessee Comprehensive Assessment Program (TCAP). Because student-level test results are aggregated into school-level accountability measures, accountability data should quantify uncertainty in some manner. The Tennessee Department of Education (TDOE) currently accounts for uncertainty using confidence intervals, which provide a range in which the "true" value of student outcomes likely falls.

Through the state accountability framework, schools (and subgroups within schools) are provided Annual Measurable Objectives (AMOs) by which the proportion of students who are classified as "On-Track" or "Mastered" in Mathematics or English/Language Arts (p) should increase in the current year. The 95% confidence interval ($CI_{95\%}$) for the current-year p may be used to determine a school or subgroup grade in the "Achievement" indicator on the state report card (Table 1).

Table 1: 95% Confidence Intervals and State Report Card Grades: Achievement

State Report Card Grade	Criteria
A	Current-year p ≥ prior-year p + 2x AMO
B	Current-year p ≥ prior-year p + AMO
C	Upper $CI_{95\%}$ of current-year p ≥ prior-year p + AMO
D	Upper $CI_{95\%}$ of current-year p > prior-year p
F	Upper $CI_{95\%}$ of current-year p ≤ prior-year p

This report will examine the properties of the 95% confidence intervals for student achievement generated by the current TDOE methodology and a competing methodology. The intent of this report is to identify a preferred method to internally identify student progress estimated by the TCAP. For the sake of brevity, this report will largely concern itself with the upper 95% confidence interval, as this is the confidence interval that directly impacts school accountability designations.

Readers should also be aware that, in some cases, the current-year p value alone is used to determine the report card grade. Since this alternate pathway does not include a confidence interval, it is not discussed further in this report. Readers who are interested in learning more about the state report card should consult the TDOE Accountability methodology for the 2018-2019 school year (SY1819).

TDOE Methodology

The TDOE 95% confidence interval is computed such that the true proportion of tested students who would be considered “On-Track” or “Mastered” is contained within the interval 95% of the time (if the test were administered an infinite number of times). The equation that is used to calculate the confidence interval is contained below.

Equation 1: TDOE 95% Confidence Interval Calculation

$$CI_{95\%} = \frac{n}{n + 1.96^2} * \left(p + \frac{1.96^2}{2 * n} \pm 1.96 * \sqrt{\frac{p * (1 - p)}{n} + \frac{1.96^2}{4 * n^2}} \right)$$

Where n is the number of valid tests administered at the school (or to a subgroup of students at the school), p is the proportion of tested students who earned a performance level of “On-Track” or “Mastered”, and 1.96 corresponds to the (mean-centered) number of standard deviations containing 95% of the area under the standard normal distribution.

The state methodology is based on the principle of sampling error. This concept suggests that the “true” proportion of students who have passed the state test must be estimated since outcome data is only available from a sample of these students. It is argued that inferential statistics can be used to extrapolate annual results to estimates of student performance of similar students at a particular school over all points of time. Under this scenario, data from a given year can be considered a sample of how a specific school (or subgroup within the school) has performed over time, even if the sample data was derived from 100% of the students currently enrolled at the school. This is the case in nearly every school and/or subgroup in the Knox County Schools (KCS). In SY1819 (the most recent year in which full test results are available) 577 Knox County schools and subgroups generated valid test data for accountability purposes. The distribution of the “percentage of students tested” for these 577 groups is summarized in Table 2. The data indicate that at least half of the schools/subgroups in the Knox County Schools tested 100% of students in SY1819.

Table 2: Quartiles of Percent Tested in SY1819 in Knox County

Count	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
577	93% Tested	99% Tested	100% Tested	100% Tested	100% Tested

Equation 1 indicates that the upper 95% confidence interval is a function of p and n. The relationship between p, n, and the difference between the upper 95% confidence interval and p is shown in Figure 1. Figure 1 indicates that the maximum upper confidence interval "boost" occurs for small schools/subgroups in which p is between 30% and 40% "On-Track" or "Mastered". Additionally, Figure 1 and Table 3 indicate that as the number of students in a school or subgroup approaches infinity, the difference between p and the upper and lower confidence interval approaches zero.

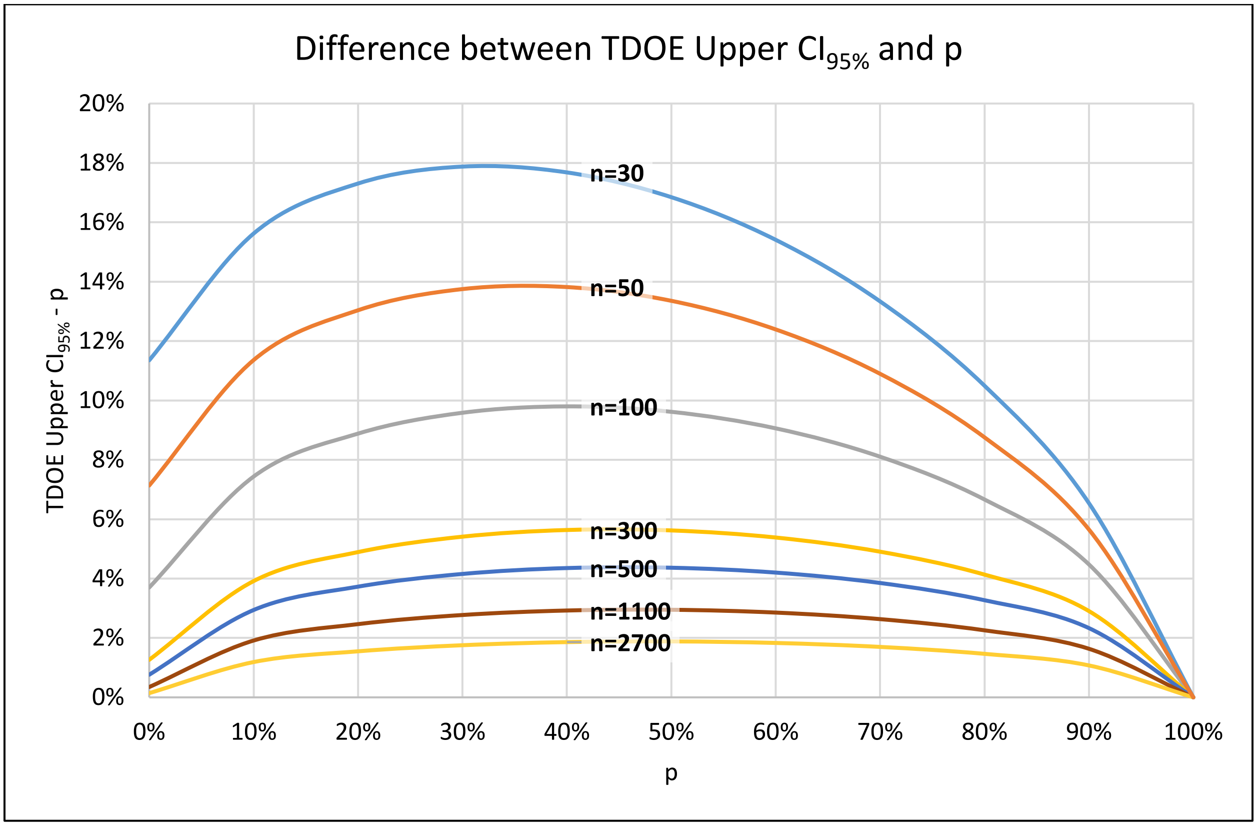


Figure 1: Difference between Upper 95% Confidence Interval and p by p and n

Table 3: Effect of n on TDOE 95% Confidence Intervals

	n (valid tests)	p (proportion "On-Track" or "Mastered")	TDOE $CI_{95\%}$ (Lower)	TDOE $CI_{95\%}$ (Upper)
School 1	10	50.0%	23.7%	76.3%
School 2	100	50.0%	40.4%	59.6%
School 3	1000	50.0%	46.9%	53.1%
School 4	10000	50.0%	49.0%	51.0%
…	…	…	…	…
School 8	100000000	50.0%	50.0%	50.0%

Simulated Sampling Distributions Methodology

The KCS Department of Research, Evaluation, and Assessment (REA) investigated an alternative method for determining the 95% confidence intervals for p. In this study, REA investigated a methodology in which individual student-level test scores were used to derive a school-level confidence interval. The methodology relies on sampling from a plausible distribution of student-level scaled scores so this approach will be referred to as the SSD (simulated sampling distributions) method.

Arguably, SSD should not be applied to TCAP data generated prior to the fall of SY1920. Prior to SY1920, uncertainty at the student level was reported as a standard error of measurement (SEM). SEM is a classical test theory measurement in which it is assumed that uncertainty exists because a test cannot contain every possible question about a given subject. However, TCAP technical documentation makes it evident that TCAP performance is measured using an item response theory (IRT) approach. In an IRT approach, uncertainty is a function of student response patterns and the type (and number) of test questions a student answers correctly. It can be argued that the SEM is not an appropriate measure of uncertainty for scaled scores generated by IRT. It is unknown as to why TDOE and/or the test vendor chose to report SEM.

The vendor used to administer the fall SY1920 TCAP assessment did not report SEM, but rather a measure of uncertainty called the conditional standard error of measurement (CSEM). The CSEM is better aligned to the IRT approach in that students with different scaled scores have different amounts of reported uncertainty. Example data from five KCS students tested in English I (E1) in the fall of SY1920 is contained in Table 4. Possible E1 scaled scores range from 200 to 450. Any student with a scaled score greater than or equal to 333 will "pass" the state test by earning a performance level of "On-Track" or "Mastered".

Table 4: SY1920 KCS English I Data Sample

Student	Content Area	Scaled Score	CSEM	Performance Level
Student 1	E1	371	12	Mastered
Student 2	E1	331	2	Approaching
Student 3	E1	334	2	On-Track
Student 4	E1	301	6	Below
Student 5	E1	333	2	On-Track

In the SSD approach, each student-level TCAP scaled score is treated as uncertain. It is assumed that the observed scaled score is the center of a normal distribution of plausible scores and that the standard deviation of the plausible score distribution is the CSEM. A computer program was written to pull a random score from each student's plausible distribution and compare it to the minimum scaled score required for a student to be classified as “On-Track”. If the “simulated” score is greater than or equal to the “On-Track” cut, a student is coded with a 1 for “proficiency”. Otherwise, they are coded with a 0. Once this procedure is complete for each student and content area, the results are aggregated to the school level and one plausible value for p (the proportion of students in the school who are “On-Track” or “Mastered”) is calculated. This process is repeated tens, hundreds, or thousands of times to generate plausible values for p at a school. The 95% confidence intervals are constructed from the 2.5% and 97.5% quantiles of the simulated p values for the schools.

Please note that although SSD uses random sampling from distributions, it is generally desirable to be able to exactly reproduce simulation results. Unless specifically mentioned, these investigations used a seeded randomized process so that the results would be replicable. All random sampling was done through R (version 3.6.1) running on RStudio (version 1.2.1335)

For illustrative purposes, SSD was applied to the students in Table 4, sampling 100 times (N_{sim} = 100) from each student's distribution to generate 100 plausible scaled scores. The distributions are shown as violin plots in Figure 2. Wider portions of a violin plot indicate that a score from this portion of the distribution is more likely to be observed than a score from a narrower portion of the distribution. The blue dots indicate the reported (observed) scaled score and the red line corresponds to the "On-Track" cut score.

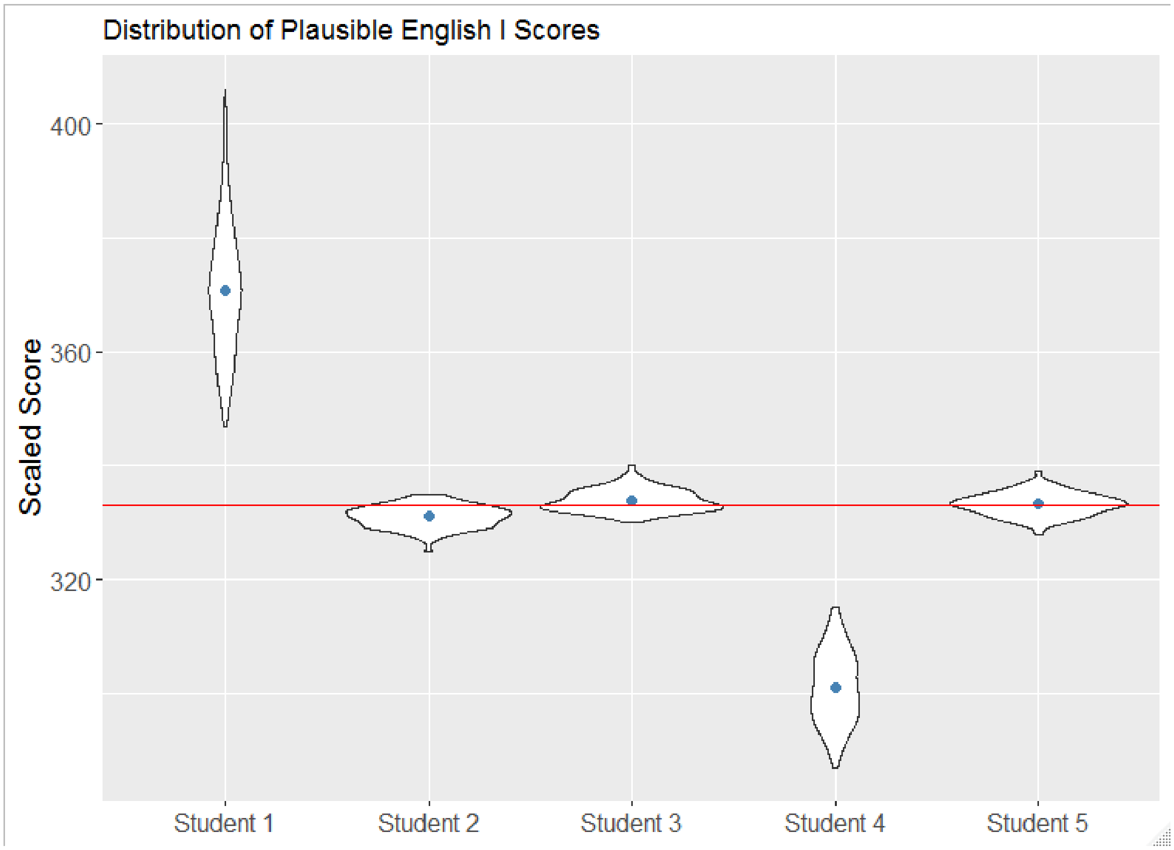


Figure 2: Distribution of Plausible English I Scaled Scores for Five Students

The likelihood that a student has a scaled score that exceeds the "On-Track" cut can be computed from the simulated distribution of scaled scores. The likelihood that a student would have an E1 scale score greater than or equal to 333 for the five example students is provided in Table 5 (based on 100 random samples per student). These simulated results consider Student 2 to be "On-Track" 24 times when sampled (i.e. tested) 100 times. Student 5 would be considered "On-Track" 65 times when sampled 100 times.

Table 5: SY1920 KCS English I Sample SSD Results; N_{sim} = 100

Student	Observed Scaled Score	Observed CSEM	Observed Performance Level	Likelihood of Passing (Scaled Score >= 333)
Student 1	371	12	Mastered	100%
Student 2	331	2	Approaching	24%
Student 3	334	2	On-Track	72%
Student 4	301	6	Below	0%
Student 5	333	2	On-Track	65%

The 95% confidence interval generated by the SSD methodology is a function of the number of simulated data points (N_{sim}). Confidence intervals were generated for school-level results from the fall of SY1920 using N_{sim} = 50, 100, 250, 500, 750, 1000, 1500, 2000, 2500, 3000, and 3500. A measure of stability, the percent deviance, was calculated for each N_{sim} as:

Equation 2: Percent Deviance from the Mode

$$\%\ Deviance = \frac{|Mode\ Upper\ CI_{95\%} -\ Calculated\ Upper\ CI_{95\%}|}{Mode\ Upper\ CI_{95\%}}$$

Where Mode Upper $CI_{95\%}$ is the mode upper 95% confidence interval from the eleven simulations using different N_{sim} and Calculated Upper CI95% is the upper 95% confidence interval generated for a specific N_{sim}.

The mode values for the upper 95% confidence interval are presented in Table 6. The percent deviance for each school is presented in Figure 3. The percent deviance rapidly approaches zero for the majority of schools as N_{sim} increases. Five-hundred appears to be the minimum number of data points required to reach a stable estimation of the upper 95% confidence interval at the majority of schools. In this study, an estimation is considered stable when the percent deviance is generally less than 1%. It should be noted that the upper 95% confidence interval at Austin-East High/Magnet continually fluctuates across all values of N_{sim} used in this investigation.

Table 6: SSD Mode Upper CI$_{95\%}$ from 11 Trials with varying N$_{sim}$

School	Mode Upper CI95%
Austin East High/Magnet	10.8%
Bearden High School	45.0%
Career Magnet Academy	50.0%
Carter High School	38.5%
Central High School	34.5%
Dr. Paul L. Kelley Volunteer …	5.6%
Farragut High School	66.1%
Fulton High School	15.8%
Gibbs High School	33.3%
Halls High School	44.9%
Hardin Valley Academy	55.2%
Karns High School	36.9%
Powell High School	38.4%
Richard Yoakley School	12.1%
South Doyle High School	32.3%

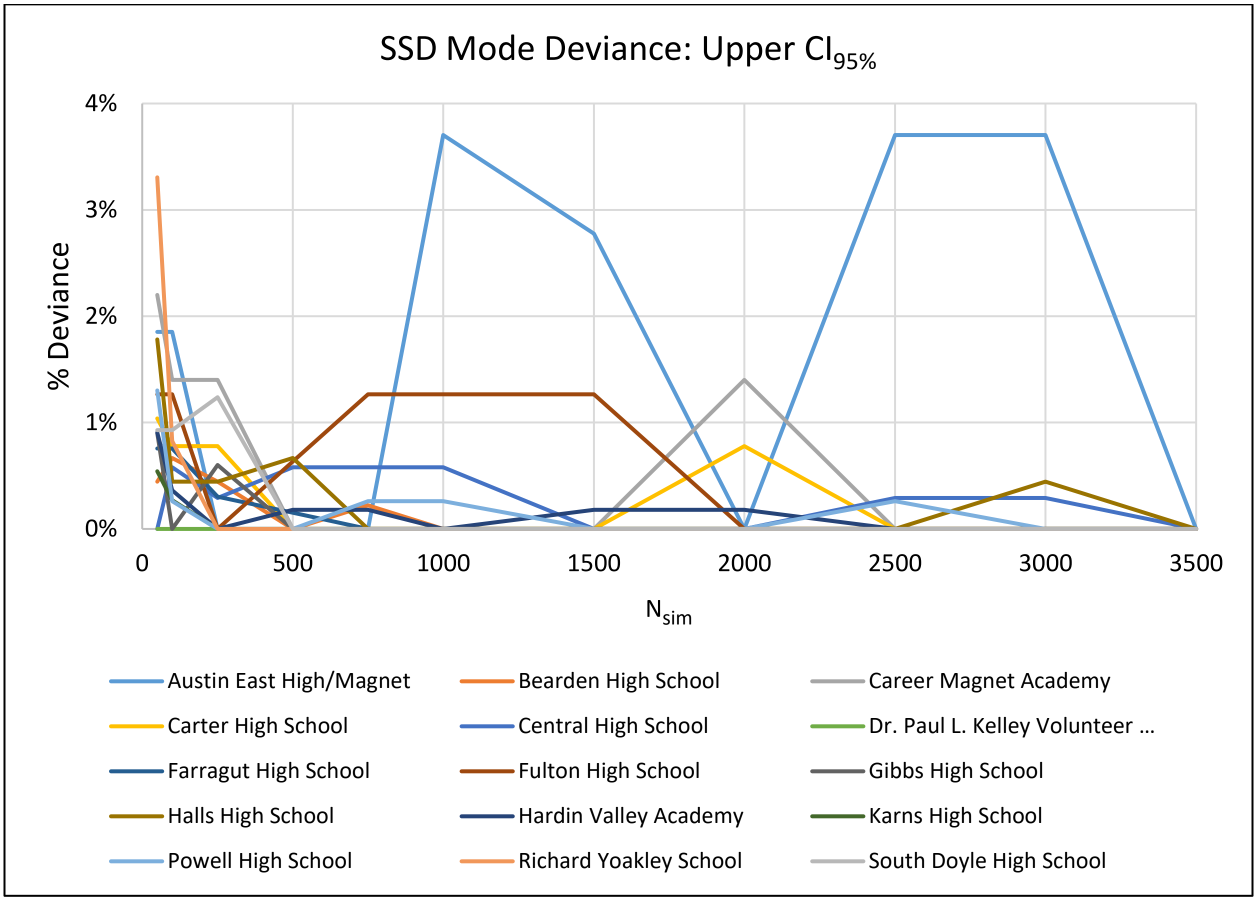


Figure 3: Dependence of SSD Upper CI$_{95\%}$ on N$_{sim}$

There are a variety of SY1819 metrics that differentiate Austin-East High/Magnet from the KCS average. Notably, Austin-East has much lower p and lower n than the KCS average. Also, the proportion of students with scaled scores near the "On-Track" cut score that are classified as "On-Track" is far lower at Austin East (~11%) than the district as a whole (~45%).

The sensitivity of the confidence intervals on the random number generator used in SSD was also investigated. Five fully random (unseeded) trials were simulated with N_{sim} = 3500. The results are presented in Table 7 and do raise some concerns about the reproducibility of the SSD confidence interval for some schools: notably Austin-East High/Magnet and Dr. Paul L. Kelley Volunteer Academy. The data in Table 7 suggest that it would be preferable to use a seeded randomization process if the SSD confidence intervals had to be replicated exactly. Like Austin-East High/Magnet, the Dr. Paul L. Kelly Volunteer Academy has a p value and n value much lower than the KCS average.

Table 7: Mode Deviance for Five Random Trials with N_{sim} = 3500

			SSD Upper $CI_{95\%}$ % Deviance, N_{sim}=3500				
School	TDOE Upper $CI_{95\%}$	Trial Mode Upper $CI_{95\%}$	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
Austin East ...	8.6%	10.8%	0.0%	2.8%	0.0%	3.7%	3.7%
Bearden ...	45.0%	45.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Career Magnet ...	51.0%	50.7%	1.4%	1.4%	0.0%	0.0%	0.0%
Carter ...	38.6%	38.5%	0.0%	0.8%	0.0%	0.0%	0.0%
Central ...	35.2%	34.6%	0.3%	0.0%	0.0%	0.0%	0.0%
Dr. Paul L. Kelley ...	9.6%	8.3%	32.5%	0.0%	32.5%	0.0%	0.0%
Farragut ...	66.0%	66.1%	0.0%	0.0%	0.0%	0.0%	0.0%
Fulton ...	14.3%	15.8%	0.0%	0.0%	0.0%	0.0%	0.0%
Gibbs ...	31.9%	33.5%	0.6%	0.0%	0.0%	0.0%	0.3%
Halls ...	45.2%	44.9%	0.0%	0.4%	0.7%	0.4%	0.0%
Hardin Valley	55.4%	55.2%	0.0%	0.0%	0.0%	0.0%	0.2%
Karns ...	34.8%	36.9%	0.0%	0.0%	0.0%	0.0%	0.3%
Powell ...	38.2%	38.5%	0.3%	0.5%	0.0%	0.3%	0.0%
Richard Yoakley ...	18.6%	12.1%	0.0%	0.0%	0.0%	0.0%	0.0%
South Doyle ...	32.1%	32.3%	0.0%	0.0%	0.0%	0.0%	0.3%

REA created a design of experiments (DOE) in an attempt to better understand the factors leading to instability in the SSD confidence interval. The DOE used four variables with two levels per variable (4x2 DOE). The low and high values for the variables were chosen from the ranges observed in the KCS fall SY1920 TCAP data. The variables investigated included n (n_{low} = 300, n_{high}=900), p (p_{low} = 5.3%, p_{high} = 53.3%), the percentage of students who would be considered "cusp" students (% $Cusp_{low}$ = 5%, % $Cusp_{high}$ = 10%), and the percentage of students classified as "cusp" who earned a scaled score greater than or equal to the "On-Track" cut (%Prof. $Cusp_{low}$ =13.3%, % Prof. $Cusp_{high}$ = 46.7%). "Pseudo" schools were generated using stratified sampling from the fall SY1920 TCAP data. N_{sim} was set to 3500 and five trials were executed, each using a different randomization scheme for sampling from the student-level distributions. The results of the trials are contained in Table 8. The "Mode Upper $CI_{95\%}$" column provides the mode upper 95% confidence interval from the five trials, the "% Trials Mode" indicates the percentage of the five trials that produced an upper 95% confidence interval equal to the mode, and the "Max % Deviance" column provides the maximum percent deviance (Equation 2) of the five trials.

Table 8: Results from Five Random Trials for 4x2 Design of Experiments

School	n	p	% Cusp	% Prof. Cusp	Mode Upper $CI_{95\%}$	% Trials Mode	Max % Deviance
School A	300	5.3%	5%	13.3%	13.0%	80%	2.3%
School B	300	5.3%	5%	46.7%	11.0%	60%	2.7%
School C	900	5.3%	5%	13.3%	10.4%	80%	1.0%
School D	900	5.3%	5%	46.7%	9.3%	60%	1.1%
School E	300	53.3%	5%	13.3%	57.7%	60%	0.5%
School F	300	53.3%	5%	46.7%	57.7%	100%	0.0%
School G	900	53.3%	5%	13.3%	56.6%	40%	0.5%
School H	900	53.3%	5%	46.7%	55.8%	100%	0.0%
School I	300	5.3%	10%	13.3%	14.0%	100%	0.0%
School J	300	5.3%	10%	46.7%	11.7%	100%	0.0%
School K	900	5.3%	10%	13.3%	11.5%	40%	1.7%
School L	900	5.3%	10%	46.7%	9.6%	100%	0.0%
School M	300	53.3%	10%	13.3%	59.7%	60%	0.7%
School N	300	53.3%	10%	46.7%	58.0%	100%	0.0%
School O	900	53.3%	10%	13.3%	57.9%	80%	0.2%
School P	900	53.3%	10%	46.7%	55.9%	100%	0.0%

The results to not provide definitive evidence regarding the mechanism that leads to stability issues when determining the 95% confidence limits using SSD. All of the schools with maximum percent deviance greater than or equal to 1 are schools with low p. However, Schools I and J also have low p but did not exhibit stability issues in the five trials. The schools in which the mode was not produced each time (i.e. "% Trials Mode" ≠ 100%) were largely the schools in which the percent of “cusp” students who were classified as “On-Track” were low. However, there are exceptions (Schools B, D, and I).

The author suspects that the stability issue is related to low p and the presence of cusp students. Low p values were observed at two schools in which stability issues were identified (Austin-East High/Magnet and Dr. Paul. L. Kelley Volunteer Academy). Theoretically, the SSD methodology should produce varying confidence intervals in the presence of students with plausible score distributions that span the “On-Track” cut score. However, the overall impact of “cusp” students may be negligible if 50% of these students are classified as “On-Track”, and 50% are classified as “Approaching”. An imbalance in the distribution of “cusp” students in the performance categories that span the “On-Track” cut score could yield unstable confidence limits.

The inability to determine the exact cause of instability using the simulations in Table 8 may stem from the definition being used to classify a student as “cusp”. Recall that in the current study a student is categorized as a “cusp” or “bubble” student if their scaled score estimate plus or minus their CSEM spans the “On-Track” cut in a given subject (see Students 2, 3, and 5 in Figure 2). This is a somewhat arbitrary definition. The application of this arbitrary classification may not allow the construction of a pseudo school that is as sensitive to the presence of “cusp” students as actual KCS schools. Further study is required to determine the exact cause of the instability in the upper 95% confidence interval.

Comparisons of Methodologies

The upper 95% confidence intervals generated by the two methodologies from the fall SY1920 TCAP data were compared in a series of trials. The SSD methodology used $N_{sim} = 3500$ to minimize the observed deviance at all schools. A seeded random process was used for the SSD methodology so that the results could be replicated.

The comparison between the school-level SSD and TDOE 95% confidence intervals is contained in Table 9. The upper confidence interval, which is used to derive the report card grade, was larger using the SSD method in 50% of the schools. The mean absolute difference in the upper 95% confidence interval was 1.4 percentage points.

Table 9: Comparison of TDOE Method and SSD with N_{sim} = 3500

			SSD 95% CI		TDOE 95% CI	
School	n	p	Lower	Upper	Lower	Upper
Austin East High/Magnet	297	5.4%	4.7%	10.8%	3.4%	8.6%
Bearden High School	1046	42.0%	39.5%	45.0%	39.0%	45.0%
Career Magnet Academy	134	42.5%	37.3%	50.0%	34.5%	51.0%
Carter High School	369	33.6%	29.3%	38.5%	29.0%	38.6%
Central High School	743	31.8%	28.0%	34.5%	28.6%	35.2%
Dr. Paul L. Kelley Volunteer …	36	0.0%	0.0%	5.6%	0.0%	9.6%
Farragut High School	1061	63.1%	60.3%	66.1%	60.2%	66.0%
Fulton High School	467	11.1%	9.0%	15.8%	8.6%	14.3%
Gibbs High School	481	27.7%	24.7%	33.3%	23.9%	31.9%
Halls High School	523	40.9%	35.9%	44.9%	36.8%	45.2%
Hardin Valley Academy	1010	52.3%	48.9%	55.2%	49.2%	55.4%
Karns High School	721	31.3%	29.8%	36.9%	28.0%	34.8%
Powell High School	683	34.6%	31.0%	38.4%	31.1%	38.2%
Richard Yoakley School	58	8.6%	3.4%	12.1%	3.7%	18.6%
South Doyle High School	564	28.2%	24.6%	32.3%	24.6%	32.1%

A comparison of the differences between the upper 95% confidence interval and the observed p value (by school) is contained in Figure 4. The TDOE methodology results in larger differences at small n. As n increases, the difference decreases in both methodologies. The difference between p and the upper 95% confidence interval should approach zero as n approaches infinity using the TDOE methodology. This is not necessarily true for the SSD methodology. The difference between p and the 95% confidence interval will approach zero using the SSD approach in any school or subgroup where performance is sharply divided (i.e. the school or subgroup consist only of students with 0% and/or 100% likelihood of meeting or exceeding the "On-Track" cut score).

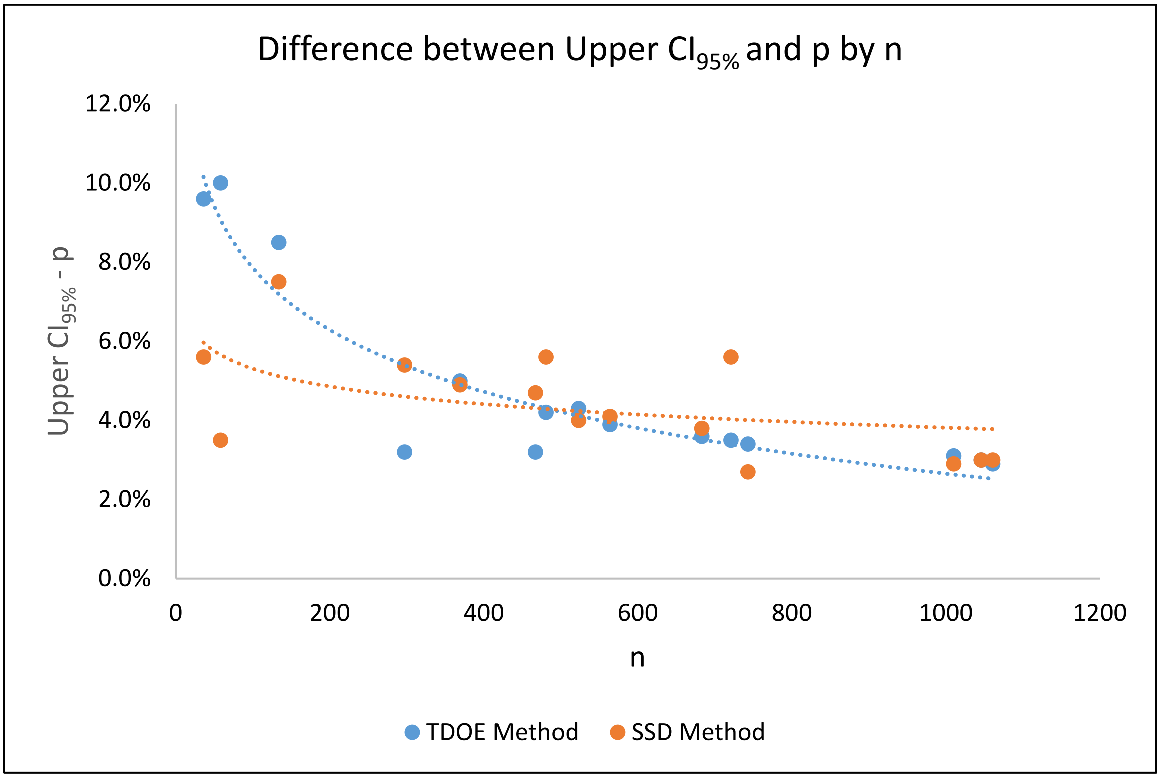


Figure 4: Difference between Upper $CI_{95\%}$ and p by n

Theoretically, the SSD methodology should have a stronger dependency on the percentage of “cusp” students in a school or subgroup. The difference between the upper 95% confidence interval and p by the percentage of “cusp” students is provided in Figure 5. It is evident that as the proportion of cusp students in a school increases, the methodology that produces the largest difference between p and the upper 95% confidence switches from the TDOE methodology to the SSD methodology. It is somewhat surprising to see that the slope of the line through the SSD is slightly negative. This finding is counter to SSD theory, but may be due to a relatively small number of schools with a low percentage of “cusp” students, or interactions between n, p, and the proportion of “cusp” students.

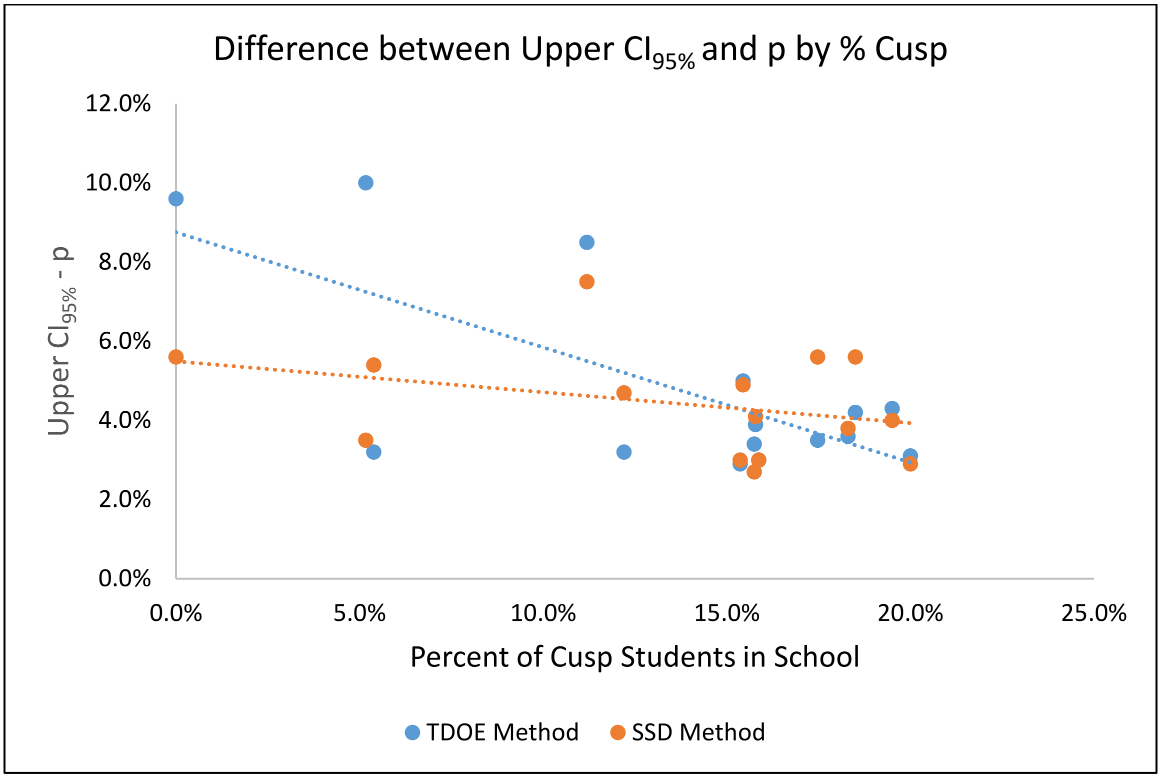


Figure 5: Difference between Upper $CI_{95\%}$ and p by % Cusp

For completeness, Figure 6 displays the relationship between the difference between p and the upper 95% confidence interval for schools with varying p. The relationship appears to be largely independent of the methodology used. Any trend visible in the data is likely related to the fact that schools with low p tend to be smaller KCS schools.

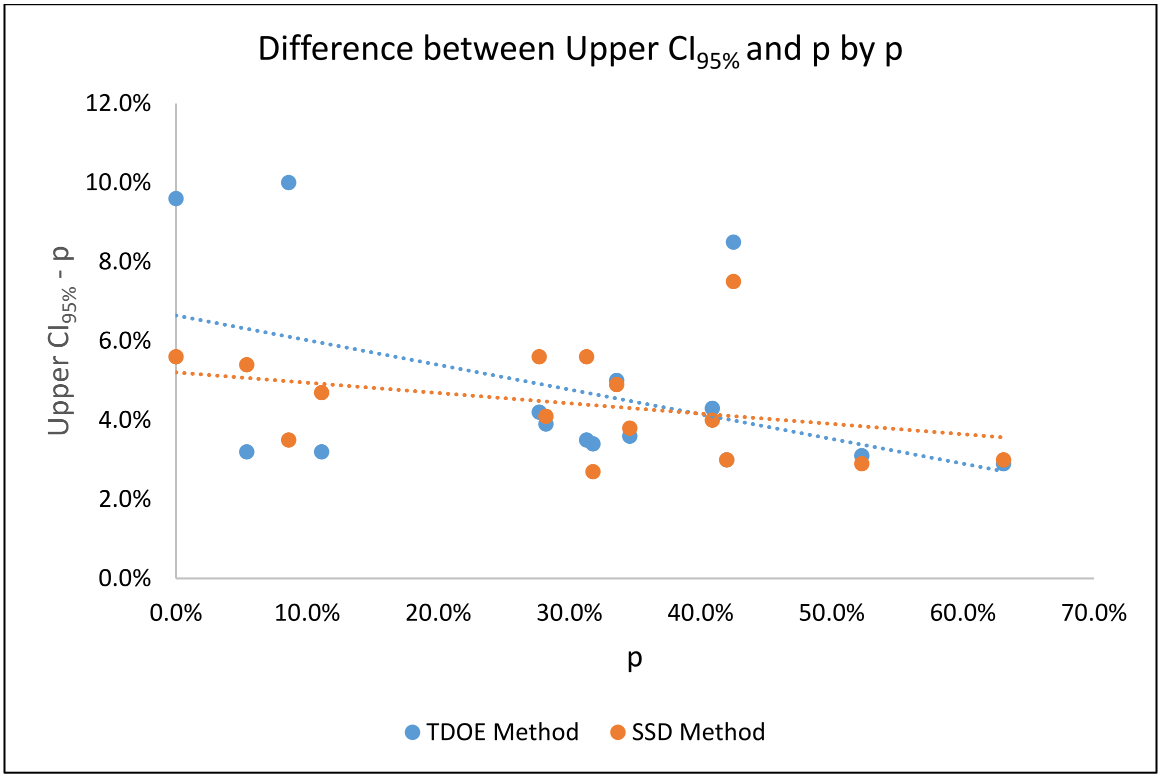
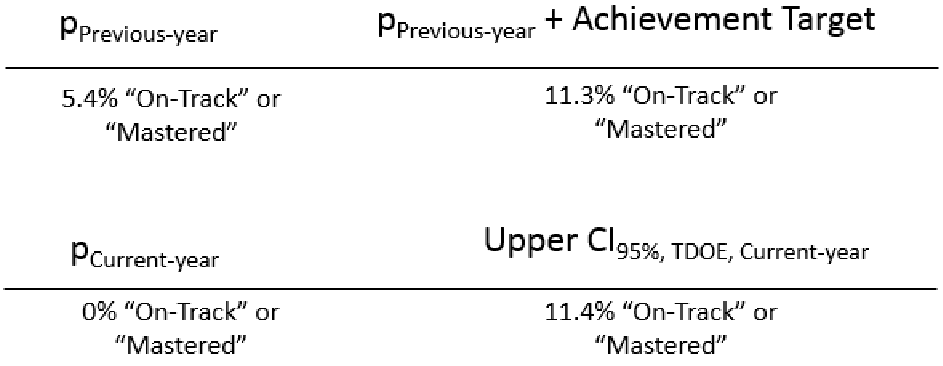


Figure 6: Difference between Upper $CI_{95\%}$ and p by n

Conclusions & Considerations

Both methods used to determine a confidence interval for the proportion of students who are “On-Track” or “Mastered” have strengths and weaknesses. The TDOE method assumes that data coming from a small group of students is a poor approximation of the population statistics since a small group is less likely to reflect the universe of results. This leads to large confidence intervals that can inflate report card grades under certain circumstances (especially schools or subgroups with a very low proportion of “On-Track” or “Mastered” students). Figure 7 illustrates this for a hypothetical school or subgroup with n = 30. Even though the measured percentage of students who were “On-Track” or “Mastered” has decreased to 0% in the current year, the TDOE methodology would produce an upper confidence interval of 11.4% “On-Track” or “Mastered”, allowing the school or subgroup to score a “C” on their report card.



$p_{\text{Previous-year}}$	$p_{\text{Previous-year}}$ + Achievement Target
5.4% “On-Track” or “Mastered”	11.3% “On-Track” or “Mastered”

$p_{\text{Current-year}}$	Upper $\text{CI}_{95\%,\text{ TDOE, Current-year}}$
0% “On-Track” or “Mastered”	11.4% “On-Track” or “Mastered”

Figure 7: Hypothetical School or Subgroup with n=30

The TDOE methodology only considers sampling error when estimating uncertainty. Uncertainty in the individual student-level measurements are assumed to be negligible in comparison to the sampling error (regardless of n or p). The TDOE methodology is, therefore, insensitive to students who meet or fail to meet the “On-Track” cut score by very thin margins. For example, the two hypothetical schools contained in Figure 8 would have the same confidence interval for the proportion of students who were “On-Track” or “Mastered” because n and p are equal in both schools.

School A	School B
25 Students "Mastered" 50 Students "On-Track"	25 Students "Mastered" 50 Students "On-Track"
150 Students "Approaching" *(but 1 less correct question and all 150 would be "Below")*	150 Students "Approaching" *(but 1 more question correct and all 150 would be "On-Track")*
100 Students "Below"	100 Students "Below"

Figure 8: Results from Two Hypothetical Schools

Additionally, analysis utilizing data from the TDOE methodology will likely increase the rate of Type II errors (concluding no discernable difference between groups when a difference exits) in small schools/subgroups (Figure 1). The TDOE methodology will also likely increase the rate of Type I errors (concluding a difference between groups when none exists) in very large schools/subgroups (Table 3).

Despite these issues, the TDOE methodology is computationally simple and replicable assuming n and p are provided (as they are in state accountability files). The calculation methodology is straight-forward and can be easily explained to large cross-sections of stakeholders. The sampling error approach may support the use of statistical inference regarding time-invariant performance at a school. This may make this type of measurement better suited to the long-term monitoring of student performance data through the state accountability system, despite potential issues with non-independence in year-over-year sampling.

Conversely, the SSD method utilizes the inherent uncertainty in the student-level data to create school or subgroup aggregate levels of uncertainty. This method is theoretically more accurate for estimating performance of the current cohort of students because of how students who are performing near the "On-Track" cut impact the confidence interval. The schools in Figure 8 would have different confidence intervals when the SSD method is deployed.

However, the SSD method is considerably more computationally expensive than the TDOE methodology, may be difficult for some district-based staff to replicate (though not impossible if seeded randomization is utilized), may be difficult to explain to stakeholders, and results may be unstable for small schools with a low proportion of students classified as "On-Track" or "Mastered" ($\lesssim$10%) and/or schools with large populations of students near the "On-Track" cut score. Additionally, the SSD method ignores error associated with sampling from the universe of student outcomes. Therefore, the SSD methodology provides information only about the tested cohort of students and does not support inferences beyond that cohort. Analysis using data generated by the SSD method may be more prone to Type I or Type II errors when a current-year cohort is radically different than prior-year cohorts.

The author recommends that the SSD methodology be deployed within the district to inform tactical decision making within the district. The author feels that the SSD method provides a more accurate estimate of current-year performance because it relies on fine grained information from scaled scores data rather than coarse categorization of student performance. Even though confidence intervals produced for small schools/subgroups with a low proportion of "On-Track" or "Mastered" students may be unstable, they are less likely to over-estimate performance to the same extent as the TDOE methodology. Although the SSD methodology does not lend itself to inferences about the longitudinal population of the school, school-leaders tend to use the state test data to make tactical programmatic and staffing decisions.

There is a concern that the TDOE confidence interval would be considered “more correct” than an unstable SSD confidence interval simply because the TDOE interval is reproducible. However, no matter which method is used, readers are cautioned when interpreting confidence intervals for small schools, schools/subgroups with low proportions of “On-Track” or “Mastered” students, and schools with a large proportion of students near the “On-Track” cut. The “true” measures for these schools/subgroups are difficult to estimate regardless of the methodology used to derive the confidence intervals. Progress or regression in these schools will be difficult to detect using measures related to the proportion of students who are “On-Track” or “Mastered”. Therefore, it may be prudent for the district to develop more sensitive indicators to monitor student performance at these schools.