Thank you Chair Cupp, Ranking Member Fedor, members of the Joint Education Oversight Committee and Director Jones for the opportunity to testify on the Ohio state report card Value-Added metric.

My name is Beth Osyk. I'm a parent in the Twinsburg school district and a software engineer. I see a report card system with tremendous potential, which we begin to unlock by aligning the metrics with the intent behind accountability.

Accountability is an important concern that is somewhat different from the requirements for teacher evaluation. The value-added metric used by Ohio today, EVAAS, was designed to measure the relative contribution of teachers to growth. The mismatch to accountability needs causes gaps in accountability, especially for at-risk students.

Accountability legislation in Ohio encompasses two key items:

- Measuring a district's ability to educate students, with special concern for "failing" districts
- Mandatory high school graduation requirements, by passing a test or other achievement

A growth metric should measure how much a student has learned in a year and check that this is an acceptable for eventually reaching graduation, given external factors such as poverty.

The central idea behind a value-added metric is to compare a student's performance with a target. The key choices are how to set targets, how to assign letter grades, and how to design the test itself. Ohio's current metric has the following concerns:

- The actual amount of growth is not reported. (It is not needed to be known for relative teacher evaluations. This is a problem for accountability, however.)

- There are two different methods to calculate the value-added scores, depending on subject and grade level. Neither checks that adequate material has been learned by all students.

- The URM method works by predicting a test score for a student based on the student's own past test scores. If the new score matches the prediction, in the same way that the new average state score matches the previous year's, this is deemed acceptable (a 'gain' of zero). For a student with low prior-year scores, the target is the same low score.

- The MRM method works by taking a student's test score, computing the student's relative position to others, and checking if the student has the same relative position as last year. For the student who comes in "last place", the target is for him to stay in last place.

- Because these two methods are relative measures (URM with respect to state average, MRM with respect to relative position), approximately half of the students will over-perform or equal, and half will under-perform or equal. This leads to district grades where roughly half are A to C, and half are C to F.

- The gain values are then divided by the standard error to produce an Index that reports "confidence" in being over, at or under target. Standard error is proportional to the number of students. The Index value will generally be skewed high or low for large districts.

- Finally, cut scores are defined to map Index values to letter grades. The current cuts are at +2, +1, -1 and -2 for the Index. This results in most districts receiving either an "A" or an "F".

- As tests are given at grade level, there is a lack of information for students performing below grade level No metric can compensate for missing input.

These are the underlying reasons behind the confusion and frustration with the value-added metric. Essentially, the equations are not aligned with the intent behind accountability. Picking different choices for these items will produce a more understandable and more actionable report card.

The next steps are to get a handle on what the target values currently are, define what is intended to be measured, and select equations that match the intent.

- Ask SAS for more information on student score targets. Check on targets for low-performing students. SAS has made the average score target available in the URM school report [SAS15, page 4].

- Clarify what amount of growth is deemed acceptable, and what is “failing”. This requires linking metric results to material learned. Clarify the factors acceptable for adjusting a target (for example, poverty).

- Work out a report card design that meets the intent. A “try before you buy” approach could create an updated value-added analysis for a sample of three or four districts.

The report card has tremendous potential, as it has high visibility (especially among teachers and administrators), but people struggle to make sense of it. It is important to have an accurate and fair school accountability system as many voters use report card grades when deciding to approve or deny school tax levies.

More details are included in sections below, for reference.

State Representative Duffey included some excellent language on value-added metric transparency in proposed House Bill 591, which is copied below.

Thank you for your time today. I truly appreciate the opportunity to present this information.

Sincerely,
Elizabeth Osyk
Ph.D., Electrical and Computer Engineering

1) Measuring Progress in Addition to Achievement

The decision to use a progress metric is a choice itself. One does not necessarily have to include a progress metric.

The motivation behind using a progress metric is that students enrolling in the district may start out at different achievement levels. With a progress metric, a district can demonstrate high quality instruction by showing a high level of student growth despite a low starting achievement level.

Users of the report card are looking for the total amount of growth in terms of material learned in a school year. Since this information is absent, it is very difficult to relate the report card grades to curriculum gaps.

2) Introducing Targets with Value-Added Methods

Value-added methods report if some item is at, over or under its target. This item may be for a student, teacher or district. Many variations are possible regarding the items measured and how the targets are set.

The public has certain expectations when a progress metric is used for accountability assessment on a report card. **People intuitively look for a progress "goalpost" in terms of amount of material learned in a year,** where favorable letter grades are given to districts meeting or exceeding this goalpost. It is a surprise to be graded relative to others, where the goalpost is unknown and may change every year.

Ohio's value-added system uses two different models depending on subject and grade level. Both models set student item targets based on a student's past individual performance with respect to other students in the state. These are types of *norm-referenced* models, which implicitly adjust the student targets for external factors without requiring specific factors to be specified, or even known.

One downside of a pure norm-referenced approach is that the targets are set unconditionally, in the sense that no "reason" is required. This is a poor fit from the accountability perspective, as students with low past performance are given low future targets, without any consideration of the factors responsible for the low performance.

It's possible to define a model in a different way, called a *criterion-referenced* model, where the amount of growth is reported and a specific goal is set for the amount of growth, where the goal may be adjusted for external factors such as poverty [Martineau16].

3) Using Past Individual History for a Predicted Score Target (URM)

Looking in more detail at each method provides greater insight into why the value-added metric is not producing the calculation that people expect to see.

One model is the Univariate Response Model (URM). According to the SAS documentation, the URM growth model is used “when a test is given in non-consecutive grades, such as OST science assessments in grades 5 and 8 or any End-of-Course tests.” [SAS19].

In the URM, “a predicted score can be calculated for each individual student **based on his or her own prior testing history**” [SAS19] (bold added). Then, the URM “measures the difference between students’ predicted scores for a particular subject/year with their observed scores”.

This predicted score can be quite low, and can even be zero, when the student has had low scores in the past. Please see Appendix A for an example calculation. This is due to the URM model being originally designed for teacher evaluations. For teacher evaluations, it is the intention to predict a low score for students that have previously scored low - otherwise the teacher would garner an undeserved low teacher effect coefficient.

This is problematic for accountability since the model may consider scoring low (even zero) on the test for two consecutive years as acceptable.

The predicted score does not tell if a student surpassed a certain amount of growth. The actual amount of growth is not reported.

4) Using Relative Test Score Position for a Predicted Position Target (MRM)

For other grade levels and subjects, a second model called the Multi-Variate Response Model (MRM) is used. From the SAS documentation, “The Multivariate Response Model (MRM) is used for tests given in consecutive grades, like OST math and reading assessments in grades 3–8” [SAS19].

The target for a student is to maintain the same **relative position** score-wise with respect to other students. The relative position is computed with a Normal Curve Equivalent [SAS19], which is similar to a percentile with the added property of being an equal-interval scale, meaning that (for example) 10 to 11 is the same distance in some units as 90 to 91.

The growth measure is computed as the change in a student’s **relative position** [SAS19]:

- “MRM simplified example: If students’ achievement was at the 50th NCE in 2014 grade 4 math, based on the 2014 grade 4 math scale score distribution, and their achievement is at the 50th NCE in 2015 grade 5 math, based on the 2015 grade 5 math scale score distribution, then their estimated gain is 0.0 NCEs.”
- “The growth measure for these students is year 2 NCE – year 1 NCE, which would be 50 – 50 = 0. “

This is again problematic for accountability since the model considers it acceptable for a student to **remain at the same position** regardless of what the position is. The student coming in last place in the prior year is expected to remain last this year.

The relative position does not tell if a student surpassed a certain amount of growth. The actual amount of growth is not reported.

5) Half-and-Half Letter Grade Distributions Due to Relative Growth Expectations

“Growth expectations” are defined in a relative way by both models.

For URM the “growth expectation” is [SAS19]:

- “URM definition: Students with a district, school, or teacher made the same amount of progress as students with the average district, school, or teacher in the state for that same year/subject/grade.“

For MRM the “growth expectation” is [SAS19]:

- “MRM definition: Students maintained the same relative position with respect to the statewide student achievement that year.”

Since the “growth expectation” is computed in a relative way, **roughly half of the students will be at or above expectation, and roughly half will be at or below.**

This is a surprise for report card users (especially within the school district) who are looking for a growth goalpost in terms of the amount of material learned in a year. Currently, the value-added metric is a zero-sum game where students moving up mean that other students move down. (Or vice-versa, where students moving down result in others moving up, giving an false impression of improvement where there is none).

This is cited as a feature in the SAS documentation [SAS19]:
“Key feature: The value-added measures tend to be centered on the growth expectation every year with approximately half of the district/school/teacher estimates above zero and approximately half of the district/school/teacher estimates below zero.”

This is reflected in the district scores shown in Figure 1 from Thomas B. Fordham Institute, where roughly half of the districts receive a C or above, and half receive a C or below [Churchill17].

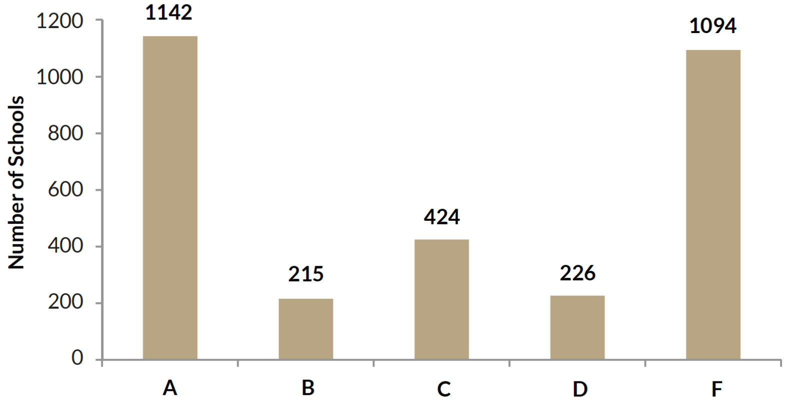

Figure 1: Distribution of A-F Grades on Overall Value Added, Ohio Schools, 2016-17 [Churchill17]

6) Grading Certainty as Part of the Letter Grade Skews Grades for Large Districts

Next, the report card computes an Index, which is used to determine the letter grade. The Index is computed as the "Growth Measure" field divided by the "Standard Error" field, as shown in the example in Table 1.

The result is the amount of "certainty" in the computed values, and not any of the values themselves. For example, the Red rating on the report card website is "**Significant evidence** that the district's students made less progress than the Growth Standard" (bold added).

Subject	Grade	**Number of Students**	Growth Measure	**Standard Error**	**Index**
Mathematics	7	**4866 Large**	-1.5574	**0.1515 Small**	**-10.27 Very High or Very Low**

Table 1: A Large Number of Students Can Lead to a Small Standard Error, Inflating the Index Grade 7 Mathematics Value-Added Score, Olentangy Local Schools (046763), 2018

This generally **skews the letter grades (either very high or very low) for large districts**, since the standard error is generally measured by taking the standard deviation divided by the number of students, then taking the square root. Larger districts have lower standard error, all else being equal. This makes the Index field quite different from the Growth Measure field for this example, driving the Index value into the "F" range.

7) Cut Score Selection for Letter Grade Assignment is Currently Tight

Finally, cut scores are used to find a letter grade for a particular Index. The current cuts are at +2, +1, -1 and -2 for the Index, as shown in Figure 2. **This results in most districts receiving either an "A" or an "F".**

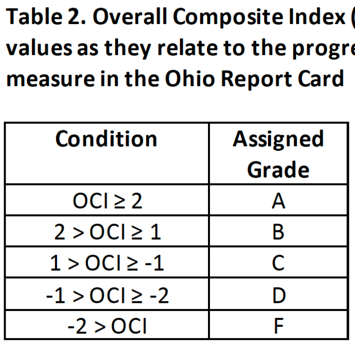

Table 2. Overall Composite Index (OCI) values as they relate to the progress measure in the Ohio Report Card

Condition	Assigned Grade
OCI ≥ 2	A
2 > OCI ≥ 1	B
1 > OCI ≥ -1	C
-1 > OCI ≥ -2	D
-2 > OCI	F

Figure 2. ORC 3302.03(A)(1)(e) cuts applied to the distribution of District OCI Cuts shown in Red

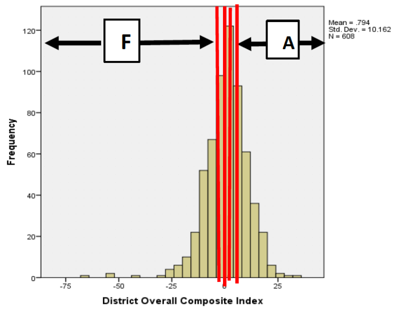

Figure 2: Cut Score Criteria and Grade Letter Distribution, JEOC Brief [JEOC18]

8) Testing for Growth of Below-Grade-Level Performers

Students are currently given tests at grade level, for a few reasons, including that the achievement metric is measured on an at-grade-level basis and due to perceived federal requirements.

However, a very low at-grade-level test score cannot really be said to reflect any learning. **An at-grade-level test essentially returns no information for a below-grade-level performer.** Fundamentally, no metric can work properly if the input is missing.

Measuring growth requires knowing the level of test questions that a student can successfully answer. Approaches like the Measures of Academic Progress test use progressive test questions and a cumulative score system, starting at zero when a student enters school and climbing up to a final value throughout successive grade levels [NWEA]. The progressive test questions are tailored based on the student's current material mastery, and the cumulative score system allows growth to be expressed despite a student performing below-grade-level.

The text of federal requirements should be carefully checked to see what test design can satisfy both the federal requirements and the mathematical need to somehow offer below-grade-level test questions to below-grade-level performers in order to measure the amount of material they have learned.

A careful design would meet any federal requirements while still gathering the necessary information for calculating a student's amount of growth.

9) Prior Legislative Work

State Representative Duffey included some excellent language on metric transparency in the proposed HB 591 Section 3302.03.B.3 [Duffey]:

"(3) Student growth. This measure shall do all of the following:
(a) Convey the amount of progress a student has made over the school year toward either having the knowledge necessary to perform proficiently in the next grade level or toward being college or career ready after graduation;
(b) Consist of a methodology that allows the measure's results to be validated and replicated by school districts. The department shall provide a district with the data necessary to validate or replicate the measure's results upon the district's request;
(c) Not consist of or contain a proprietary formula or method for measuring student growth. The department may contract with another entity to perform service work related to the measure.
(d) Include an explanation of the factors that influence student growth beyond the classroom, including parental and community influence and student attitude."

References

[Churchill17] A. Churchill, Back to the Basics: A plan to simplify and balance Ohio's school report cards, Thomas B. Fordham Institute, December 2017, https://fordhaminstitute.org/ohio/research/back-basics-plan-simplify-and-balance-ohios-school-report-cards

[Duffey] State Representative M. Duffey, Ohio House Bill 591, Revise report card rating system for schools,
https://www.legislature.ohio.gov/legislation/legislation-summary?id=GA132-HB-591

[JEOC18] JEOC Brief - April 19, 2018 - Value-Added,
http://www.jeoc.ohio.gov/Assets/Files/62.pdf.

[Martineau16] J. Martineau, A Guide to Understanding and Selecting Measures of Growth for Smarter Balanced Members, Smarter Balanced Assessment Consortium, May 2016, https://www.nciea.org/library/guide-understanding-and-selecting-measures-growth-smarter-balanced-members

[NWEA] NWEA, Student Profile Report, Measures of Academic Progress Help Center, Retrieved April 29 2019, https://teach.mapnwea.org/impl/maphelp/Content/Data/SampleReports/StudentProfile.htm

[Raudenbush12] S. Raudenbush and M. Jean, How Should Educators Interpret Value-Added Scores?, Oct. 2012, http://www.carnegieknowledgenetwork.org/briefs/value-added/interpreting-value-added/#footnote-3

[Renaissance] Renaissance EdWords, Lexile Measure, Retrieved April 29 2019, https://www.renaissance.com/edwords/lexile-measure/

[SAS15] SAS EVAAS, URM Modeling Approach for Value-Added, March 2015, http://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Ohio-Report-Cards/Value-Added-Technical-Reports-1/URM-Modeling-Approach.pdf.aspx

[SAS19] SAS EVAAS Statistical Models and Business Rules of OH EVAAS Analyses, http://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Ohio-Report-Cards/Value-Added-Technical-Reports-1/Technical-Documentation-of-EVAAS-Analysis.pdf.aspx

[White18] J. White and J. Bell, Value-Added Reporting in Ohio, JEOC Testimony, April 2018, http://jeoc.ohio.gov/Assets/EventFiles/342.pdf

Appendix A: URM Model Predicted Score and Growth Computation

The URM model is represented as follows, according to the SAS documentation:

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i$$

Figure 3: Formula 14 from [SAS19]

Where:

- The score to be predicted serves as the response variable (y, the dependent variable).
- The covariates (x's, predictor variables, explanatory variables, independent variables) are scores on tests the student has already taken.
- The categorical variable (class variable, factor) are the teacher(s) from whom the student received instruction in the subject/grade/year of the response variable (y).

In particular [SAS2019]:

"The μ terms are means for the response and the predictor variables. $\alpha j$is the teacher effect for the j_{th} teacher—the teacher who claimed responsibility for the i_{tth} student. The β terms are regression coefficients."

The predicted score for s student is calculated as follows:

$$\hat{y}_i = \hat{\mu}_y + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots$$

Figure 4: Formula 18 from [SAS19]

Where:

- “The y_i term is nothing more than a composite of all the student’s past scores. It is a one-number summary of the student’s level of achievement prior to the current year.”
- “Note that the $\alpha_j$term is not included in the equation. Again, this is because y_i represents prior achievement before the effect of the current district, school, or teacher.”
- “The different prior test scores making up this composite are given different weights (by the regression coefficients, the βˆs) in order to maximize its correlation with the response variable.” Note that the beta terms are calculated for a particular student as per equation 17 in [SAS19].

As a simplified example, the predicted score for a student, assuming:

- 3 prior test scores (as required), assuming one score each year for convenience
- A statewide average student score of 50 in each prior year
- A student score of 0 in each prior year
- Equal weight of all past three test years (for convenience - the weights can be any positive numbers that sum to 1)

The predicted score is:

$$y_i = \mu_y + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \beta_3(x_{i3} - \mu_3)$$

$$y_i = 50 + (1/3)(0 - 50) + (1/3)(0 - 50) + (1/3)(0 - 50)$$

$$y_i = 0$$

Figure 5: Example Predicted Score Calculation

Then, “Growth is the *change in achievement* over time for a group of students, compared to an expectation based on students’ prior testing history” [White18], as shown in Figure 6.

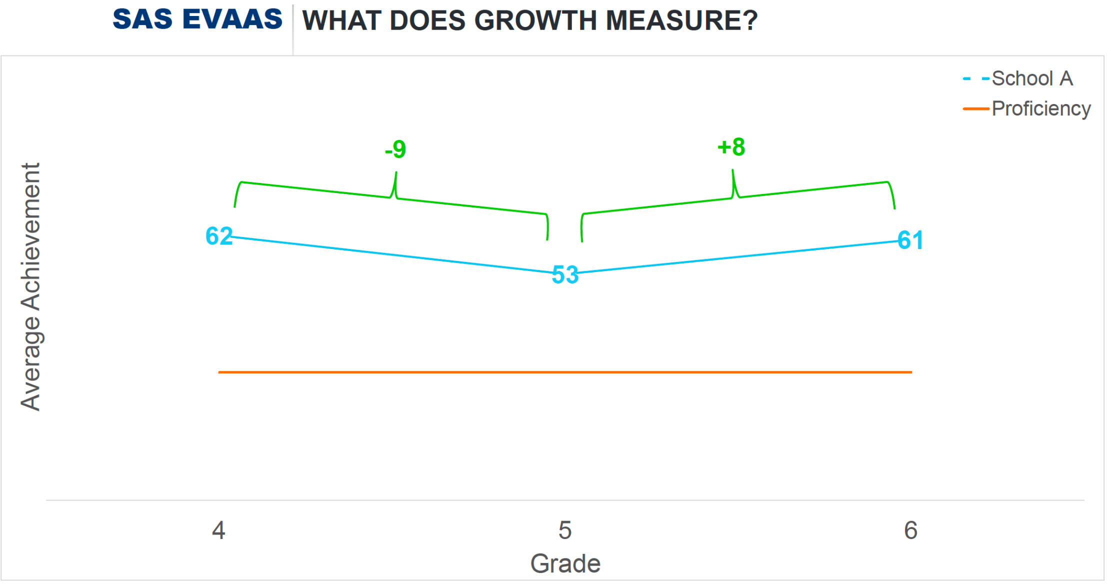

Figure 6: Growth Example from [White18]

And the growth expectation for URM is [SAS19]:
“URM definition: Students with a district, school, or teacher made the same amount of progress as students with the average district, school, or teacher in the state for that same year/subject/grade.”
“Typically, the “expected” growth is set at zero, such that *positive* gains or effects are evidence that students made *more* than the expected progress and *negative* gains or effects are evidence that students made *less* than the expected progress. “

In our example, the average student made a gain of (50 - 50) = 0 in the past two years.

The student with a zero teat score also made a gain of (0 - 0) = 0 in the past two years. Thus, a student with a **zero test score is determined as meeting the growth expectation** according to URM in this example. In reality, a zero test score gives no information about what a student has learned.

This doesn’t have to be a hypothetical question. SAS has the predicted scores for students already calculated [SAS15]. One only need ask for the data. A histogram by score bins should suffice. Note that any scores lower than 20% of the maximum score are equivalent to guessing for a 5-item multiple choice test, and would not show any evidence of material mastery.

Fundamentally, **growth for below-grade-level students cannot be measured with an at-grade-level test** (unless the student achieves such large growth as to put the student at grade level). The current URM metric assumes equal learning occurred for ANY pair of matching year-over-year test scores, whether the scores are 100, 50, or even 0.

If a cumulative score from lower grades through high school is used, such as in MAP testing [NWEA], and a progressive test is used offering questions at a level that a student can mostly answer correctly, then the rate at which students are growing becomes clear (in terms of test units, which must then be related back to content mastery). It’s possible to distinguish between students that are below-grade-level with high growth vs. students below-grade-level with low growth.

MRM is not immune to these problems. The MRM expected growth is:
“MRM definition: Students maintained the same relative position with respect to the statewide student achievement that year.” [SAS19].

A student maintaining a very low position relative to others, year-over-year, can hardly be assured to be learning anything.