4-18-2019

NEPC Review: The Michigan Context and Performance Report Card: High Schools 2018 (Mackinac Center for Public Policy, January 2019)

John T. Yun
Michigan State University, jyun@msu.edu

Recommended Citation

NEPC Review: The Michigan Context and Performance Report Card: High Schools 2018 (Mackinac Center for Public Policy, January 2019)

Reviewed by:

John T. Yun
Michigan State University

April 2019

National Education Policy Center

Acknowledgements

NEPC Staff

Kevin Welner
Project Director

William Mathis
Managing Director

Alex Molnar
Publications Director

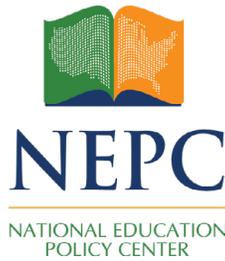# NEPC Review: The Michigan Context and Performance Report Card: High Schools 2018 (Mackinac Center for Public Policy, January 2019)

Reviewed by:

John T. Yun
Michigan State University

April 2019

Executive Summary

A recent Mackinac Center for Public Policy report, *The Michigan Context and Performance Report Card: High Schools 2018*, seeks to measure and publicize high school performance by ranking schools according to their test scores. Although this has been done previously in many contexts, this publication touts as its major contribution taking socioeconomic status into account in its school rankings. While the stated goal of the report is laudable, the reality falls far short due to at least four shortcomings. First, the study lacks both justification for and explanation of its methodological decisions. Second, the validity and reliability of combining disparate tests across different years without proper equating invalidates the findings – particularly for the high-stakes applications presented in the report. Third, the free-lunch percentage measures used in the study have a great deal of measurement error. Fourth, the use of a single predictor with unacceptably low correlations for this type of usage grossly oversimplifies and biases the estimates. Given these four shortcomings, the rankings presented in this report should be given no weight in any discussions of policy or practice. In fact, this report does a disservice by introducing questionable information in an easily readable form that is not substantiated by any credible analysis.

NEPC Review: The Michigan Context and Performance Report Card: High Schools 2018 (Mackinac Center for Public Policy, January 2019)
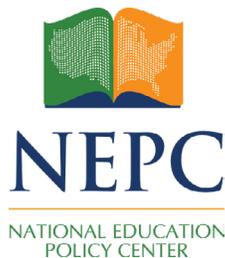
Reviewed by:

John T. Yun
Michigan State University

April 2019

I. Introduction

A recent Mackinac Center for Public Policy report by Ben DeGrow and Ronald Klingler, *The Michigan Context and Performance Report Card: High Schools 2018,* seeks to measure and publicize high school performance in Michigan by ranking schools according to their test scores. This publication is the fourth in a series published by the Mackinac Center focused on ranking schools, and touts as its major contribution to the conversation about school effectiveness its methodology for school ranking which, the authors suggest, takes into account socioeconomic status. As the report states: "The goal of this report card is to provide another helpful dimension to identifying the relative effectiveness of schools so that parents, educators and school leaders can have a more accurate and well-rounded picture of school performance.[1]" This report does not offer analysis or provide specific recommendations; it simply creates tables ranking schools by their created index, which they refer to as the "Context and Performance Score" or "CAP Score." Thus, the main innovation and purpose of this report is to suggest that this CAP Score provides information otherwise unavailable to the public, and which is more accurate than other sources of information in the state.

It is important to examine this report since it and the previous reports in the series have been used in discussions of the relative successes or failures of traditional public and charter schools in Michigan. Understanding how credible the report's rankings actually are will determine the value of the report to the broader policy argument around school performance in the state.

The text of the report simply describes the methods by which the CAP Score is calculated and describes tables based on the CAP Score. Since there are no stated conclusions or specific recommendations in the report, it is left to readers to draw their own inferences from the tables. Such inferences are inevitably inexact at best and can distract from a direct analysis

of the specifics of the report. Therefore I will confine my discussion to the key areas of the report that determine the degree to which the methods, and therefore the findings, are valid, reliable and informative.

The key issues are:

- The degree to which the methods used accurately model the relationship between socioeconomic status and school performance
- The degree to which these models can be used to create a measure that accurately represents the performance of an individual school

As discussed below, there are too many critical issues that undercut the ability of the general model to credibly rank Michigan's schools. These issues include problems with the CAP Score (including inconsistent measures over time), the measurement of free lunches (including reliable measurement over time), and the model used to estimate performance itself, which suffers from an oversimplification of the relationship between school context and performance.

II. Findings and Conclusions of the Report

The report is quite short and basic in its approach. Using the CAP Score, the report simply ranks all schools on overall CAP Score as well as determining the top and bottom 25 schools on several key variables: overall CAP Score, school type (regular district, charter, and selective), long-term performance, and finally, the largest positive and negative changes from the first report in 2012 to this one in 2019. Within each of these categories the tables contain eight pieces of information: the school's rank on their CAP Score, school name, school type (district, charter, or selective), school district/municipality, school locale (suburban, rural, urban), CAP Score, CAP Percentile Rank, and a grade (A, B, C, D, or F) based on percentile. These variables were presumably selected because they are important factors in understanding school performance and useful for characterizing the school's performance relative to other institutions. Oddly, the tables do not include the key variable used by the report to fit their regression models, the percent of free-lunch-eligible students. It is not clear whether the addition of the free-lunch data would change the rankings, but no explanation is provided for its omission from the tables and seeing the level of free-lunch eligibility in the schools would have been useful in examining the degree to which this metric contributed to the rankings.

Graphic 3 in the report ranks overall performance on the CAP Score. The report calls out charter schools and selective public schools, stating, "Public charter and selective high schools are both disproportionately represented, each making up eight of the top 17."[2] In discussing Graphic 4 (the bottom 25 schools on CAP Score), the report highlights that 5 of the bottom 25 are charter schools, but does not go on to comment about what these placements might mean for either charter schools or regular public schools. When discussing the top 25 district schools, which can be thought of as standard public schools, the report calls

out individual schools for excellence without comment on the specifics of the schools. But, in discussing the bottom 25 schools, it highlights that 60 percent are schools in large cities, which is disproportionate to their share among all public high schools in the state (around 4 percent). Again, this statement is followed by no further discussion or analysis.

The rest of the report follows this pattern. Where charter schools are present to contrast with district and selective schools, the contrasts are highlighted, but the focus is on individual schools and their successes or their need to improve without comment as to why they were performing well or poorly. To the extent that a trend can be discerned in the discussion of the results, it is that where tables had both charter and non-charter schools, a charter school was either on top or close to the top of the high-performance tables, and charters were not present in large numbers on the poor-performance tables. Similarly, when schools in large cities were located on the tables, the report highlights the degree to which the city schools were disproportionately present on the poor-performance lists and disproportionally absent from the good-performance lists.

III. The Report's Rationale for Its Findings and Conclusions

The report does not offer any conclusions or explanations of high or low school rankings. Thus the entire rationale for the report apparently rests on the credibility of the CAP Score itself. If the CAP Scores are highly reflective of the actual status of schools across the state then the information in the tables could provide an additional perspective for all stakeholders in Michigan. If the analysis is likely not to provide such estimates, the report's usefulness for any purpose is virtually nil. To test this foundational requisite requires exploring how the report calculated its rankings and how that approach was informed by the relevant literature.

IV. The Report's Use of Research Literature

The report does not cite any literature beyond the previous reports in the series. This is somewhat understandable given that this is the fourth in a series. However, previous reports did not offer rationale for the central assumption of their analysis—the use of the share of free-lunch-eligible students as the only measure of school context. This lack of any justification is highly problematic considering that the analysis is based entirely on this single measure.

There is ample research literature about the characteristics of good educational measures, as well as a number of research papers that have used residuals in a regression to measure school over/underperformance as an indicator of school quality.[3] The inclusion of a literature review or details of the CAP Score's construction could have provided explanations for the methodological decisions. As is, we are left with a decontextualized set of tables purporting to provide additional context for understanding school performance.

V. Review of the Report's Methods

Creating the CAP Score

The basic approach used to create the CAP Score and school ranking is quite straightforward and easily understood, but the details are highly questionable. The report uses state data from the 11th grade tests included in the Michigan Merit Examinations (mathematics, English language arts, science, and social studies) over four years. The report created an examination index that mapped the different examinations onto the same scale with a mean of 100 and a standard deviation of 15 and averaged across all of the examinations. This resulted in each school having a composite test score averaged across 15 tests (four tests over four years minus the 2018 science test which was withheld because it did not appropriately measure proficiency, a clear red flag for the use of the M-STEP[4] examination).

It is important to note that the 2015 tests were different from the tests used in 2016, 2017 and 2018 because Michigan replaced M-STEP math and language arts tests in 11th grade with the math and evidence-based reading and writing tests of the SAT in the Michigan Merit Examination. This is problematic since it is unclear whether the tests are measuring equivalent constructs across the years. To address this concern, the report estimated that the state language and mathematics tests explained roughly 53 and 66 percent respectively of the SAT language and mathematics tests. While this explains a moderately large share of the variation, it still means that nearly half to a third of the variation between the tests remains unexplained, suggesting that the tests may not be suitable substitutes for one another. In addition, in order to normalize the test scores on the scale that they did in this report, the distribution of the scores by school should be normal. There was no discussion of the distribution, just the application of the normalization formula, which leads to a lack of transparency about these key assumptions. Finally, normalization of the scores does not mean the tests are equated. There are established ways to equate tests,[5] but the report does not explain the decision not to pursue these avenues. This is particularly troubling given the myriad examinations that are included in the CAP Score across different times and within a single CAP Score average.

According to the report, the second step in creating the CAP Score and rankings was to fit a regression predicting the school's composite test score using as the sole predictor the percentage of 11th grade students who applied and qualified for the federal free-lunch program.[6] The actual composite test score for the school is then divided by the predicted composite test score and multiplied by 100. This results in an index that takes the values of 100 when schools perform as predicted, greater than 100 when schools perform better than predicted, and less than 100 when they perform worse than predicted—this index is the CAP Score. This approach to estimating the performance of schools is completely dependent on the validity of the composite test score as well as the quality of the model used to predict that composite.

Unfortunately, the report's methods fail to either provide a strong justification for the quality of the initial composite test score (as discussed above) or to create a high-quality model that appropriately predicts the composite test score. There are key two reasons for this fail-

ure. First, models are only as good as the variables used in them, and the free-lunch variable and the composite test score are both flawed. Second, the use of free lunch as the only predictor opens the analysis to omitted variable bias, which renders the predictions for each of the schools highly questionable—particularly those schools at the higher and lower end of the free-lunch distribution where bias in the estimates will have the highest leverage. These schools are also of the most interest, since they include the schools with the highest poverty levels and are likely to be in the greatest need of assistance and support.

Problems with Free-Lunch Eligibility

The report acknowledges that the percent of free-lunch students is not a great measure of school poverty since many eligible students (particularly in high schools) do not apply for free lunch, while some ineligible students do. However there is very little discussion of why this is the case, and particularly how the measure's application over this time period and in this way could be problematic. Owens, Reardon and Jencks[7] found that the Common Core of Data over-counted FLE students and that these over-counts were more prevalent in schools with higher shares of actual free-lunch-eligible students. This relationship suggests that by using free lunch as the only indicator of poverty, the report may overestimate the relationship at higher levels of free lunch and underestimate it for lower levels. Another key issue is that the Healthy, Hunger-Free Kids Act, passed by Congress in 2010, included the Community Eligibility Provision (CEP) allowing all students in a school to receive free meals if 40% of its students met the community met minimum free-lunch eligibility requirement. Also, if the school's free-lunch percentage is 62.5 percent or more, the school will receive 100 percent reimbursement at the federal "free" rate.[8] Thus, there is no incentive for schools who participate in the CEP to collect free-lunch numbers above the 62.5 rate since those 62.5 is all that is necessary to receive full benefits. Michigan was eligible for this provision starting in the 2011-2012 school year and currently has 692 schools receiving the benefit, many of them high schools.[9] These schools are likely to have incorrect values of free-lunch eligibility and thus incorrect values of predicted composite test scores, which would affect their CAP Score and subsequent ranking.

Finally, the authors use only the free-lunch-eligible students, not the free- and reduced-lunch-eligible students. The authors cite a previous report[10] that suggests that free-lunch eligibility is correlated with test scores but reduced-lunch numbers are not. However, the evidence in the earlier report shows a very distinct pattern which shows an actual relationship between reduced lunch and test scores. This apparent contradiction occurs because there are two reasons schools can have low numbers of students on reduced-price lunches— either they are relatively wealthy and have no one who qualifies for reduced-price lunch, or they are very poor and everyone qualifies for free lunch. Thus, at low values of reduced-price lunch there are both very poor and very wealthy schools, resulting in very high and very low CAP Scores. As the percentage of reduced-lunch individuals increases, wealthy schools become less wealthy (since the share of poorer students increases), while poor schools become less poor since fewer of their students qualify for free lunch. This means that on average there is no relationship between reduced lunch and test scores, but there is clearly important information in the reduced-lunch numbers which the report ignores. When combined with the free-lunch numbers, this problem of relationship is mitigated and provides a better

overall measure of school poverty. This misinterpretation of their own analysis means that the report's choice to omit the reduced-lunch numbers was based on faulty assumptions. It also suggests that the level of poor and near-poor students are likely underestimated in some schools, resulting in incorrect predicted scores for those schools—directly affecting their ranking using the CAP Score.

Problems with the Model

Finally, the report did not use any other variables to predict school performance, only the flawed measure of free-lunch eligibility. This makes its analysis susceptible to omitted variable bias. Which could make the critical estimate of predicted CAP Score incorrect. Omitted variable bias occurs when variables that are critical to predicting outcomes in a model are left out of the regression and those variables are correlated with both the outcome and the key predictor[11]—in this case free-lunch eligibility. There are many variables that meet these criteria. In fact, the report identifies other important variables by creating tables that are disaggregated on just such variables; for example, school locale (rural, suburban, and urban) and school type (district, charter, selective). In addition, there are a large number of additional, and readily accessible, variables that could have been included in their predictive model such as school racial composition, percent of English language learners (ELL), percent of special education students, and school size.

These alternative variables would likely result in changes in the schools' ranking—this is particularly true when you note that the difference between number 10 and 25 on the top ranking schools in the state is only 10 points on the CAP Score (which ranges from 75 to 143). Omitted variable bias plus poor measurement of free-lunch eligibility could easily result in a large change in this index, rendering the reliability of these rankings questionable.

Finally, the metric used in Graphics 12 and 13 examined the changes in school's CAP Scores from their first report issued in 2012 to this most recent report in 2019 (which included data from 2008 to 2018). This analysis purports to identify schools with the largest positive and negative changes during this time period. Their approach to this is questionable at best for several key reasons. First, the tests used in the CAP Score from the earlier report were completely different from those used in the most recent report. Michigan used the Michigan Educational Assessment Program (MEAP) exams from 1969 until 2015, when they were replaced by the M-STEP. There is no attempt in the report to demonstrate that the MEAP examinations are equivalent to the M-STEP, nor was there any attempt to equate the scores. As a result, any comparison using these very different scores is completely unwarranted. In addition, between the two reports used for their change calculations the Community Eligibility Provision for free and reduced lunch was also implemented, introducing even more variability into individual school's measures of free-lunch eligibility. Since this is the only metric used to calculate the predicted CAP Scores, it is likely that some portion of the change found in the report between these years is due to measurement and estimation error. How much is impossible to approximate, so relying on the model estimates to identify schools that have improved or declined the most is at best unwarranted and at worst misleading.

VI. Review of the Validity of the Findings and Conclusions

The use of a single variable in a regression may be justifiable (but not advisable) if only intended to illustrate a general relationship between free-lunch eligibility and examination score. However, to then use that single predictor regression to report a reliable estimate of that relationship is questionable due to the issues of omitted variable bias and measurement error. More egregious is the use of this flawed model to estimate values of over- or under-performance for specific schools since these individual schools' estimates are likely to be substantially different if one includes additional measures of school context.

This is the core criterion on which these rankings must be measured: to what degree do the estimates underlying the rankings accurately represent school performance. Given the issues described above, the answer must be that these rankings are deeply flawed and cannot be seen as viable measures of school quality or performance. At the extremes and across very large categories, the rankings may distinguish between the very worst and the very best, but the report suggests it is achieving a level of detail in the ranking that is well beyond the capabilities of the approach used in the report.

VII. Usefulness of the Report for Guidance of Policy and Practice

The purpose of the report is laudable. It is important to produce additional information that includes school context to control for the relationship between socio-economic status and school performance. However, incorporating that context is neither simple nor straightforward—particularly not at the level of rating an individual school's performance and ranking it against all other schools. Much more methodological care and justification must be provided for any ranking to be credibly used for decision-making. Given the lack of both justification and explanation for the methodological decisions made in the paper, as well as the questionable validity and reliability of the CAP Score, the measurement error inherent in the free-lunch metric, the incorrect justification for not including reduced-lunch students, and the oversimplified single predictor regression, the details of these rankings should be given no weight in any discussion of policy or practice. These problems are blatant and should be obvious to any trained analyst. The lack of any discussion addressing these shortcomings is deeply troubling. In fact, instead of positively contributing to the policy discussion around schooling, this report does a disservice by introducing questionable information in an easily readable form that is not substantiated by any credible analysis.

Notes and References

1   DeGrow, B., & Klingler, R. (2019). *The 2018 Michigan public high school context and performance report card*. Midland, MI: The Mackinac Center for Public Policy. Retrieved February 28, 2019, from https://www.mackinac.org/archives/2019/s2019-01.pdf

2   DeGrow, B., & Klingler, R. (2019). *The 2018 Michigan public high school context and performance report card*. Midland, MI: The Mackinac Center for Public Policy. Retrieved February 28, 2019, from https://www.mackinac.org/archives/2019/s2019-01.pdf

3   Linn, R.L. (1993). *Educational measurement, Third Edition*. New York, NY: American Council on Education: Oryx Press;

     Kane, T.J., & Staiger, D.O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*(4), 91-114;

     Pituch, K.A. (1999). Describing school effects with residual terms: Modeling the interaction between school practice and student background, *Evaluation Review, 23*(2), 190-211.

4   The M-STEP is the state examination used to test student proficiency in the 3-8th and 11th grades.

5   Livingston, S.A. (2014). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service. Retrieved March 20, 2019, from https://www.ets.org/Media/Research/pdf/LIVINGSTON2ed.pdf

6   The adjusted R^2 for the 2015-2018 regressions ranges from a low of 0.52 to a high of 0.68. While these are relatively large values for R^2, the real question remains whether the model correctly predicts a CAP Score for the purpose of ranking.

7   Owens, A., Reardon, S., & Jenks, C. (2016). Income segregation between schools and districts. *American Educational Research Journal, 53*(4), 1159-1197.

8   US Department of Education (2015). *The community eligibility provision and selected requirements under Title I, Part A of the Elementary and Secondary Education Act of 1965, as amended*. Washington, DC: Department of Education. Retrieved March 15, 2019, from https://www2.ed.gov/programs/titleiparta/15-0011.doc

9   Author's tabulation at 2018 School Level CEP Information. Retrieved March 15, 2019, from https://www.michigan.gov/mde/0,4615,7-140-66254_50144---,00.html

10   Spalding, A (2013). *The Michigan Context and Performance Report Card: Public Elementary and Middle Schools, 2013*. Midland, MI: The Mackinac Center for Public Policy. Retrieved March 15, 2019, from https://perma.cc/S8EP-7BHJ

11   Green, W. (2017). *Econometric Analysis*, 8th Edition. New York, NY: Pearson.