NEPC Review: Still a Good Investment: Charter School Productivity in Nine Cities (University of Arkansas Department of Education Reform, November 2023)

Reviewed by:

Mark Weber

Rutgers, The State University of New Jersey
New Jersey Policy Perspective

January 2024

National Education Policy Center

School of Education
University of Colorado Boulder
nepc.colorado.edu

Acknowledgements

NEPC Staff

Faith Boninger
Publications Manager

Francesca López
Academic Editor

Elaine Duggan
Production Design

Alex Molnar
NEPC Director

Kevin Welner
NEPC Director

Suggested Citation: Weber, M. (2024). *NEPC review: Still a good investment: Charter school productivity in nine cities*. Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/review/charter-productivity

Funding: This review was made possible in part by funding from the Great Lakes Center for Educational Research and Practice.

The National Education Policy Center (NEPC), a university research center housed at the University of Colorado Boulder School of Education, sponsors research, produces policy briefs, and publishes expert third-party reviews of think tank reports. NEPC publications are written in accessible language and are intended for a broad audience that includes academic experts, policymakers, the media, and the general public. Our mission is to provide high-quality information in support of democratic deliberation about education policy. We are guided by the belief that the democratic governance of public education is strengthened when policies are based on sound evidence and support a multiracial society that is inclusive, kind, and just. Visit us at: http://nepc.colorado.edu

NEPC Review: Still a Good Investment: Charter School Productivity in Nine Cities (University of Arkansas Department of Education Reform, November 2023)

Reviewed by:

Mark Weber

Rutgers, The State University of New Jersey
New Jersey Policy Perspective

January 2024

Summary

In an ongoing series of reports, The Department of Education Reform at the University of Arkansas has argued that charter schools suffer from inequitable funding compared to public district schools. The latest report, *Still a Good Investment: Charter School Productivity in Nine Cities,* goes further: It claims that charters get better results on standardized tests even as they receive less in revenues, making them more cost-effective. To support this claim, the report combines its own database of school revenues with results from another think tank that compares charter school students' academic growth with matched students in public schools. There is substantial reason to question both the fiscal and academic outcome comparisons, casting doubt on the report's conclusions. But the report makes an additional critical error: When transforming academic outcomes into a common scale, it conflates test scores with academic "growth," rendering its outcome measure invalid. It then compounds this error by using a simplistic "return on investment" measure that eschews the complexity found in serious research on education cost modeling. The report, consequently, is yet another methodologically flawed document that should be ignored by policymakers and stakeholders.

NEPC
NATIONAL EDUCATION POLICY CENTER

NEPC Review: Still a Good Investment: Charter School Productivity in Nine Cities (University of Arkansas Department of Education Reform, November 2023)

Reviewed by:

Mark Weber

Rutgers, The State University of New Jersey
New Jersey Policy Perspective

January 2024

I. Introduction

The Department of Education Reform at the University of Arkansas has produced a series of reports that allege charter schools are inequitably funded compared to public district schools. For nearly a decade, National Education Policy Center (NEPC) reviewers have found serious and persistent flaws in these reports, calling into question their validity and usefulness for informing policymaking.[1]

The latest report in this series is *Still a Good Investment: Charter School Productivity in Nine Cities,* by Alison H. Johnson, Josh B. McGee, Patrick J. Wolf, Jay F. May, and Larry D. Maloney (hereafter referred to as the "report").[2] This report makes a significant leap beyond its predecessors: It claims that charter schools are more "cost-effective," achieving better results for less money. As proof of the alleged superior performance of charter schools, the report cites another series of reports from the Center for Research on Educational Outcomes (CREDO).[3] The report transforms CREDO's results into points on the National Assessment of Educational Progress (NAEP), a common exam administered to a sample of students in all states. It then claims charter school students in nine cities get more NAEP points per $1,000 spent than similar students in traditional public schools (TPSs).

Unfortunately, the same methodological flaws found in The University of Arkansas's earlier reports—along with a glaring new error—once again render this report's conclusions invalid.

II. Findings and Conclusions of the Report

Based on the nine cities included, the report asserts that charter school students perform better than traditional public school (TPS) students on standardized tests, as measured in "estimated" points on the eighth-grade National Assessment of Educational Progress (NAEP) exam in reading (2.4 points) and math (1.3 points). At the same time, the report cites data from a previous University of Arkansas report to claim that charter schools receive less revenue per pupil than TPSs.[4] Because of their allegedly better performance and lower cost, the report concludes that charter schools are 40 percent more cost-effective than TPSs. The report goes on to calculate a higher "return on investment" (ROI) for charters, asserting that higher test scores lead to higher lifetime earnings.

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

The validity of this report relies on five key assumptions:

- The CREDO findings are valid measures of the differences in charter and traditional public school (TPS) student performance, validly accounting for differences in student characteristics.
- These measures can be transformed into estimated NAEP points, thereby creating a common scale.
- Valid measures of charter and TPS finances are used for comparisons.
- A simple calculation of "cost-effectiveness," using student outputs (standardized test outcomes, measured in estimated NAEP points) per inputs (revenue per pupil), validly assesses the relative efficiency of the charter and TPS sectors.
- This efficiency measure can then be multiplied by projections of lifetime earnings, based on test outcomes, to calculate a "return on investment" (ROI) for charters and TPSs.

All of these assumptions must hold up to scrutiny for the report's conclusions to be valid.

IV. The Report's Use of Research Literature

As in previous University of Arkansas Department of Education reports, most of the citations are to the authors' own works, or other reports released by the department's School Demonstration Project. The report cites a few pieces of outside research that address charter school outcomes; however, it cites no outside research related to charter school spending differences, education cost-modeling, or critiques of the CREDO reports.

V. Review of the Report's Methods

The report's methodology is based on the calculation of the "short-term cost-effectiveness" of the traditional public school (TPS) and charter school sectors. The report defines this measure as "average NAEP points earned per $1,000 in funding allocated per pupil." Essentially, the report is creating a ratio between an output (scores on a test) and an input (revenue per pupil).

The "Output": Equated NAEP Scores

Annual student achievement tests are administered by states; each state uses its own set of exams, making it impossible to directly compare the results across different states. The only nationally administered common exam across states is the National Assessment of Educational Progress (NAEP), which is administered to samples of fourth and eighth graders every two years.[5] Theoretically, if state assessment outcomes could be "linked" to NAEP outcomes, different state exam results could be compared.

Researchers at the Educational Opportunity Project at Stanford University have, in fact, produced datasets that use NAEP linking to allow for cross-state comparisons.[6] Their methods are complex, requiring the use of NAEP cut-score data for each state. Notably, because of the nature of the data and the modeling used to construct these equated scores, these researchers have explicitly cautioned *against* using these data to make comparisons between charter and public district schools.[7] That said, linking state test scores to NAEP outcomes is at least theoretically possible.

The current report, however, does not directly compare charter school and district school test outcomes. Instead, it relies on reports from the Center for Research on Educational Outcomes (CREDO) that focus on specific cities, showing the differences in "growth" between students enrolled in charter schools and "matched" students in public district schools.[8]

On multiple occasions, NEPC authors have questioned the validity of CREDO's methods.[9] Particularly relevant to the University of Arkansas report are the following critiques:

- The variables used to match charter and public district students are crude and do not capture important observable differences between students.
- Even more refined variables would not capture important unobserved differences between students (such as parental involvement), biasing the estimates of a school's effectiveness.
- Because only students who can be "matched" are included in the study, the results can't be validly generalized to the entire student population.

Any of these points, even on their own, would be enough to call into question the use of the CREDO findings in the current report. There is, however, an additional, serious problem with the report's methods: The report attempts to convert CREDO effect sizes—*not* state test scores—from different states and grade levels into NAEP points, based on an unvalidated

assumption that an effect size of 1.0 standard deviations in the CREDO reports is equal to 38 NAEP points.

The issue is complex (see a full explanation in the Appendix to this review). In brief, it is enough to note that nowhere in the report is there any explanation as to why a standard deviation in CREDO effects is equal to 38 NAEP points, let alone any empirical evidence to support the conversion. Serious research that uses educational testing outcomes should include, at the very least, a discussion of the validity issues involved in making such a conversion. The report, however, simply sidesteps the issue, leaving readers with no basis upon which to accept the validity of its methods.

The "Input": Revenue per Pupil

Setting aside the problems with the output measure, the report still needs a valid fiscal measure to calculate "short-term cost-effectiveness." It relies on a previous University of Arkansas report that claims to have assembled a proprietary dataset of charter and TPS revenue data. Earlier this year, I outlined the problems with the methods for collecting these data[10]:

- The collection and assembly of the data is inadequately documented.
- Revenues from non-governmental sources are reported with little to no documentation of the sources.
- Revenues for activities, programs, and functions conducted by school districts, but not charter schools, are included in district totals.
- No attempt is made to adequately account for student population differences that drive differences in school spending.

This last point is particularly problematic. The premise of the CREDO studies is that they measure learning gains by comparing students who have similar characteristics. Again, there is good reason to believe the CREDO methodology is inadequate to the task, but at least the methods attempt to account for student differences that affect learning outcomes.

In contrast, no effort is made to adjust the financial figures used in the report to account for student differences. This is in spite of the fact that the earlier report concedes that special education classification rates affect school spending.[11] It is well understood that student characteristics affect both academic outcomes and costs; adjusting one and not the other produces measures of efficiency that are, again, invalid.

The Problem With "Short-Term Cost-Effectiveness"

Education cost modeling has a decades-long history.[12] Because of the intricate relationship between school finances, school outcomes, student characteristics, and schooling contexts, the models used in high-quality research are often complex.[13] In contrast, the report's "short-term cost-effectiveness" measure simply takes a dubious outcome measure, translates it into

another scale with questionable validity, and sets it in a ratio to an equally dubious revenue figure. That it does so for a convenience sample of nine cities only further calls into question the validity of the report's findings.

VI. Review of the Validity of the Findings and Conclusions

The report uses its "cost-effectiveness" measure to calculate a "return on investment" (ROI): the lifetime earnings of a student divided by the amount spent on their schooling. The report alleges that charter students realize a higher ROI than TPS students because they get higher test scores, which translate into higher earnings. The ROI, however, is based on the "cost-effectiveness" measure, which is invalid. In addition, the report relies on a single study that gives a "conservative estimate" of "13 percent per standard deviation of achievement," based on a sample of previous studies that show varying results.[14] The report provides no further empirical evidence to justify using this estimate; consequently, the ROI figure, even if it were based on valid inputs and outputs, is at best little better than a guess.

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

By themselves, the use of either the CREDO findings or a small, weakly documented dataset of district and charter revenues would be enough to call this report's findings into question. But the erroneous transformation of the CREDO results into NAEP points, combined with a simplistic ROI figure, thoroughly invalidates any of the report's findings. As with the previous reports of the School Demonstration Project that purport to show funding inequities between charter and public district schools, this report should be ignored by policymakers and stakeholders when forming charter school policies.

Appendix: Issues With Equated NAEP Points

In its technical appendix, the University of Arkansas report lays out its method for converting CREDO's findings into NAEP-equated scores, using Indianapolis's traditional public schools (TPS) as an example:[15]

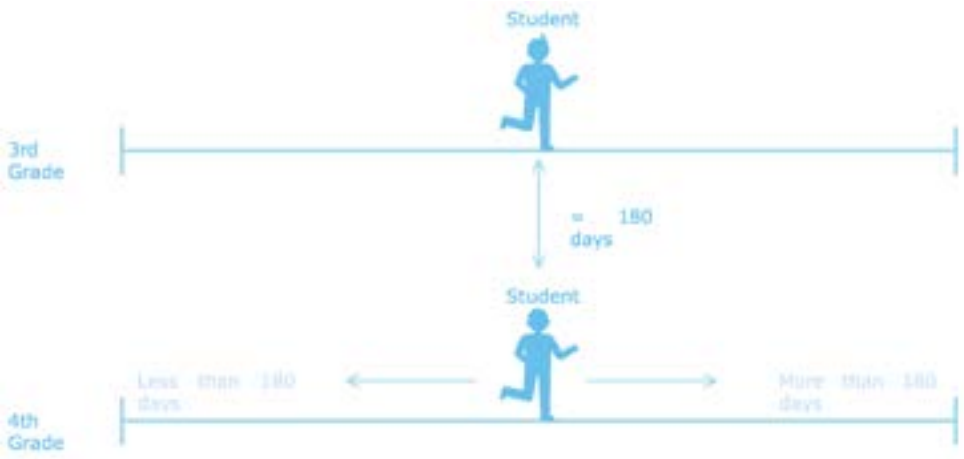
> *City TPS NAEP Reading Score*
> *= State Average NAEP Reading Score*
> *+ (SD difference between City TPS and State Average NAEP Reading Score*
> ** 1 SD on NAEP Reading Exam) = 266 + (-0.08 * 38) = 263.01*

To make the conversion between the CREDO findings and NAEP scores, the report uses a common statistical measure of variation: a standard deviation (SD). Assuming a "normal" or bell-curve distribution of scores, about two-thirds of any test's scores will be within one standard deviation of the mean (or average). Because test outcomes are measured on different scales, converting them to SDs allows differently scaled tests to be compared.[16]

Here, the report converts the effect sizes found by CREDO (i.e., the "treatment" of attending a charter school as opposed to a public school in Indianapolis), which are reported in standard deviations, into NAEP reading scores. In this example, CREDO reports an effect size of -0.08 SD for Indianapolis Public School students.[17] The report multiplies this by 38—allegedly, the NAEP point equivalent of a 1.0 SD CREDO effect size—and adds that negative number to the state average NAEP (266), calculating a NAEP score of 263.01 for Indianapolis Public Schools.

The critical assumption in this conversion—one glossed over in the report—is that a standard deviation in CREDO effect sizes is equal to 38 NAEP points. It is true that, in 2022, the Grade 8 NAEP scores had a standard deviation of 38 points.[18] But this was based on a distribution of a single set of scores, and not a distribution of *changes* in scores over two or more years—which is what the CREDO studies effect sizes actually measure.

As CREDO explains in its Technical Appendix, the outcome variable in its model is the difference in a student's "z-score"—its place in a distribution of scores, as measured in standard deviations—from one year to the next.[19] CREDO explains this in its reports by using figures like this one, found in its Indianapolis report[20]:

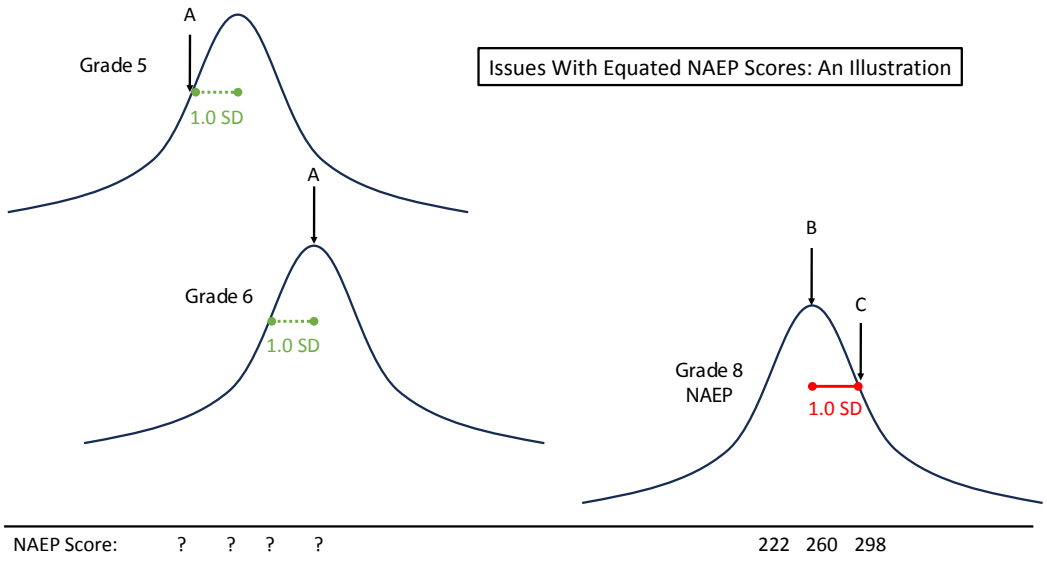


A student who scores at the mean in both third and fourth grade is designated as having "typical" growth: In other words, that student's change in z-scores is 0.0 SDs. The effect CREDO measures is not, then, of test outcomes per se; it is an effect measuring outcome *growth*, reliant on assessing the performance of students, relative to others, in two or more time periods.

In the figure above, the students in third and fourth grade are at the same place in the distribution of scores; this is because they occupy the same position *relative to other students in their grade*. If a student moved up (or down) in position relative to other students, the CREDO reports would show an effect: not because their scores were higher or lower, but because their position relative to other students had shifted.

To illustrate this further, consider the figure below (this is my own illustration and not a figure from either the CREDO or University of Arkansas reports). Student "A" takes a state test in fifth grade; their score is 1.0 SD below the mean (average) score, as shown in the distribution of the scores for that grade. Note the distribution is a bell curve, also called a "normal" distribution.

The next year in sixth grade, Student A scores right at the mean for their cohort. The CREDO reports would say that student has shown 1.0 SD of *growth*; Student A moved up in their position relative to other students in their cohort. Notice, however, what is not known: the scores for Student A on either the fifth- or the sixth-grade tests, as well as the mean score for their cohort on either test. Neither is needed, as the CREDO methods measure growth *relative to other students*.



In contrast, consider the third distribution: the NAEP Grade 8 reading scores for 2022. Students who take this test take it only once, and it is only given in eighth grade. The mean for the 2022 cohort was 260 points; this is Student B's score. The scores have a standard deviation of 38: This means that students who scored at the 84th percentile—which is 1.0 SD

above the 50th percentile—earned a score of 298 (260 plus 38). Student C represents these students.

Note that the NAEP scores are reported along a *scale*: the underlying range of numeric values representing various levels of achievement on the test. In the same way that 298 is the score 1.0 SD above the mean, 222 is 1.0 SD below. Importantly, we do not know the fifth- or sixth- grade means on this scale, because fifth and sixth graders do not take the test.

To make its conversion, the University of Arkansas report makes a critical assumption: that the difference between Student A's relative position in the fifth- and sixth-grade distribution of scores is equivalent to the score difference between Students B and C—*in NAEP points*. Yes, it's true both are 1.0 SD; however, the conversion of standard deviations to NAEP points can't be made in the lower grades because we don't know where they fall on the Grade 8 NAEP scale.

The report has simply assumed that the difference in NAEP points between Students B and C is the same as the difference between Student A in fifth and sixth grade, because the difference in both is 1.0 SD. But without a common underlying scale, that difference cannot be assumed; serious research using this method should present some evidence that the two are equivalent. The report provides no such evidence.

Now, it is true that the NAEP is also given in Grade 4, and that the scores given for Grades 4 and 8 might appear to be on a common or "vertical" scale.[21] In 2022, the Grade 4 mean score on the reading NAEP was 217, with a standard deviation of 40.[22] If one assumed an imaginary NAEP test were given in Grades 5, 6, and 7, and assumed the scores were normally distributed with a standard deviation of 40, and assumed growth along the NAEP scale was linear, one could argue that the NAEP scale used in Grade 8 is common across all grades, and that 1.0 SD in any grade was equivalent to another in NAEP points.

The problem, as psychometricians have noted, is that there is no empirical evidence to back up these assumptions.[23] While the NAEP reading scale has the appearance of a vertical scale, issues with its validity have been described by testing experts, and NAEP does not use cross-grade scaling on more recent assessments.[24] Assuming a common scale across grades on the NAEP—the core assumption of the report's methodology—is, therefore, contradicted by best practices in the field.[25]

One further point: The CREDO reports make their own conversion of effect sizes into "days of learning," based on a poorly documented assumption that a grade level of growth is equal to about one-quarter of a standard deviation.[26] This conversion, however, is, by its own creator's admission, a "rule of thumb," and "has not been extensively researched."[27]

According to the *Standards for Educational and Psychological Testing, "Evidence of validity, reliability, and fairness for each purpose for which a test is used in a program evaluation, policy study, or accountability system should be collected and made available."*[28] Best practices require the University of Arkansas report to fully document its validity argument in favor of converting CREDO effect sizes into NAEP eighth-grade reading scores. In particular, the argument must address the questionable validity of assuming a common

scale across all tested grades, and explain why the NAEP eighth-grade reading score scale is an appropriate choice for a common scale. Unless and until this argument is convincingly made, the validity of the report's methods and conclusions remain in doubt.

Notes and References

1 For an overview of the NEPC reports, see: Weber, M. (2023). *Evaluating research that alleges funding disparities between charter and district schools*. Boulder, CO: National Education Policy Center. Retrieved November 29, 2023, from https://nepc.colorado.edu/publication/funding-disparities

2 Johnson, A.H., McGee, J.B., Wolf, P.J., May, J.F., & Maloney, L.D. (2023). *Still a good investment: Charter school productivity in nine cities*. Fayetteville, AR: School Choice Demonstration Project, University of Arkansas Department of Education Reform. Retrieved November 29, 2023, from https://scdp.uark.edu/still-a-good-investment-charter-school-productivity-in-nine-cities/

3 Center for Research on Educational Outcomes (CREDO). (n.d.).[Webpage]. Retrieved January 10, 2024, from https://credo.stanford.edu

4 Johnson, A.H., McGee, J.B., Wolf, P.J., May, J.F., & Maloney, L.D. (2023). *Charter school funding: Little progress towards equity in the city*. Fayetteville, AR: School Choice Demonstration Project, University of Arkansas Department of Education Reform. Retrieved November 29, 2023, from https://scdp.uark.edu/charter-school-funding-little-progress-towards-equity-in-the-city/

5 National Assessment of Educational Progress (NAEP). Washington, D.C.: Institute of Education Science, National Center for Education Statistics. Retrieved December 4, 2023, from https://nces.ed.gov/nationsreportcard/

6 The Educational Opportunity Project at Stanford University. (n.d.). [Webpage]. Retrieved December 4, 2023, from https://edopportunity.org/

7 "SEDA-style cohort growth measures may overstate charter school growth in the three states we examine, suggesting that these estimates should not be used to draw comparisons between charter and traditional public school sectors." (p. 4). Reardon, S.F., Papay, J.P., Kilbride, T., Strunk, K.O., Cowen, J., An, L., & Donohue, K. (2019). *Can repeated aggregate cross-sectional data be used to measure average student learning rates? A validation study of learning rate measures in the Stanford Education Data Archive*. (CEPA) Working Paper No. 19-08. Palo Alto, CA: Stanford Center for Education Policy Analysis. Retrieved December 4, 2023, from http://cepa.stanford.edu/wp19-08

8 Center for Research on Educational Outcomes (CREDO). (n.d.). *City studies*. Retrieved December 4, 2023, from https://credo.stanford.edu/research-reports/city-studies-3/

9 Maul, A. (2015). *NEPC review: Urban charter school study report on 41 regions 2015*. Boulder, CO: National

Education Policy Center. Retrieved November 29, 2023, from http://nepc.colorado.edu/thinktank/review-urban-charter-school

Miron, G. & Shank, C. (2017). *NEPC review: Charter management organizations 2017*. Boulder, CO: National Education Policy Center. Retrieved November 29, 2023, from http://nepc.colorado.edu/thinktank/review-CMOs

Ferrare, J.J. (2023). *NEPC review: As a matter of fact: National charter school study III 2023*. Boulder, CO: National Education Policy Center. Retrieved November 29, 2023, from https://nepc.colorado.edu/review/charter-study

10 Weber, M. (2023). *Evaluating research that alleges funding disparities between charter and district schools*. Boulder, CO: National Education Policy Center. Retrieved November 29, 2023, from https://nepc.colorado.edu/publication/funding-disparities

11 Johnson, A.H., McGee, J.B., Wolf, P.J., May, J.F., & Maloney, L.D. (2023). *Charter school funding: Little progress towards equity in the city* (p. 15). Fayetteville, AR: School Choice Demonstration Project, University

of Arkansas Department of Education Reform. Retrieved November 29, 2023, from https://scdp.uark.edu/still-a-good-investment-charter-school-productivity-in-nine-cities/. From the report: "Controlling for the percentage of students in special education reduces the disparity by 68 percent (to about $1,707), though that factor is imprecisely estimated."

12 Duncombe, W.D. & Yinger, J. (2011). Are education cost functions ready for prime time? An examination of their validity and reliability. *Peabody Journal of Education, 86*(1), 28-57.

13 As an example, see: Gronberg, T.J., Jansen, D.W., & Taylor, L.L. (2012). The relative efficiency of charter schools: A cost frontier approach. *Economics of Education Review, 31*(2), 302-317.

14 Hanushek, E.A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466-479, 473.

15 Johnson, A.H., McGee, J.B., Wolf, P.J. May, J.F. & Maloney, L.D. (2023). *Still a good investment: Charter school productivity in nine cities* (p. 28). Fayetteville, AR: School Choice Demonstration Project, University of Arkansas Department of Education Reform. Retrieved November 29, 2023, from https://scdp.uark.edu/still-a-good-investment-charter-school-productivity-in-nine-cities/

16 I note here that transforming any two tests into a common scale does not guarantee that the tests are, in fact, measuring a student's ability in the same learning domain. Two tests may, for example, be used to assess Grade 8 reading and yield a normal distribution of scores, but each may measure very different aspects of what it means to read at that grade level. On this basis alone, the validity of the report's methods in questionable; at the least, it should be mentioned as a caution to readers in interpreting results. For more on test score linking, see:

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). Chapter 5: Scores, scales, score linking, and cut scores. In *Standards For Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

17 Center for Research on Educational Outcomes (CREDO) (n.d.). *City study 2022: Indianapolis* (p. 40). Retrieved November 29, 2023, from https://credo.stanford.edu/wp-content/uploads/2022/06/Indianapolis_slide_deck_FINAL_20220601.pdf

18 Digest of Education Statistics (2022). Table 221.75: Average National Assessment of Educational Progress (NAEP) reading scale score and standard deviation, by selected student characteristics, percentile, and grade: Selected years, 1992 through 2022. Washington, D.C.: Institute of Education Science, National Center for Education Statistics. Retrieved November 29, 2023, from https://nces.ed.gov/programs/digest/d22/tables/dt22_221.75.asp

19 Center for Research on Educational Outcomes. (2013). *National charter school study technical appendix, 2013* (p. 10). Retrieved November 29, 2023, from https://credo.stanford.edu/wp-content/uploads/2021/08/ncss2013_technical_appendix.pdf

20 Center for Research on Educational Outcomes (CREDO) (n.d.). *City study 2022: Indianapolis* (p. 34). Retrieved November 29, 2023, from https://credo.stanford.edu/wp-content/uploads/2022/06/Indianapolis_slide_deck_FINAL_20220601.pdf

21 Young, M.J. & Tong, Y. (2015). Vertical scales, Chapter 23. In S. Lane, M.R. Raymond, & T.M. Haladyna (Eds.), *Handbook of test development* (Second edition). New York: Routledge.

22 Digest of Education Statistics (2022). *Table 221.75: Average National Assessment of Educational Progress (NAEP) reading scale score and standard deviation, by selected student characteristics, percentile, and grade: Selected years, 1992 through 2022*. Washington, D.C.: Institute of Education Science, National Center for Education Statistics. Retrieved November 29, 2023, from https://nces.ed.gov/programs/digest/d22/

tables/dt22_221.75.asp

23 Thissen, D. (2012). *Validity issues involved in cross-grade statements about NAEP results*. NAEP Validity Studies (NVS) Panel, American Institutes For Research. Retrieved November 29, 2023, from https://files.eric.ed.gov/fulltext/ED528992.pdf

24 "Based on the recommendation of the NAEP Technical Review Panel, in 1991, the NAGB adopted the position that in future assessments, when possible, NAEP would report scores on a within-grade scale rather than on an across-grade or vertical scale. This position was adopted based upon the controversy regarding the across-grade scale surrounding the NAEP 1990 Mathematics Assessment (Haertel, 1991). Haertel noted that the mix of content and process categories represented in the different areas of the vertically equated scale were likely to differ, and he cited this as one reason psychometricians were not supportive of the vertical scale. This scaling issue is, in essence, a validity issue because NAEP frameworks often specify different content and process category allocations by grade level."

Huynh, H. & Schneider, C. (2005). Vertically moderated standards: Background, assumptions, and practices. *Applied Measurement in Education, 18*(1), 99-113, 102. Retrieved January 3, 2023, from https://doi.org/10.1207/s15324818ame1801_6

25 "Interpreting score differences as gains is more justified for test levels at adjacent grades than for test levels that are widely separated. It was this issue among others that led to the scrutiny of the across-grade scaling at Grades 4, 8 and 12 for NAEP."

Young, M.J., & Tong, Y. (2015). Vertical scales, Chapter 23. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (Second edition) (p. 452). New York: Routledge.

26 The CREDO reports do not adequately document their method for converting standard deviations into "days of learning." As an example, its 2022 report (p. 19), *Meta-analysis of academic recovery after COVID-19*, states, "We transform days of learning to z-scores [sic] using a standard transformation procedure. This procedure was built by Dr. Eric Hanushek and Dr. Margaret Raymond based on the 2017 4th and 8th grade test scores from the National Assessment of Educational Progress (NAEP). Using a standard 180-day school year, each one standard deviation (s.d.) change in effect size was equivalent to 590 days of learning in this study. For more details on this methodology, see Eric A. Hanushek, Paul E. Peterson, and Ludger Woessmann. 'Achievement Growth: International and US State Trends in Student Performance. PEPG Report No.: 12-03.' Program on Education Policy and Governance, Harvard University (2012)."

Following the citation leads to a report in which the only relevant passage is: "On most measures of student performance, student growth is typically about 1 full std. dev. on standardized tests between 4th and 8th grade, or about 25 percent of a std. dev. from one grade to the next" (pp. 3-4). No citations back up this claim; it is simply assumed to be true.

In fact, in another study, the lead author, Eric Hanushek, calls into question the validity of his own conversion: "Differences across grades and ages on the vertically linked NAEP tests support the rough rule of thumb that one standard deviation of achievement is equal to three to four years of schooling; see Hanushek, Peterson, and Woessmann (2012a, 2012b). Note, however, that this correspondence has not been extensively researched and is likely to vary by grade level, position in the test distribution, and other factors" (p. 610, footnote 3). The first citation here is actually a reference to the very report CREDO cites to assert the validity of its methods. These circular references do little to establish validity and are not strong evidence to justify CREDO's methods.

Center for Research on Education Outcomes (2022). *Meta-analysis of academic recovery after COVID-19* (p. 19). Retrieved November 29, 2023, from https://credo.stanford.edu/wp-content/uploads/2023/04/meta_analysis_20230420_tr.pdf

Hanushek, E.A., Peterson, P.E., & Woessmann, L. (2012, July). *Achievement growth: International and U.S. state trends in student performance, PEPG Report No.: 12-03*. Harvard's Program on Education Policy and

Governance & Education Next. Retrieved January 4, 2023, from https://files.eric.ed.gov/fulltext/ED534652.pdf

Hanushek, E.A., Light, J.D., Peterson, P.E., Talpey, L.M., & Woessmann, L. (2022). Long-run trends in the U.S. SES—achievement gap. *Education Finance and Policy, 17*(4), 608-640. Retrieved January 4, 2023, from https://doi.org/10.1162/edfp_a_00383

27 Hanushek, E.A., Light, J.D., Peterson, P.E., Talpey, L.M., & Woessmann, L. (2022). Long-run trends in the U.S. SES—achievement gap. *Education Finance and Policy*, *17*(4), 608-640. Retrieved January 4, 2023, from https://doi.org/10.1162/edfp_a_00383

28 American Educational Research Association, American Psychological Association, National Council on Measurement in Education (Ed.). (2014). *Standards for educational and psychological testing* (p. 210). Washington, DC: American Educational Research Association.