EXHIBIT A

ANALYSIS OF RELATIONSHIPS BETWEEN ACADEMIC PERFORMANCE STANDARDS AND EDUCATIONAL ADEQUACY

Expert Report of Gregory J. Cizek, PhD

In Re Delaware Public Schools Litigation (C.A. No. 2018-0029-VCL)

March 11, 2020

I. INTRODUCTION

The purpose of this report is to provide descriptive information, analyses, and expert opinions relevant to the appropriateness of drawing conclusions about educational adequacy from student performance on large-scale educational achievement testing in Delaware. The balance of this report provides: an overview of the author's qualifications (Section II); a list of other cases in which the author has provided expert testimony (Section III); Summary Opinions related to the present matter (Section IV); a list of Materials Reviewed (Section V); Background, Analyses, and Conclusions (Section VI); Summary (Section VII), and References (Section VIII). Figures and tables presented in the body of the report are reproduced individually in Appendix B; the author's curriculum vita is provided as Appendix C.

II. QUALIFICATIONS

I am Guy B. Phillips Distinguished Professor of Educational Measurement and Evaluation at the University of North Carolina at Chapel Hill, where I teach courses in assessment, applied psychometrics, program evaluation, statistics, and research methods. My scholarly work centers on large-scale student achievement testing, setting performance standards, validity, and test security. I am the author of over 300 journal articles, book chapters, conference papers, and other publications. My work has been published in journals such as *Educational Researcher, Educational Assessment, Review of Educational Research, Journal of Educational Measurement, Educational Measurement: Issues and Practice, Educational Policy, Phi Delta Kappan, Education Week* and elsewhere. I am a contributor to:

* *Handbook of Classroom Assessment* (1998) and
* *Handbook of Test Development* (2006, 2014);

editor of and contributor to:

* *Handbook of Educational Policy* (1999)
* *Setting Performance Standards: Concepts, Methods, and Perspectives* (2001)
* *Handbook of Formative Assessment* (with H. Andrade, 2010)
* *Setting Performance Standards: Foundations, Methods, and Innovations* (2012)
* *Handbook of Quantitative Methods for Detecting Cheating on Tests* (with J. Wollack, 2017)
* *Handbook of Formative Assessment in the Disciplines* (with H. Andrade & R. Bennett, 2019);

and author of:

* *Filling in the Blanks* (1999)

* *Cheating on Tests: How to Do It, Detect It, and Prevent It* (1999)

* *Detecting and Preventing Classroom Cheating: Promoting Integrity in Educational Assessment* (2003)

* *Addressing Test Anxiety in a High Stakes Environment* (with S. Burg, 2005)

* *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests* (with M. Bunch, 2007)

* *Validity: An Integrated Approach to Test Score Meaning and Use* (in press).

I am a member of the American Educational Research Association (AERA) where I have served as the Secretary for Division D (Measurement and Research Methodology) and as an officer of the AERA Special Interest Group on Professional Licensure and Certification. I am a member of the National Council on Measurement in Education (NCME), where I have served on the Standards and Test Use Committee and as Vice President and President. I am the recipient of the AERA Division D award for Significant Contribution to Educational Measurement and Research Methodology (2006) and recipient of the NCME award for Outstanding Dissemination of Educational Measurement Concepts (2007).

I am an advisory reviewer for relevant professional journals; I currently or previously have served on editorial boards for journals such as *Applied Measurement in Education*, the *Journal of Educational Measurement*, and *Educational Researcher*. I have served on the advisory board of the Buros Institute of Mental Measurement, from which I earned its "Distinguished Reviewer Award" (2005) for my technical reviews of tests. Currently, I provide expert consultation on standard setting, validity, test security, testing policy, and other applied psychometric issues to numerous local, state, and national organizations. Since 2010 I have served as a member of the Smarter Balanced Assessment

Consortium (SBAC) technical advisory committee; from 2012-2014, I assisted in the design and evaluation for setting achievement levels on SBAC assessments. I am currently serving my second term as a member of the National Assessment Governing Board, which oversees the National Assessment of Educational Progress (NAEP).

I received my PhD in Measurement, Evaluation, and Research Design from Michigan State University in 1991. I have managed national licensure and certification testing programs, primarily for medical and health-related professions, for American College Testing (ACT) in Iowa City, Iowa. Previously, I served as a test development specialist for the Michigan Educational Assessment Program (MEAP). From 1997-1999, I was elected to and served as vice-president of a local board of education in Ohio. I began my career working for five years as an elementary school teacher. A complete resume is provided in Appendix C to this report.

III. COMPENSATION AND OTHER CASES IN WHICH THE WITNESS HAS TESTIFIED AS AN EXPERT

I am compensated at the rate of $300/hour for my work on the present case. Table 1, shown below, provides a list of other cases in which I have provided expert testimony during the past five years and the topic of the expert testimony.

Table 1
Cases in which the Witness Has Testified as an Expert in the Last 5 Years

Case Name/Citation [Court]	Year	Topic of Expert Testimony
Commonwealth of Pennsylvania v. Ary Sloane, Docket No. CP-51-CR-0009924-2014 [Philadelphia Municipal Court, Criminal Section]	2014	psychometrics, validity of test scores
Commonwealth of Pennsylvania v. Barbara McCreery, Docket No. CP-51-CR-0011071-2014 [Philadelphia Municipal Court, Criminal Section]	2014	psychometrics, validity of test scores
Commonwealth of Pennsylvania v. Arthur Melton, Docket No. CP-51-CR-0011945-2014 [Philadelphia Municipal Court, Criminal Section]	2014	psychometrics, validity of test scores
Commonwealth of Pennsylvania v. Lola Marie Davis-O'Rourke, Docket No. CP-51-CR-0000916-2015 [Philadelphia Municipal Court, Criminal Section]	2015	psychometrics, validity of test scores
State of Georgia v. Dr. Beverly Hall et al., Case No. 13-SC-1179S4 [Fulton County (GA) Superior Court]	2015	psychometrics, validity of test scores

IV. SUMMARY OF OPINIONS

I have carefully reviewed relevant student performance data, documentation, and relevant research in the area of educational performance standards, accountability systems, and educational policy. I have also relied on my extensive experience as a technical advisor to state assessment programs, as advisor on the design of the *No Child Left Behind* Act, my experience designing, facilitating, and evaluating test development, standard setting and validation activities, my experience with assessment consortia including the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Career (PARCC), and professional testing guidelines, such as those found in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). It is my overall conclusion, to a reasonable degree of professional certainty, that it is inappropriate to use the percentages of students achieving any specified performance level on mandated state achievement testing programs as a sole or major indicator of the adequacy of a state's educational system.

Central to my overall conclusion regarding the impropriety of using percentages of students in a performance level as a measure of educational adequacy is the fact that large scale assessments such as those administered in Delaware and other states in compliance with federal accountability regulations are developed according to psychometric procedures that are congruent with the purpose of those tests. A common purpose of educational testing is the measurement of student achievement. There are many ways in which the measurement of student learning occurs, such as classroom assessments, teacher questioning of students, homework assignments, and others. One prominent way in which student achievement is measured comprises large-scale assessments such as those comprising the Delaware State Testing Program (DSTP), the Delaware Comprehensive Assessment System (DCAS), and the Smarter Balanced (SB) assessments. It is important to recognize that these assessments are

intentionally developed as measures of student achievement. They are *not* designed as measures of school quality, instructional quality, or system adequacy. Importantly, they are also validated for the purpose for which they were designed. That is, they have been evaluated and judged according to professional psychometric standards to be valid for obtaining high quality, dependable information about student achievement. They have not been validated for additional purposes.

Among several supporting conclusions described in the following sections are:

1) All tests are designed for a specific purpose. The assessments used in Delaware were designed and validated as measures of student achievement. They were not designed or validated as measures of instructional quality, educator effectiveness, or educational system adequacy. It is inappropriate to use test data for unintended and unvalidated purposes.

2) Educational reforms initiated in the 1970s have focused on increasing the rigor of content and performance expectations in American schools. Key aspects of these reforms have been the introduction of mandated assessments and accountability systems intended to spur increases in student achievement.

3) States have exhibited wide latitude in adopting content and performance expectations that reflect greater or lesser degrees of rigor. However, a common, primary policy goal of adopting challenging content and performance expectations has been to incentivize progress and exhort those affected to higher levels of educational achievement.

4) Over three generations of standards and assessments (DSTP, DCAS, and Smarter Balanced), the standards and assessments adopted by Delaware reflect a history of increasing expectations

for students and Delaware schools.

5) Because of variation *across* states in the rigor of their content and performance expectations, and because of variation *within* states when content and performance expectations are changed over time, it is necessary to refer to stable, external benchmarks to gauge educational progress. One such external benchmark is the National Assessment of Educational Progress (NAEP) which provides a common "yardstick" for such comparisons.

6) Since the 1990s, Delaware has made steady educational progress in reading and mathematics on the NAEP, demonstrating progress not only for students overall, but for at-risk/ disadvantaged students, in many cases ranking first nationally in gap reduction.

7) Into the future, Delaware has established continuing, challenging achievement expectations for all students. As part of its goals required by the federal Every Student Succeeds Act, Delaware has established achievement targets for 2030 in grades 3-8 of 76.1% proficiency in English language arts and 70.3% proficiency in mathematics for general population students. These targets represent increases of 23.96% and 29.76%, respectively. Long term goals for students with special needs have been set as 56.7% and 55.2% proficient in ELA and mathematics, respectively (increases of 43.3% and 44.8%); the proficiency targets for English language learners in ELA and mathematics are 57.6% and 59.1%, respectively (representing increases of 42.4% and 41%, respectively).

In the following sections, I provide background information on content and performance standards, including how performance standards are established; I provide background on recent federal educational reform initiatives involving assessment and accountability systems; I describe the

relationship between performance standards on assessments and accountability systems; I state several conclusions related to educational assessment and accountability in Delaware; and I provide the evidence and reasoning that formed the basis for those opinions.

V. MATERIALS REVIEWED AND BACKGROUND

In connection with my work in this matter, I reviewed relevant background, including court filings and records (i.e., second amended complaint; state's answer; and court's order on the motion to dismiss). I also reviewed reports and data from Delaware state testing programs and I conducted various analyses based on data from Delaware and other states obtained from the National Assessment of Educational Progress. Tables and figures using these data are included in the body of the report and are provided individually in Appendix B. The specific materials I relied upon are cited in the body of this report or the references listed in Section VIII. Finally, I relied on my experience in the area of large-scale student achievement testing, standard setting, and U.S. educational testing policy.

My opinions in this matter may be supplemented and are subject to amendment to the extent that additional information or material becomes available.

VI. BACKGROUND, ANALYSES, AND CONCLUSIONS

A. Background on Recent Federal K-12 Education and Testing Policy

A1. Background on the "Standards Movement" in the U.S.

Although testing has been a feature of American public schooling since its inception, federal requirements for statewide, every-pupil, high-stakes accountability testing are a relatively recent phenomenon. What has come to be referred to as the "standards movement" has its roots in a study requested by then-Secretary of Education Terrel Bell who formed the National Commission on Excellence in Education (NCEE). The NCEE final report expressed grave concern about the state of education in the U.S. The urgency of the need for reform was highlighted in the report's famous assertion that "If an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war" (1983, p. 9).

The alarm prompted a number of responses, reports, and legislation. Among them, the National Governors Association report, *Time for Results* (1986), the *Goals 2000: Educate America Act* (1994), and the *No Child Left Behind Act* [NCLB, (2001)]. Strong themes of the education reform efforts of the late 1980s and 1990s included the need to increase the global competitiveness of American students, to raise achievement expectations for all students, to promote consistency of those high expectations within and across the states, and to develop mechanisms for better monitoring and reporting of achievement progress.

As one initiative to accomplish these goals, NCLB required (among other things) that each state:

1) adopt rigorous *content standards* in English language arts, mathematics, and science;

2) develop and administer annual summative tests measuring those content standards to all students in Grades 3-8 and high school (for ELA and Mathematics; once each in elementary school, middle school, and high school for Science);

3) establish *performance standards* on the assessments for one level called "Proficient" and for at least one level above Proficient and one level below Proficient; and

4) develop plans for, and demonstrate "adequate yearly progress" (AYP) toward the goal of 100% of its students reaching the Proficient level of performance by the year 2014.

The most recent iteration of the standards movement has focused on increasing the rigor of required grade-level content in English language arts and mathematics. This emphasis began in 2009 when three organizations (the Council of Chief State School Officers, the National Governors Association, and Achieve, a non-profit, educational reform advocacy group) collaborated to produce sets of more challenging learning expectations known as the Common Core State Standards (CCSS). Subsequently, more rigorous assessments aligned to those standards were also developed. Extending the themes of earlier education reform efforts, the CCSS and related assessments have been motivated by intentions to incorporate more rigorous content in core subjects, raise achievement expectations for all students, and increase economic competitiveness.

(Additional information on content standards and performance standards, including the CCSS is provided in Sections B1 and B2 of this report.)

CONCLUSIONS: The standards movement, which began nearly 40 years ago, has been characterized by legislative and policy initiatives designed to increase the rigor of content requirements and performance expectations for students in American public schools.

A2. Background on Accountability Systems

A second primary feature of the standards movement has been the incorporation of accountability systems that mandate specific student achievement goals and reporting requirements for states. The assessments mandated by NCLB were instituted as a means for holding states accountable

for the use of federal education funding and, more importantly, for improving achievement results. NCLB assessments came to be known as *high stakes assessments* because of the accountability consequences, also a part of the legislation, that were associated with failure to demonstrate AYP.

In response to NCLB, all states (including Delaware) adopted ambitious programs and instituted various reforms to meet the requirements of NCLB and implemented plans to reach the goal of 100% proficiency.

However, although perhaps admirable, the goal of 100% students achieving a Proficient level of performance was soon recognized to be unreasonable. Within a few years of the passage of the NCLB legislation, educational researcher Robert Linn observed that "there is a zero percent chance that we will ever reach a 100 percent target" (quoted in Paley, 2007). As 2014 approached, no state had reached the goal of 100% proficiency, and many states failed to reach their AYP targets toward that goal. Such states faced NCLB-mandated sanctions for failing to achieve their AYP accountability targets, including loss of federal funding. As a result, in 2012, President Obama announced that the U.S. Department of Education would grant states waivers related to NCLB's 100% proficiency requirement and other aspects of the law (The White House, 2012).

CONCLUSIONS: In addition to increased content and performance expectations for students, the standards movement has also been marked by the introduction of accountability requirements for states, districts, and schools, and educators. The accountability requirements have also been intended primarily to spur increases in student achievement. Problems associated with some aspects of accountability systems have required changes to those systems.

A3. Background on the Purposes of Tests and Test Validation

Large-scale educational achievement testing has been a feature of the standards movement;

however, it has been ubiquitous in the United States prior to that movement and continues today. K-12 educational systems routinely administer state- or federally mandated tests to all students in a state in several grades and content areas. It is essential to understand the principle that such tests are designed to serve a specific purpose.

Whereas tests may be *used* for multiple purposes, they are *designed* with a primary purpose in mind. Any test designed with a primary purpose in mind typically serves that purpose more effectively than it does other, ancillary purposes.

Again, the common and primary purpose of all achievement testing (e.g., classroom quizzes, teacher questioning, end-of-semester examinations, etc.) is to measure student learning. As achievement measures, mandated statewide testing programs share this primary purpose of measuring student learning. Technical documentation for statewide testing programs such as the Delaware Student Testing Program (DSTP), the Delaware Comprehensive Assessment System (DCAS), the Smarter Balanced Assessment Consortium (SBAC) tests, or other large-scale assessment programs provides evidence that this primary purpose is accomplished well.

Admittedly, achievement tests may also be pressed into use for purposes for which they were not intended. For example, in K-12 education contexts, large-scale tests intended to be used primarily for measuring student achievement are sometimes used for purposes of personnel evaluations, compensation decisions, or in evaluations of instructional or educational system quality. The set of best practices for the field of testing embodied by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) caution against unintended and unvalidated uses when a test was not designed for such use or when sufficient information supporting the unintended or secondary use has not been gathered and analyzed. According to the *Standards*: “If validity for some common or likely interpretation for a given use has not been evaluated...potential users should be strongly cautioned about making unsupported interpretations” (p. 23) and “the improper use of tests...can cause considerable

harm to test takers and other parties affected by test-based decisions" (p. 1).

The statewide achievement tests administered in Delaware (e.g., DSTP, DCAS, SBAC) have been developed for the purpose of measuring student learning with respect to the grade level content standards they have been designed to assess. Technical data available for these tests indicates that they provide highly reliable and valid information about the learning goals they target. The state of Delaware makes available for public review information supporting the reliability of such tests and evidence supporting the validity of scores on the tests relevant to their intended purposes of measuring student learning.

The statewide achievement tests administered in Delaware were *not* developed to serve unintended purposes such as education personnel evaluation, evaluation of instructional quality, evaluation of the adequacy of the state's system of public schools, or other potential purposes. Before being used for such or other unintended purposes, the reliability and validity of data for those purposes should be gathered and evaluated.

CONCLUSIONS: Tests are designed for a specific purpose. Test data (e.g., student test scores) are most useful and defensible when a test is used for the purpose for which it is intended. Unintended uses of a test are possible; however, in such cases information justifying and supporting the additional uses must be gathered and evaluated.

B. Background on Standards in K-12 Educational Achievement Testing

Beyond a general understanding of the history of standards-based testing and accountability systems, the purpose(s) of testing, and the professional requirement for evidence supporting an intended use, it is also essential to understand the different ways in which the term *standards* is used in educational achievement testing and accountability systems. There are two primary usages. The first

usage is *content standards.*

B1. Content standards

Content standards refer to extensive compendia of statements specifying the knowledge, skills, or abilities that a state has established as learning targets for its students. Of note is that states typically only mandate the content that must be taught in given grades and subjects, not the instructional methods for teaching that content, nor any specific curriculum or sequence of instruction; these are typically issues of local control.

Content standards are organized into collections of fine-grained learning objectives for specific grades and subjects, specifying what should be taught at each grade and subject area covered by the standards. Examples of state-wide content standards applicable to general education students include:

* the *PA Academic Standards for Mathematics* (Pennsylvania Department of Education, 2014);
* the *New York State Next Generation Learning Standards* (New York State Education Department, 2017);
* the *Maryland Social Studies State Standards* (Maryland State Department of Education, 2008); and
* the *New Jersey Student Learning Standards for Comprehensive Health and Physical Education* (New Jersey Department of Education, 2014).

Relevant to the standards movement described in Section A1 is a recent and prominent set of content standards called the *Common Core State Standards* (CCSS) for English language arts and mathematics (NGA & CCSSO, 2010). The CCSS resulted from another federal educational policy initiative called “Race to the Top”, a competitive grant program that was a component of the *American*

Recovery and Reinvestment Act [ARRA (2009)]. Among other things, the ARRA continued the trajectory of the standards movement by encouraging states to experiment with innovative approaches to increasing the rigor of student performance expectations. The CCSS were adopted by a majority of states (including Delaware) and are widely regarded as representing more challenging expectations in terms of *what* students should be taught compared to then-extant content standards, and as requiring substantial changes in approaches for *how* the CCSS should be taught.

The ARRA also encouraged states to experiment with common assessments. Several groupings of states (called "consortia') were formed in response to this encouragement. States join consortia to not only share common sets of content standards, but also as potential means of cost-saving with respect to the development of assessments to measure the common content standards. Following several iterations, two consortia were established that are still in existence: the Partnership for Assessment of College and Career (PARCC) and the Smarter Balanced Assessment Consortium (SBAC or "Smarter Balanced"). In addition to adopting the CCSS, Delaware joined the Smarter Balanced consortium and began administering Smarter Balanced assessments in reading and mathematics in the 2014-2015 school year. The content standards covered by the Smarter Balanced assessments are detailed in the *Content Specifications for the Summative Assessment of the Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects* (Smarter Balanced Assessment Consortium, 2015b).

Most recently—though not resulting from a federal policy requirement—more challenging content standards for science, called the "Next Generation Science Standards" [NGSS; National Research Council (2012)] have been introduced and many states are currently engaged in efforts to develop assessments based on those content standards.

Table 2 provides examples of content standards for general student populations. Illustrated in the table are the kinds of specific, fine-grained learning objectives that comprise sets of content

standards. Examples from each of the state sources mentioned previously, as well as examples from the SBAC/Common Core standards and the NGSS are shown in Table 2.

Table 2
Examples of General Education Content Standards

State/Group	Grade	Subject	Specific Content Standard
Pennsylvania	8	Mathematics	CC.2.2.8.B.1: *Apply concepts of radicals and integer exponents to generate equivalent expressions.*
New York	4	English Language Arts	4R1: *Locate and refer to relevant details and evidence when explaining what a text says explicitly/implicitly and make logical inferences.*
Maryland	7	Social Studies/ Geography	7A1: *Use geographic tools to locate places and describe the human and physical characteristics in the contemporary world.*
New Jersey	2	Health	2.1.2.A.2: *Use correct terminology to identify body parts and explain how body parts work together to support wellness.*
SBAC	8	English Language Arts/Literacy	Target 14, Language Use: *Interpret understanding of figurative language, word relationships, nuances of words and phrases, or figures of speech (e.g., verbal irony, puns) used in context and the impact of those word choices on meaning.*
NGSS	5	Science	*5-PS1-2: Measure and graph quantities to provide evidence that regardless of the type of change that occurs when heating, cooling, or mixing substances, the total weight of matter is conserved.*

Sources: Pennsylvania Department of Education, (2014); New York State Education Department, (2017); Maryland State Department of Education, (2008); New Jersey Department of Education, 2014); Smarter Balanced Assessment Consortium, (2015a); Next Generation Science Standards, (2013)

Finally, in addition to the sets of state-level content standards listed above and the broader CCSS and NGSS content standards that have been developed primarily for students in general educational programs, content standards for populations of students with special needs have been developed. For example, states that participate in the World-class Instructional Design and Assessment (WIDA) consortium—of which Delaware is a founding member and participant—have adopted specialized content standards for English language learners (Board of Regents of the University of Wisconsin System, 2014). The states that participate in the Dynamic Learning Maps (DLM) consortium—which

Delaware joined for the 2017-2018 school year—have adopted specialized content standards for students with severe cognitive disabilities (University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems, 2019).

CONCLUSIONS: Content standards are sets of specific learning objectives that define the knowledge and skills students in a state are to be taught in given grades and subjects. States have adopted various sets of content standards, with many states (including Delaware) adopting increasingly rigorous content standards and assessments aligned to those standards.

B2. Performance standards

The second usage of the term standards is in reference to *performance standards*. In general, performance standards describe the judged levels of achievement on an assessment covering the content standards to which students should perform in order to be classified into one of several ordered categories. An example of a system of achievement levels is the categories of *Basic, Proficient, and Advanced*, where "Basic" is a label typically used to describe the level of performance of students with entry-level or beginning mastery of the set of content standards, and "Advanced" is a label commonly used to describe the performance of students with exceptional or very strong mastery of the content standards.

Alternatively, and as used in Delaware, a system of ordered achievement levels may be described using the labels, *Level 1, Level 2, Level 3,* and *Level 4.* These achievement level labels are ascribed with fixed meanings, called *achievement level descriptions* (ALDs) that describe required student performance at each level. ALDs may be written to different levels of specificity. At the highest level, ALDs can be written to generally describe achievement level categories applicable to any grade or subject; these ALDs are sometimes referred to as *policy ALDs*. Further refined ALDs can be

developed to describe differing degrees of achievement for specific grades and subjects. Table 3 provides the policy ALDs applicable to Delaware's 8th grade mathematics assessments. For comparison, the policy and grade/subject specific ALDs for mathematics used by the National Assessment of Educational Progress (NAEP) are also provided. (NAEP assessments are federally mandated tests, regularly administered to nationally representative samples of U.S. students at grades 4, 8, and 12.)

Table 3
Delaware and NAEP Achievement Level Descriptions, Grade 8 Mathematics

	Policy ALDs	Grade/Subject Specific ALDs
SBAC/Delaware Achievement Levels[1]		
Level 1 (*Does Not Meet/ Minimal)*	The Level 1 student demonstrates minimal understanding of and ability to apply the English language arts and literacy (mathematics) knowledge and skills needed for success in college and career, as specified in the Common Core State Standards.	
Level 2 (*Nearly Meets/ Partial*)	The Level 2 student demonstrates partial understanding of and ability to apply the English language arts and literacy (mathematics) knowledge and skills needed for success in college and career, as specified in the Common Core State Standards.	
Level 3 (*Meets/ Adequate*)	The Level 3 student demonstrates adequate understanding of and ability to apply the English language arts and literacy (mathematics) knowledge and skills needed for success in college and career, as specified in the Common Core State Standards.	
Level 4 (*Exceeds/ Thorough*)	The Level 4 student demonstrates thorough understanding of and ability to apply the English language arts and literacy (mathematics) knowledge and skills needed for success in college and career, as specified in the Common Core State Standards.	
NAEP Achievement Levels		
NAEP Basic	This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for performance at the NAEP Proficient level.	Eighth-grade students performing at the Basic level should exhibit evidence of conceptual and procedural understanding in the five NAEP content areas. This level of performance signifies an understanding of arithmetic operations—including estimation—on whole numbers, decimals, fractions, and percents.
NAEP Proficient	This level represents solid academic performance for each NAEP assessment. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.	Eighth-grade students performing at the Proficient level should apply mathematical concepts and procedures consistently to complex problems in the five NAEP content areas.
NAEP Advanced	This level signifies superior performance beyond NAEP Proficient.	Eighth-grade students performing at the Advanced level should be able to reach beyond the recognition, identification, and application of mathematical rules in order to generalize and synthesize concepts and principles in the five NAEP content areas.

Sources: Smarter Balanced Assessment Consortium (2013)/Delaware Department of Education (2019, September), National Assessment Governing Board (2018).

[1] Original SBAC Policy ALDs were phrased in slightly different terms. In addition to allowing member states flexibility in naming their performance levels, some states also adopted slight revisions of the Policy ALDs.

Operationally, performance standards refer to the cut scores applied to test results, where cut scores define the boundaries of the achievement levels. The number of points a student earns on a test covering the content standards (called a *raw score*) is transformed mathematically into a *scaled score*. Because achievement levels capture a range of performance, not all students classified into an achievement level have the same level of knowledge and skill. Thus, scaled scores are used to aid in distinguishing performance within an achievement level, to allow finer distinctions between students, and to aid in monitoring achievement gains (or losses) for groups of students. (Scaled score transformations are also done to avoid confusion across years of a state's testing program.) In Delaware, scaled scores in English language arts and mathematics range from 2000 (lowest obtainable scaled score) to 3000 (highest obtainable scaled score). The performance standard defining each of the four achievement levels used in Delaware is the lowest score that must be earned to be classified into an achievement level. Table 4 provides the ranges of achievement levels for Grades 3-8 mathematics tests administered in Delaware in 2017, with the performance standards (i.e., cut scores) for each level indicated in bold.

Finally, it is important to recognize that students classified within a performance level do not all have the same level of knowledge or skill. Figure 1, taken from a hypothetical Smarter Balanced score report for parents, illustrates the ranges of performance in each performance level. The figure shows a student's performance in the lower third of the Level 3 range of performance.

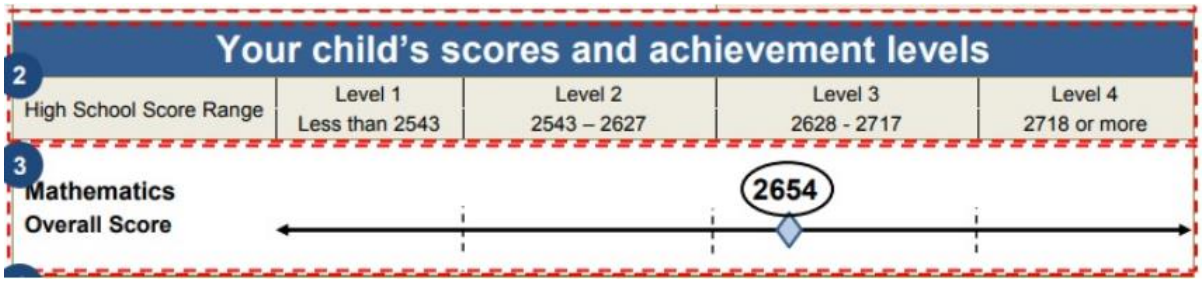
These ranges of performance can be seen in the performance bands shown in Table 4. In this case, the performance bands indicated are the actual scaled score performance ranges used in Delaware for Smarter Balanced assessments. For example, Level 3 performance in 3rd grade ranges from 2436 to 2500, with students at the lower end of the scaled score range (2436) barely mastering the required mathematics content to be classified at that performance level, whereas students scoring at the higher end of the range (2500) nearly mastered the grade 3 mathematics content well enough to be classified

at Level 4.

Table 4
Delaware Performance Standards, Grades 3-8 Mathematics

Grade	*Level 1*	*Level 2*	*Level 3*	*Level 4*
3	2380 and below	**2381** - 2435	**2436** - 2500	**2501** and above
4	2410 and below	**2411** - 2484	**2485** - 2548	**2549** and above
5	2454 and below	**2455** - 2527	**2528** - 2578	**2579** and above
6	2472 and below	**2473** - 2551	**2552** - 2609	**2610** and above
7	2483 and below	**2484** - 2566	**2567** - 2634	**2635** and above
8	2503 and below	**2504** - 2585	**2586** - 2652	**2653** and above

Source: Delaware Department of Education (2019, September)



Source: Lumos Learning *Understanding the SBAC Report Card.* Retrieved 1/9/2020 from https://www lumoslearning com/llwp/teachers-speak/understanding-the-sbac-report-card html

Figure 1.
Hypothetical Illustration of Range of Performance across Performance Levels

CONCLUSIONS: Whereas content standards specify the "what" regarding the content of instruction, performance standards specify "how well" students must perform in order to be classified into an achievement level. Performance standards can be established to reflect greater or lesser degrees of rigor. Performance standards create achievement levels (categories). Because achievement levels capture a range of performance and not all students classified into an achievement level have the same level of knowledge and skill, scaled scores are typically used in conjunction with performance levels to facilitate finer-grained reporting of student achievement and progress.

C. How and Why Performance Standards Are Set and Revised

Professional procedures have been developed to derive the performance standards (i.e., cut scores) used in educational assessment and accountability systems. Once initially set, performance standards may require periodic revision. Although one goal of setting performance standards is to delineate the achievement level categories used by a state, that is purely a technical process that is instrumental to more overarching goals related to educational achievement. The following subsections address three topics: 1) a brief review of the procedures for deriving performance standards; 2) a summary of the educational purposes of establishing performance standards; and 3) a description of the processes used by Delaware for establishing performance standards in English language arts and mathematics.

C1. Psychometric methods for setting performance standards

Researchers in the field of psychometrics have developed a variety of methods for deriving performance standards. Extensive treatment of available methods is provided in Cizek (2001, 2012) and Cizek and Bunch (2007). Fundamentally, all of the contemporary methods for setting performance standards involve the use of human judgment to identify the level of knowledge, skill, or ability students should be required to exhibit on an assessment covering a state's content standards in order to be classified as reaching one of the state's achievement levels.

The judgments are typically provided by qualified experts relevant to the grade and subject area to be tested. For example, the persons involved in setting the cut scores to demarcate achievement at the *Basic*, *Proficient,* and *Advanced* levels for a 6th grade mathematics test would be (predominately) 6th grade mathematics teachers and middle school mathematics curriculum specialists, with the possible addition of qualified persons to reflect the perspectives of, perhaps, middle school

administrators, parents, or teachers of mathematics at adjacent grades (i.e., 5th, 7th).

Although the specific procedures used by the qualified experts may differ, it is common for a large and representative group of such persons to be empaneled. The group is led through a process that includes: becoming familiar with the content standards and the assessment for which they will be setting standards; training in the specific standard setting method that will be used; reviewing and rating the items or tasks comprising the test for which performance standards are needed; considering supplemental sources of information such as historic pass rates, examinee performance, and relevant external data; and engaging in discussions and reconsideration of their judgments. Most commonly, the mean or median of the individual judgments is taken as the performance standard.

CONCLUSION: Performance standards result from the use of systematic procedures to collect informed, qualified judgments about the level of achievement necessary for a student to be classified into a given achievement level.

C2. Approval of performance standards

In preceding sections of this report, a process was described whereby qualified reviewers participate in standard setting activities. It is somewhat inaccurate, however, to state that the participants in such activities actually "set" the performance standards. Rather, it is more accurate to say that the participants *recommend* one or more performance standards. That is, in nearly all large-scale K-12 testing contexts, the participants' final mean (or median) judgment is provided to a policy making entity that alone has the authority to establish the performance standards. That entity may be a state board of education, a state chief school officer, or other policy making body. The authorized entity may accept the panelists' recommendations, adopt an adjusted recommendation, or reject the recommendation.

CONCLUSION: Although panels of qualified content specialists are typically used in standard setting procedures, those panels only recommend performance standards for assessments. Their recommended performance standards are forwarded to an entity (e.g., a state's commissioner of education, a state board of education) with the authority to actually set the standards.

C3. Establishment and use of performance standards in Delaware

The process followed for setting performance standards in Delaware followed closely the procedures just described. For its statewide assessments in English language arts (ELA) and mathematics, Delaware opted to join the Smarter Balanced Assessment Consortium (SBAC). In response to a federal initiative, SBAC sought and won funding to produce assessments based on the Common Core State Standards ("Common Core"). Delaware is one of 15 members of the Smarter Balanced consortium[2] that have adopted the Common Core content standards and which administer Smarter Balanced assessments aligned to those content standards in ELA and mathematics in grades 3-8. Delaware has adopted the SAT as its accountability measure for Grade 11.

Activities for setting the performance standards for Smarter Balanced assessments used in Delaware were conducted in 2014.[3] SBAC members decided to set four performance levels for Smarter Balanced assessments: Level 1 (Standard Not Met/Minimal Understanding), Level 2 (Standard Nearly Met/Partial Understanding), Level 3 (Standard Met/Adequate Understanding), and Level 4 (Standard Exceeded/Thorough Understanding). A four-phase process was conducted (see Smarter Balanced Assessment Consortium, 2015c). Phase I of the standard setting process comprised an online "crowd

[2] Current members of the Smarter Balanced consortium include California, Connecticut, Delaware, Hawaii, Idaho, Indiana, Michigan, Montana, Nevada, Oregon, South Dakota, Vermont, Washington, the U.S. Virgin Islands, and the U.S. Bureau of Indian Education.

[3] Since 2010, the author of this report has served as a member of the Smarter Balanced technical advisory committee and, in 2012-2014, assisted in the design and evaluation for setting achievement levels on the Smarter Balanced assessments.

sourcing" activity to gain preliminary opinion from a wide range of stakeholders regarding where the performance standard for Level 3 should be set. A total of 2,660 online responses were received from teachers, administrators, higher education representatives, and others.

Overall recommendations from Phase I were used as a source of external information for an in-person standard setting activity (Phase II) held in October 2014. The process used for Phase II followed the typical procedures described in Section B1 above. Qualified participants in the Phase II activities consisted of representatives of each member state in the SBAC consortium. A total of 16 panels were formed, with each panel focusing on a specific grade level (grades 3-8, high school) and a specific subject area (ELA, mathematics). Participants were assigned to panels based on their expertise at the relevant grade level and subject areas, with approximately 30 participants per panel. Panelists recommended performance standards to define each of the performance levels (i.e., Levels 1, 2, 3, and 4) for each grade and subject. Phase III of the process occurred at the conclusion of the individual panels' work. At that point, cross-grade panels were formed to review the recommended performance standards across grade levels for coherence. In Phase IV of the process, a final set of recommended performance standards was then referred to the appropriate policy making body (i.e., the group of chief state school officers from Smarter Balanced member states) for review and approval.

The initial set of performance standards resulting from the panelists' work was not accepted by the chief state school officers of SBAC member states. Smarter Balanced staff received input from the chief state school officers and produced modest revisions to the Phase III performance standards; ELA performance standards were developed that were slightly higher than the panelists' original recommendations; mathematics performance standards were developed that were slightly lower than the panelists' original recommendations. The SBAC staff revisions took into greater account the relevant NAEP performance standards in reading and mathematics and brought the ELA and mathematics results into closer alignment with each other. On November 14, 2015 the group of chief

state school officers voted unanimously (with two abstentions) to approve the final recommended performance standards. Like other SBAC members, Delaware has implemented the approved SBAC performance standards for all relevant grades and subjects since 2015.

CONCLUSION: The performance levels for Delaware state assessments have been established using contemporary best practices for standard setting common to state and national student achievement testing programs.

C4. Educational purposes of performance standards

All of the technical and procedural aspects of setting performance standards can sometimes distract from why performance standards are set in the first place. It would of course be possible to develop content standards describing what students should be taught in specified grades and subjects and to develop aligned assessments to measure acquisition of those content standards without establishing performance standards at all. However, the setting of performance standards serves at least two primary functions.

C4a. The exhortatory purpose of performance standards

The first primary function of setting performance standards was described by Linn (1995), who outlined four historical purposes. Foremost among those purposes is what Linn termed *exhortation*. According to Linn: "A major use of performance standards is to exhort educators, students, and the public to exert greater, and possibly different, kinds of effort to achieve established standards of performance" (p. 368). Performance standards established for this purpose are intended to stimulate actions on the part of educational organizations, personnel, and students to reach higher, aspirational achievement goals. That is, performance standards set for large-scale educational achievement tests are

not intended to merely reflect a *status quo* or to reflect current, normative levels of performance, but to represent performance goals that require stretch and effort to achieve. Setting performance standards too low—for example, setting performance standards that would be easily reached by 100% of students—would fail to inspire the intended effort and innovation.

Indeed, over the course of U.S. educational policy making, performance standards have been established that were recognized at the outset as levers for stimulating educational reforms. The performance standards adopted as part of the *No Child Left Behind Act* (2001) are perhaps the most familiar example of exhortatory performance standards. That legislation established performance expectations of 100% of students achieving the *Proficient* level or higher on mandated statewide assessments by 2014—a ten-year timeline from the first administrations of NCLB-mandated testing.

This 10-year timeline for accomplishing NCLB goals illustrates an additional characteristic of exhortatory standards: it is typically the case that such standards are intended to be *long range goals* as opposed to short term objectives. In recognition of the complex and multifaceted nature of educational systems, they are intended to drive incremental performance improvement over time. The theory of action involves several critical elements, beginning with a state's formal adoption of a new set of content standards, which then must be followed by professional development for educators in the new standards, dissemination to other key stakeholders (e.g., parents), instructional alignment and changes in classroom practices, adoption of new instructional materials, and development and administration of assessments designed to measure the new learning targets. Improvements in student performance on the new assessments characteristically—and, typically, slowly because of the complicated nature of the educational systems—follow from the implementation of each of the elements over time. Thus, the appropriate use of exhortatory standards is their use as indicators for monitoring improvement in achievement (and implementation of standards) over time.

In describing the use of performance standards for exhortation, Linn and colleagues observed

that the process of setting performance standards is typically *not* necessarily intended to result in standards that are achievable by all students. According to Linn, Baker and Betebenner:

> Panels of teachers—either alone or with other interested citizens—have generally set performance standards....The result of this judgmental standard setting process frequently has been to set the proficient level so high that it may be unrealistic to expect all students to reach that level by 2014. Certainly, no state, or country for that matter, is close to meeting the high standard set for proficient performance on NAEP or similar standards on many state assessments. (p. 4)

The performance standards for Smarter Balanced assessments were established with the same perspective of setting rigorous standards for the purpose of simulating improvement and with the explicit recognition that, at least initially, the percentages of students meeting the standards would be lower than in the past. As part of their work, Smarter Balanced panelists were presented with samples of typical student work, percentages of students answering the test questions correctly and other current performance data. Then, rather than merely setting performance levels that reflected current performance, the panelists relied on the Smarter Balanced policy achievement level descriptions. Those ALDs set forth ambitious performance goals more aligned with those of the National Assessment of Educational Progress and evolving conceptions of the levels of knowledge and skill that would be required for college and career readiness into the next century.

Importantly, panelists were presented with and explicitly considered the anticipated percentages of students that would meet the Smarter Balanced performance standards they were recommending. Figures 2 and 3 show the estimated percentages of students expected to score at or above Levels 2, 3, and 4 on the Smarter Balanced assessments in English language arts/literacy (ELA/L) and mathematics, respectively. (The data on which the figures are based is provided in Appendix A-2.)

On Smarter Balanced assessments, students scoring at performance Levels 3 and 4 are considered to be likely on-track to acquiring the knowledge and skills necessary for college and career readiness. Figure 2 shows that, in ELA/L, the educators who recommended the Smarter Balanced

performance standards recognized that only 38-41% of students would be classified as Level 3 or higher, with only around 8-18% of students expected to achieve Level 4 or higher. In mathematics, it was expected that only 32-39% of students would achieve Level 3 or higher, with 11-15% achieving Level 4 or higher. That is, the panelists were explicitly recommending exhortatory performance standards, recognizing that in all cases, the majority of students would not be expected to achieve Level 3 (i.e., Proficient) or higher when the Smarter Balanced performance standards were adopted.

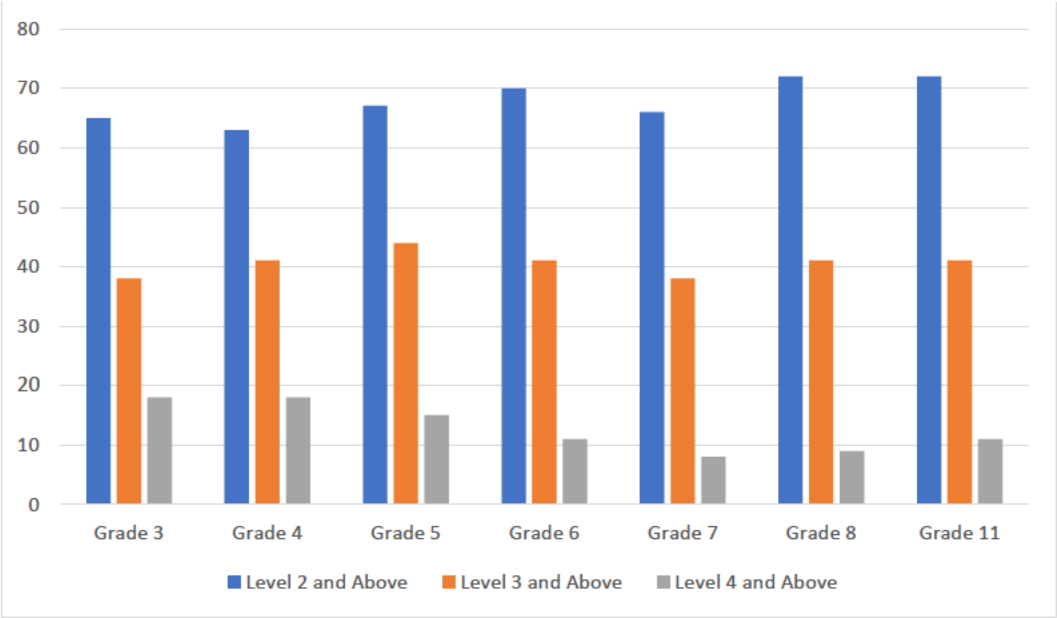


Figure 2.
Estimated Smarter Balanced ELA/L Performance by Grade, 2015

Source: Touchette, B M (2014, December 18) *Smarter achievement level setting: Presentation to the Delaware State Board of Education* Dover, DE

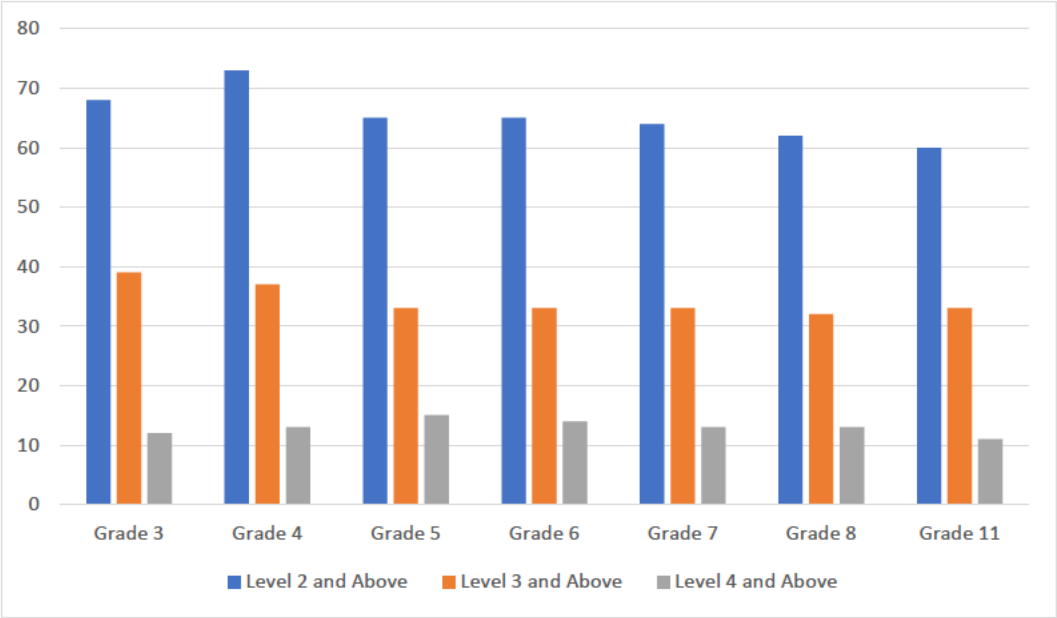


Figure 3.
Estimated Smarter Balanced Mathematics Performance by Grade, 2015

Source: Touchette, B M (2014, December 18) *Smarter achievement level setting: Presentation to the Delaware State Board of Education* Dover, DE

CONCLUSIONS: Performance standards on large-scale achievement tests such as the Smarter Balanced and NAEP assessments, are not set to reflect a status quo. Rather they are established to serve as exhortatory goals, as levers for promoting innovation, effort, and reform, and as mechanisms for monitoring progress. Such performance standards are not set with the belief that 100% of students will meet them. Rather, they are established at levels that represent "stretch" goals for educational systems, educational personnel, students, parents, and others. Ironically, if performance standards for such tests were set in such a way that they were quickly or easily reached by 100% of students, they would not only be meaningless, but would likely serve to undermine educational progress.

C4b. The monitoring uses of performance standards

A second primary function of performance standards is to serve as benchmarks for monitoring

educational progress. To the extent that performance standards remain constant over time, the achievement of all students in a state, as well as the achievement of subgroups of interest, can be tracked. The monitoring function served by performance standards can aid policy makers in directing educational resources, evaluating reforms, assessing success in reducing achievement gaps, and developing new initiatives.

This monitoring function of performance standards has been widely embraced by advocates for groups that have been historically underperforming. A strong statement advocating for the use of high-stakes assessments and performance standards was issued in 2015 by the Leadership Conference on Civil and Human Rights (LCCHR) on behalf of 12 civil rights organizations including: the American Association of University Women, the Disability Rights Education and Defense Fund, the League of United Latin American Citizens, the NAACP, the National Council of La Raza, the National Disability Rights Network, the National Urban League, and others. The statement reads in part:

> "Until federal law insisted that our children be included in these assessments, schools would try to sweep disparities under the rug by sending our children home or to another room while other students took the test. Hiding the achievement gaps meant that schools would not have to allocate time, effort, and resources to close them. Our communities had to fight for this simple right to be counted and we are standing by it." (LCCHR, 2015, p. 1)

The monitoring function alluded to by the LCCHR statement remains an explicit purpose of adopting performance standards on educational achievement tests. Many public interest groups, advocacy groups, policy makers, and the pubic rely on state educational "report cards" and accountability system data reporting for information to monitor educational progress of systems and subgroups.

CONCLUSIONS: Stable performance standards are routinely used to monitor progress toward educational achievement goals. Data resulting from longitudinal monitoring is necessary for

policy makers and the public to make informed educational decisions.

C5. Performance standards and educational accountability

In K-12 education policy, the monitoring function of educational performance standards has become closely intertwined with state and federal accountability requirements. Recall that in Section A1, the No Child Left Behind Act (2001) required states to 1) adopt rigorous content standards in English language arts, mathematics, and science; 2) develop and administer annual summative tests measuring those content standards; 3) establish performance standards (e.g., Basic, Proficient, Advanced); and 4) develop plans to ensure that 100% of students attain the Proficient level or above.

As discussed previously, the goal of 100% of U.S. students achieving the Proficient level of performance was almost immediately recognized as exhortatory, though likely unreachable. Within a few years of the passage of the NCLB legislation, educational researcher Robert Linn observed that "there is a zero percent chance that we will ever reach a 100 percent target" (quoted in Paley, 2007).

Although states adopted ambitious programs and instituted various reforms relevant to achieving it, no state accomplished the goal of 100% proficiency. In 2012, President Obama announced that the U.S. Department of Education would grant states waivers related to NCLB's 100% proficiency requirement and other aspects of the law (The White House, 2012). Despite the waivers, the aspirational nature of state level performance standards has remained via provisions of the *Every Student Succeeds Act* [ESSA (2015)] which retained a requirement that state education agencies (SEAs) "must provide baseline data (i.e., starting point data), measurements of interim progress, and long-term goals for academic achievement, graduation rates, and English language proficiency."

Importantly, although the ESSA legislation eliminated the NCLB requirement that 100% of students achieve Proficient performance, ESSA still required states to set long term achievement targets. In its ESSA plan, submitted to the U.S. Department of Education, Delaware committed to the

following long-term achievement goals for 2030 (see also Table 5):

> The DDOE is also establishing 'ambitious state-designed, long-term goals' with measures of interim progress for all students and subgroups of students.... The long term goals will increase achievement for all students from 52.09% to 76.05% for ELA and from 40.49% to 70.25% for mathematics an increase of 23.96% and 29.76% respectively. For Delaware's lowest performing subgroups, students with disabilities and English learners, the ELA proficiency goal is an increase of 43.26% and 42.43% respectively, and the mathematics proficiency goal is an increase of 44.82% and 40.95% respectively. (2019, p. 2)

Table 5
Delaware Long-term Proficiency Goals

	ELA	ELA	Mathematics	Mathematics
Subgroups	Starting Point (2015-2016)	Long-Term Goal (2030)	Starting Point (2015-2016)	Long-Term Goal (2030)
All students	52.09%	76.05%	40.49%	70.25%
Economically disadvantaged students*	35.60%	67.80%	25.42%	62.71%
Children with disabilities*	13.48%	56.74%	10.36%	55.18%
English learners	15.14%	57.57%	18.10%	59.05%
African American	36.19%	68.10%	23.39%	61.70%

Source: Delaware Department of Education (2019, June 10) *Delaware consolidated state plan under the Every Student Succeeds Act.* Dover, DE: Author

CONCLUSIONS: Performance standards can be set to reflect unrealistic goals, as was the experience with the NCLB legislation. When that is the case, they may no longer serve to stimulate effort and achievement. More recent ESSA legislation still requires achievement goals to be set to serve the exhortatory function of performance standards and to serve as accountability targets; however, the ESSA legislation permits states to establish their own long-term goals. Delaware has established such goals that commit it to achievement progress. For general population students in grades 3-8, Delaware aspires to reach approximately 76% Proficiency in English language arts and 70% Proficiency in mathematics by 2030. Similarly challenging long-term goals have been established for subpopulations such as students with

special needs (reaching approximately 57% and 55% Proficiency in ELA and mathematics, respectively), and English language learners (reaching approximately 58% and 59% Proficiency).

D. Relationships between Performance Standards and Proficiency Levels

D1. Variation in proficiency levels across states

There is dramatic variation in the percentages of students classified as "Proficient" across the states. This variation is due to three factors: 1) the rigor of the content standards adopted by each state; 2) how Proficient is defined in each state (i.e., it could be defined in a more- or less-rigorous way); and 3) where states set their cut scores for Proficient. (An additional element that relates to these factors is the level of difficulty of each state's tests.)

An illustration of the wide variation in how these factors combine to affect proficiency percentages across the states is provided by Linn, Baker, and Betebenner (2002) and reproduced in Figure 4. The figure shows the percentage of students classified as "Proficient" in reading in six selected states over a four-year period prior to the passage of NCLB. Although the figure was developed to illustrate trends, it dramatically shows the difference across states in any given year. For example, in 2001, the percentages range from approximately 25% of students classified as proficient in Maryland to greater than 90% of students classified as proficient in Texas. The extreme gulf between apparent achievement in, for example, Maryland and Texas, is almost surely due to the factors enumerated above (i.e., rigor of content standards, definition of proficiency used, rigor of state tests, and location of Proficient cut scores) and do not reflect real difference among students or the quality of their respective states' educational programs.

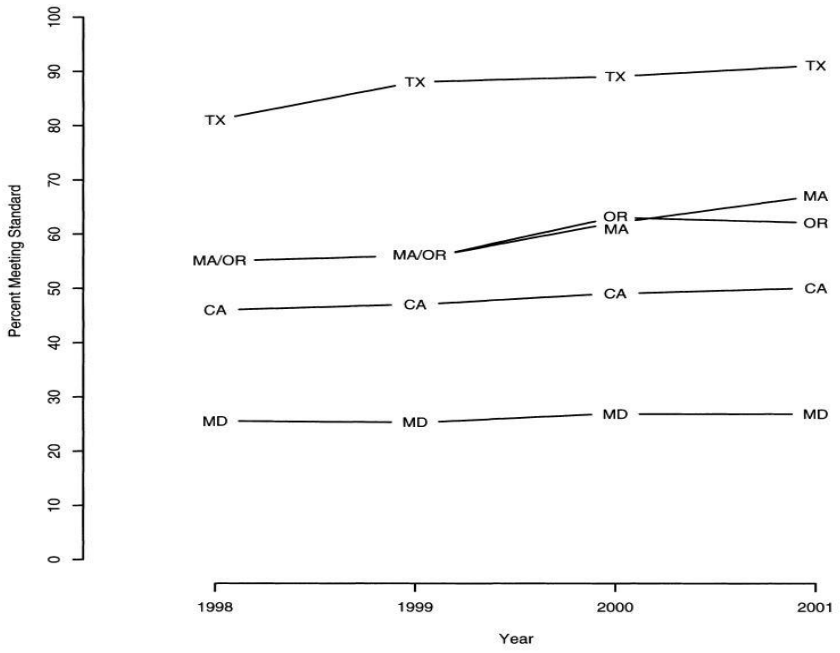


Figure 4.
Percentages Proficient in Reading in Six States, 1998-2001.

Source: Linn, Baker, & Betebenner (2002)

Linn et al's (2002) analysis was conducted based on state assessment program data *prior to* the passage of NCLB. In addition to measuring student achievement, various accountability mechanisms—including NCLB and ESSA—have been associated with their results. The linkage of accountability mechanisms and assessment results has affected the level at which performance standards are set.

For example, as indicated previously, the NCLB Act initially required that states achieve 100% of students scoring at the state's Proficient level. At least one reaction to that requirement was for states to establish differing definitions of "proficient" and each state established their own cut scores to define those differing levels. The federal mandate for all states to reach 100% of students being proficient created a dilemma for states and incentives that actually served to undermine the improvement of education. That is, it would be possible to meet the mandates of NCLB simply by establishing a sufficiently low level of performance to define "Proficient." Given the consequences of a state not

meeting its AYP goals, there was an understandable incentive to establish modest achievement goals—and clearly no incentive to set aggressive goals or raise standards. As described by Linn:

> All states had to submit plans explaining how they were going to meet the accountability requirements of NCLB by January 31, 2003. States that were in process of introducing new assessments or that had not yet set performance standards will be setting standards in quite a different context than existed prior to the enactment of NCLB. In light of the new context provided by NCLB, it reasonable to expect that they will set the standards at less ambitious levels than they would have been set a couple of years earlier. (2003, p. 5)

Across the U.S., this wide variation in how "Proficient" is defined has been demonstrated by research conducted by the National Center of Education Statistics (NCES); this work clearly shows that percentages of students labeled as "proficient" is an artifact of where states' performance levels are set.

NCES routinely conducts analyses using data from the National Assessment of Education Progress (NAEP). The NAEP program consists of periodic administration of assessments in grade 4 and 8 reading and mathematics (and other subjects) to samples of students that are representative of their states and the nation as a whole.

Through a statistical process called linking, the NAEP assessments can be used as a "common yardstick" to see how the proficiency levels on state assessments compare. These so-called *mapping studies* show how each state's performance standards in reading and mathematics compare to a single standard—the NAEP levels of *Basic*, *Proficient,* and *Advanced* (see, e.g., Bandeira de Mello, Rahman, Fox, & Ji, 2019). A sample of these biennial comparisons is provided in Figures 2-5. For example, Figures 5 and 6 show the NAEP-equivalent score of each state's 2017 *Proficient* performance level for grades 4 and 8 reading, respectively; Figures 7 and 8 show the locations of each state's 2017 *Proficient* performance levels for grades 4 and 8 mathematics. In each case, the red arrow superimposed on the graph shows the consortium of Smarter Balanced states (including Delaware), listed in alphabetical order.

For example, the results for grade 4 reading shown in Figure 5 illustrate that the definition of *Proficient* embodied by the Massachusetts (MA) state assessment program was the most rigorous in the country—at the level represented by the NAEP *Proficient* achievement level. At the other extreme, what was labeled as the "Proficient" level of grade 4 reading achievement in Texas (TX) represented the lowest level of expectations for student performance, falling into the *Below Basic* category with respect to NAEP. A collection of 12 member states in the Smarter Balanced consortium, all of whom adopted the same performance standards for that assessment, can be seen grouped near the middle of the NAEP Basic achievement level. Similar results are provided for grade 8 reading (Figure 6) with Kansas (KS) having the most rigorous performance standards, and Virginia (VA) having performance standards at the lowest level relative to NAEP.

In 4th grade mathematics (Figure 7), the performance standards adopted by members of the PARCC consortium mapped to the highest NAEP-equivalent score with Puerto Rico's (PR) performance standards mapping to the lowest level. At the 8th grade level (Figure 8), the performance standards adopted by Kansas (KS) corresponded to the highest NAEP-equivalent score and those adopted by Iowa (IA) mapped to the lowest NAEP-equivalent score.

A final point is important to note with respect to the rigor of state adopted performance standards; namely, having higher standards may result—at least initially—with smaller percentages of students meeting those standards. For example, on the 2017 NAEP 4th grade reading assessment, the state with the most rigorous NAEP-equivalent performance standard (Massachusetts) nonetheless had only 51% of its students score at or above the Proficient level and only 17% at or above the Advanced level. These results further support the conclusion that a state's proficiency percentage cannot be interpreted as an indicator of educational quality or adequacy.

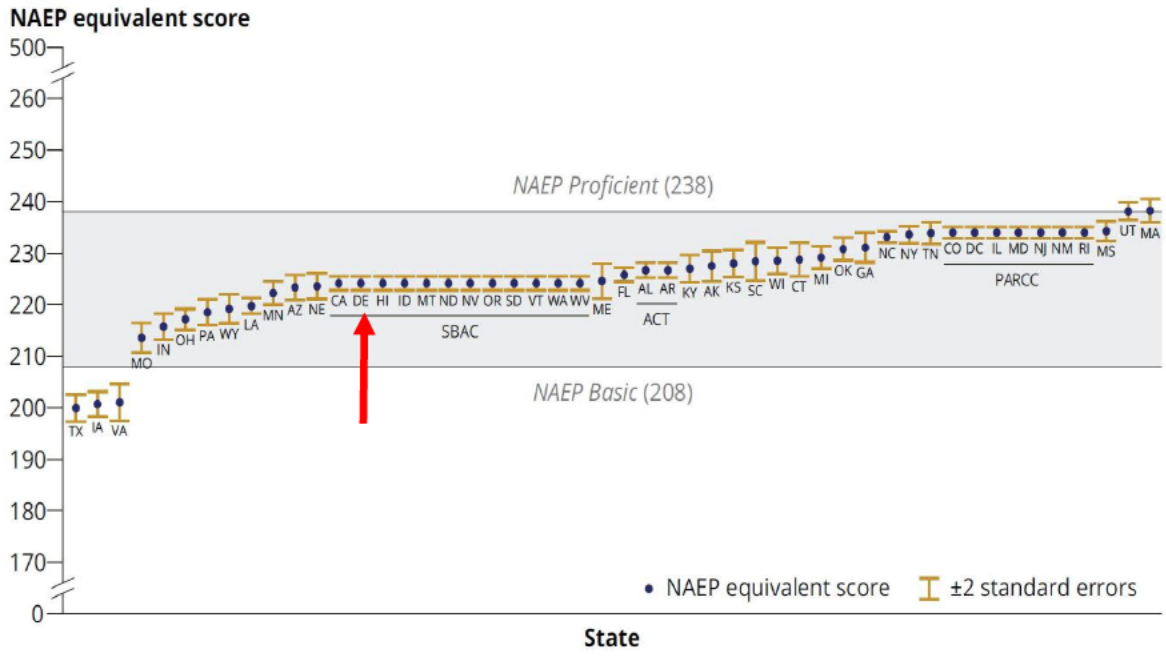


Figure 5.
NAEP Mapping Study Results, 2017, Grade 4 Reading
Source: Bandeira de Mello, V , Rahman, T , Fox, M A , & Ji, C S (2019) *Mapping state proficiency standards onto the NAEP scales: Results from the 2017 NAEP reading and mathematics assessments (NCES 2019-040)* U S Department of Education Washington, DC: Institute of Education Sciences, National Center for Education Statistics

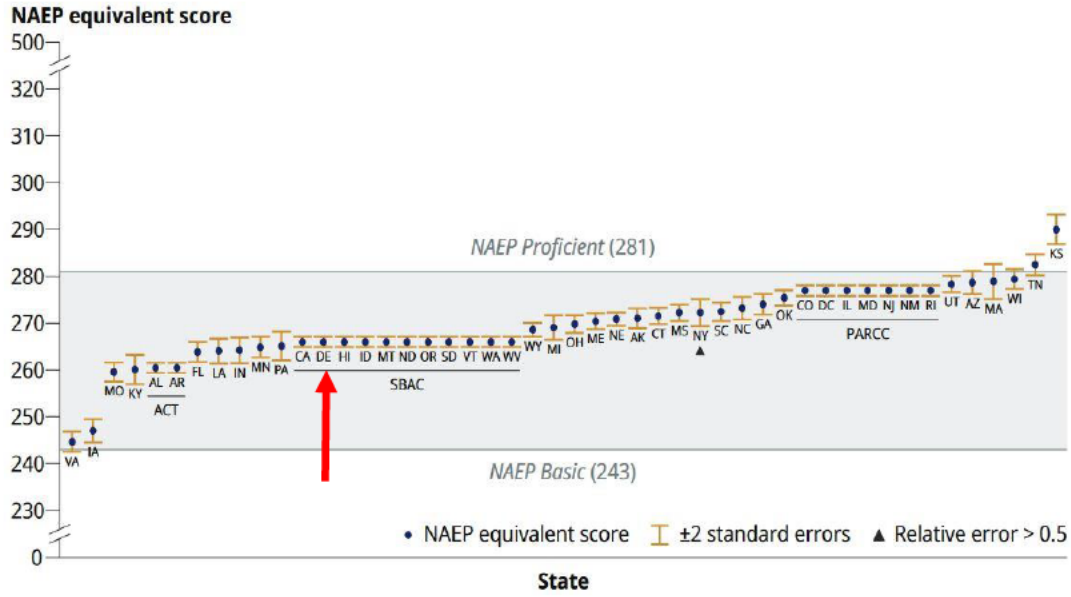


Figure 6.
NAEP Mapping Study Results, 2017, Grade 8 Reading
Source: Bandeira de Mello, V , Rahman, T , Fox, M A , & Ji, C S (2019) *Mapping state proficiency standards onto the NAEP scales: Results from the 2017 NAEP reading and mathematics assessments (NCES 2019-040)* U S Department of Education Washington, DC: Institute of Education Sciences, National Center for Education Statistics

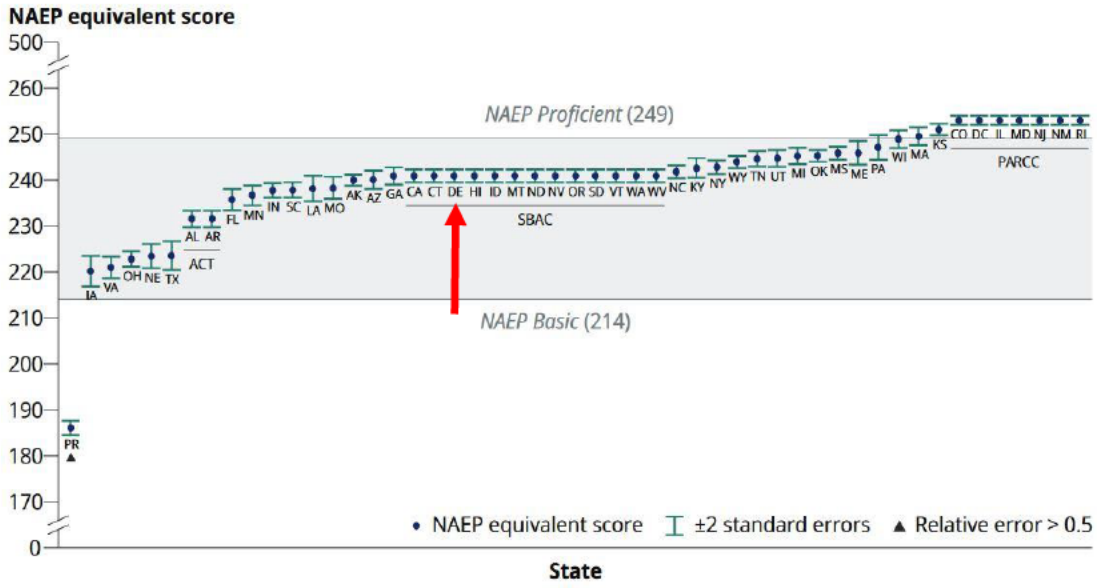


Figure 7.
NAEP Mapping Study Results, 2017, Grade 4 Mathematics
Source: Bandeira de Mello, V., Rahman, T., Fox, M.A., & Ji, C.S. (2019) *Mapping state proficiency standards onto the NAEP scales: Results from the 2017 NAEP reading and mathematics assessments (NCES 2019-040)* U.S. Department of Education. Washington, DC: Institute of Education Sciences, National Center for Education Statistics

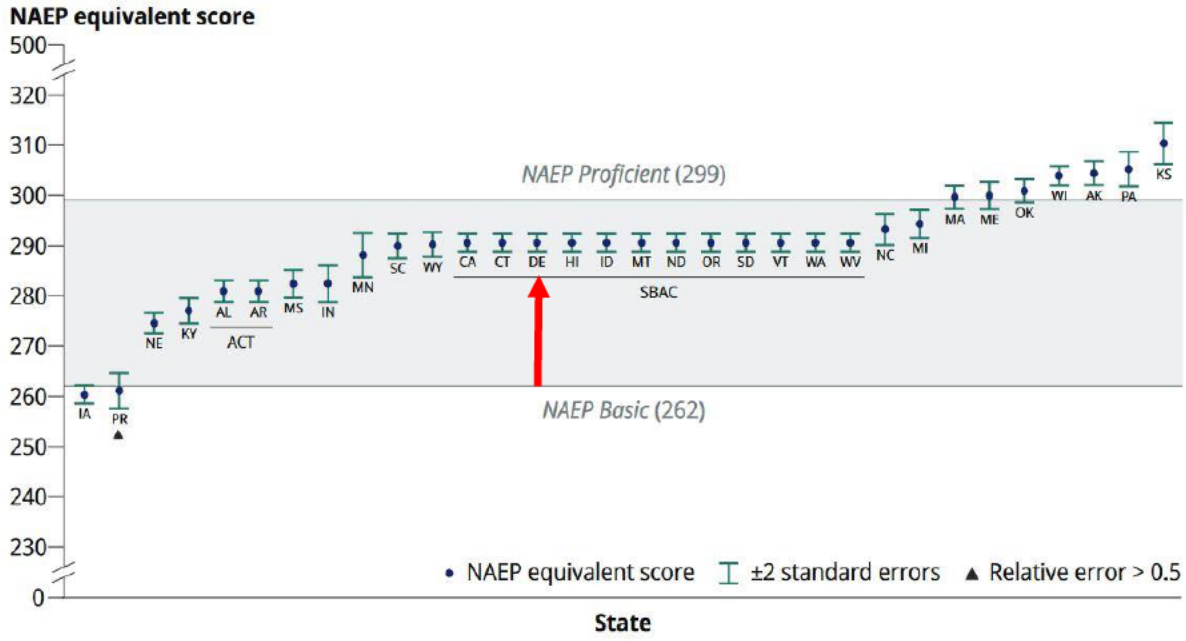


Figure 8.
NAEP Mapping Study Results, 2017, Grade 8 Mathematics
Source: Bandeira de Mello, V., Rahman, T., Fox, M.A., & Ji, C.S. (2019) *Mapping state proficiency standards onto the NAEP scales: Results from the 2017 NAEP reading and mathematics assessments (NCES 2019-040)* U.S. Department of Education. Washington, DC: Institute of Education Sciences, National Center for Education Statistics

The research on proficiency levels across the states can be summarized in two ways. First, "proficiency" is not an unambiguous or universal concept, but a "created" artifact that is dependent on three factors: 1) the *rigor of the content standards* adopted by each state; 2) the *definition of Proficient* adopted by a state (that is, the state's ALD for Proficient), and 3) the *level at which a state's Proficient cut score is set*. Because of these factors, states such as Massachusetts—although it typically has among the highest NAEP-equivalent Proficient performance standards—can often have comparatively smaller percentages of students achieving a "Proficient" level of performance because of demanding ALDs describing what it means to be proficient, and because cut scores set to establish Proficient levels of performance may be set comparatively higher than in other states. Conversely, states can have comparatively low NAEP-equivalent performance standards, but can show high percentages of student proficiency by adopting less demanding ALDs to describe proficient performance and/or by setting low cut scores for the *Proficient* achievement level.

Second, merely examining a state's percentage of students scoring at a *Proficient* (or any other achievement level) is inadequate for making any determinations about the rigor or adequacy of the state's educational programs. Again quoting Linn:

> The problem is that states set the student academic achievement standards and they have done so in ways that vary greatly in stringency. Some states have set very lenient standards such that 80% or more of their students perform at the proficient level or above on their assessments while other states have set such stringent standards that 20% or less of their students perform at that level. (2004, p. 4)

This phenomenon was detailed in a New York Times article a year later:

> Judged solely by recent statewide tests, fourth graders in Mississippi and Colorado would appear to be the best young readers in the nation. In both states, 87 percent of fourth graders passed their exams. But Mississippi came in dead last among the 50 states when fourth-grade reading was examined using a different standard, a newly mandated but decades-old test called the National Assessment of Educational Progress, or NAEP. On that test, only 18 percent of Mississippi's fourth graders achieved proficiency. (Saulny, 2005, p. 8)

CONCLUSIONS: States have demonstrated wide variability in setting performance standards

at higher or lower levels for denoting "Proficient" student performance. Lower (or higher) percentages of students scoring at the "Proficient" level in a state do not provide evidence of educational quality or adequacy of a state's educational system. Rather, percentages of students classified as "Proficient" must be interpreted in the context of other factors, such as the challenge represented by the content standards, the rigor of the performance standards, and the difficulty level of the state's tests.

D2. Variation in proficiency within states

Not only can the definitions of performance categories such as *Proficient* vary across states, but the definitions can—and typically do—vary *within* a state over time. The reason for this is that state assessment programs are constantly evolving: new sets of content standards are often adopted every 3-7 years; new achievement level descriptions may be created, representing higher (or lower) levels of performance to be classified into an achievement level; and new performance standards may be set, again representing potentially higher or lower expectations for performance. Indeed, the use of performance standards for the purpose of exhortation often follows a familiar, cyclic, chronology whereby:

1) a new set of content standards in one or more subject areas are adopted by a state;

2) challenging assessments are developed to align to the new content standards;

3) rigorous performance standards are set;

4) initially, and compared to previous years, percentages of students meeting the new performance standards are modest;

5) as experience with the new content standards grows, more students meet the new performance standards;

6) increases in percentages meeting the new performance standards are greater in the early years of adoption and begin to "plateau" in later years; and

7) content standards and/or performance standards are revised, with resulting performance

declines and the beginning of a new cycle.

The table presented in Appendix A-1 highlights this cycle. The table shows the percentages of Delaware students classified as Level 3 or above in English language arts/reading and mathematics across the school years 2009-2010 through 2018-2019. (Across these years, performance at Level 3 has generally been considered to be proficient performance.) The data show that proficiency rates were generally in the range of 70-80% for the first year shown (2009-2010; the last year of the DSTP assessment program), then fell in the range of 60% (2010-2011; the first year of the DCAS assessment program), the rose to 60-70% for the years 2011-2012 through 2013-2014 (the other years of the DCAS assessment program), before then falling into the 30-50% range for 2014-2015 through 2018-2019 school years (the first years of the Smarter Balanced assessment program).

This pattern of results is easily seen in Figure 9 which plots the percentages of students that achieved Level 3 or higher on Delaware reading tests in grades 4, 6, and 8. (Three grades and a single subject area were selected in order to illustrate the pattern. The same pattern holds for other grades in reading and for mathematics.)

Figure 9 clearly illustrates higher performance under the DSTP program in 2010, a sharp decline in performance with the introduction of new content standards, performance standards, and assessments of the DCAS program in 2011. A general rise in performance occurred in the 2011-2012 school year, followed by another sharp decline in performance in 2015 with the introduction of the more rigorous Smarter Balanced content standards and assessments. Did Delaware students become broadly less able in 2011, markedly more accomplished in 2012 and actually lose learning in 2015? No. The increases and declines shown in the figure are attributable to changes in content standards, performance standards, assessments, and policy changes that emphasized greater rigor in the state's testing program.

Indeed, increases and decreases in proficiency rates are completely predictable when the identified factors change. For example, when Delaware introduced the Smarter Balanced assessments,

it was publicly acknowledged that the more challenging Common Core content standards, and the more rigorous SBAC assessments and cut scores would produce decreases in proficiency rates. The following statements are drawn from a Delaware State Board of Education presentation:

> "Because the new standards set higher expectations for students--and the new tests are designed to assess student performance against these higher expectations--our definition of grade level performance is higher than it used to be.
>
> "As a result, it means that fewer students will meet grade level standards, especially for the first few years."
>
> "This does not mean that our students are 'doing worse' than they did last year. Rather, the scores represent a 'new baseline'." (2015, p. 27)

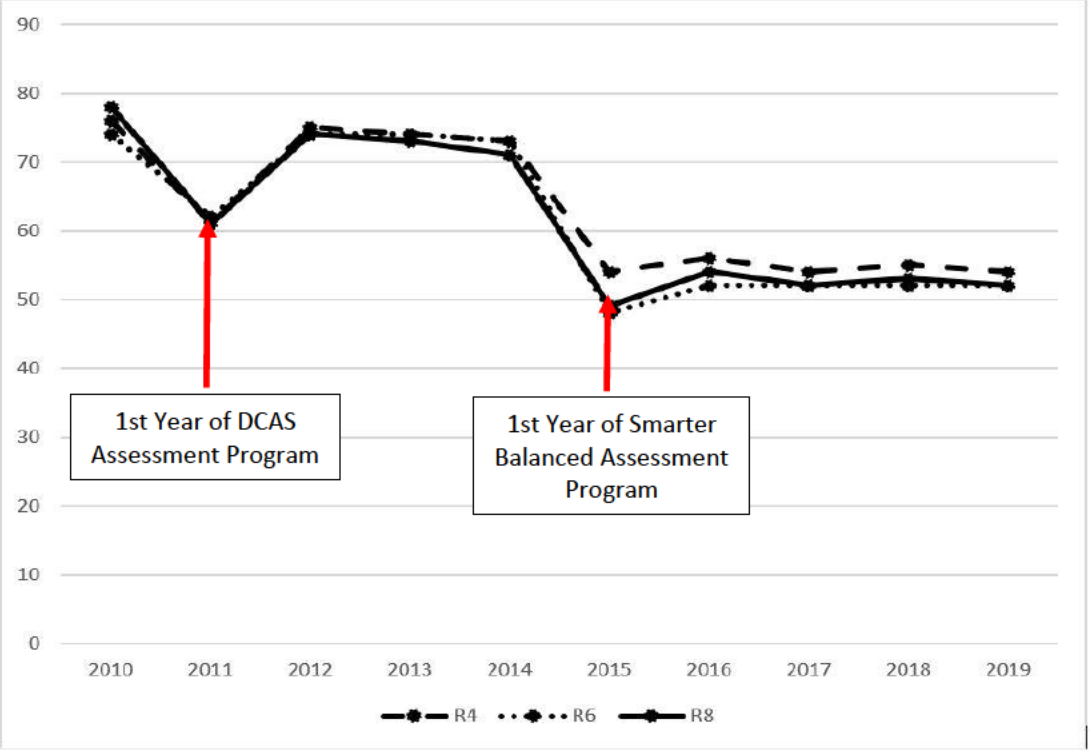


Figure 9.
Delaware Percentages Level 3 and Above, Reading, Grades 4, 6, and 8, 2009-2010 through 2018-2019

Sources: Delaware Department of Education (2010, 2011, 2012, 2014, 2019, September); State of Delaware (2019)

CONCLUSIONS: Predictable patterns exist in percentages of students classified at a given performance level within a state that are unrelated to general quality of students' educational experiences. The adoption of more rigorous content standards and/or exhortatory performance standards by a state is typically followed by declines in performance, followed by increases in the early years of adoption, followed by more stable percentages in subsequent years.

E. Relationships between Proficiency Levels and Educational Progress

As indicated in Section D1 Conclusions, percentages of students scoring at a Proficiency level are not appropriately used as indictors of educational quality or adequacy of a state's educational system. Because content standards, performance standards, and the rigor of state assessments change over time, the percentages over time refer to different contexts and cannot be directly compared. In practice, this means for example that *lower* percentages of current students in a state reaching a Proficiency level may be associated with *greater* achievement compared to higher percentages in the past. Thus, although routine increases in the rigor of a state's content standards and assessments may serve as important drivers for improvement, a stable, external indicator is necessary for monitoring a state's progress over time.

The National Assessment of Educational Progress (NAEP) is such an indicator. NAEP assessments are given every two years in reading and mathematics to samples of students across the U.S. in grades 4 and 8. In addition to reporting on the performance of students in the U.S. (the NAEP assessments are sometimes referred to as "The Nation's Report Card") using the familiar performance level categories of *Below Basic*, *Basic*, *Proficient*, and *Advanced*, NAEP permits monitoring of progress for individual states. NAEP data for Delaware are available for approximately the past 30 years.

Figures 10-13 provide the average NAEP scaled scores for students in Delaware compared to all public school students in the U.S. Results are shown for Grades 4 and 8 reading and mathematics

for NAEP assessments administered since 1990 (or when first administered in Delaware).

Two lines are shown in each of the following figures. The darker line plots the average scaled scores for students in U.S. public schools; the lighter line plots the average scaled scores for Delaware students. Each of the figures show that Delaware students have performed as well as, or in many instances, superior to students across the U.S. for as long as the NAEP assessments have been administered. Although performance in Delaware in reading has remained fairly stable (as it has across the U.S.), longitudinal performance in mathematics demonstrates a noticeable trajectory of increased achievement.

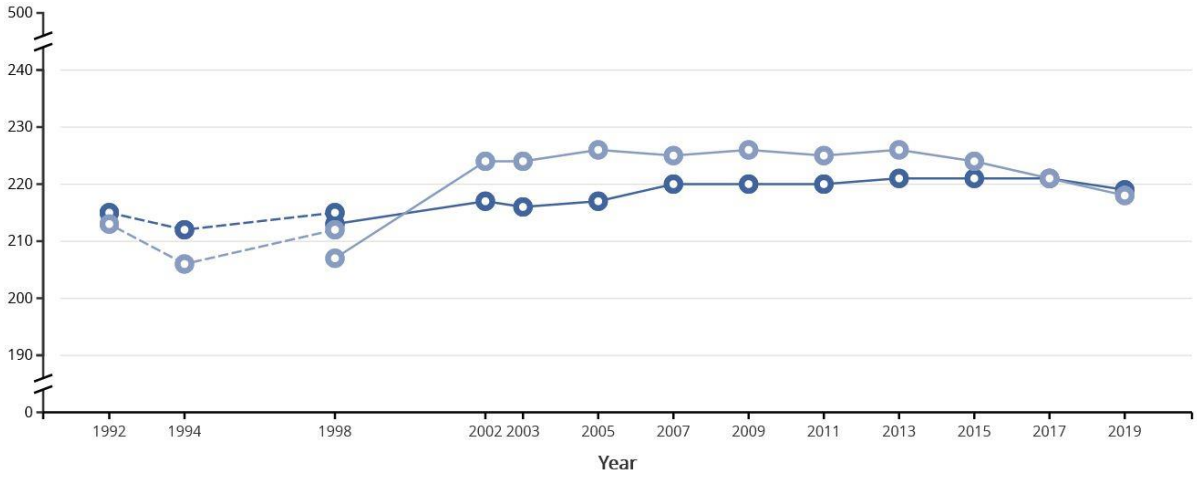


Figure 10.
NAEP Results for U.S. Public Schools and Delaware, Grade 4 Reading

Source: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore)

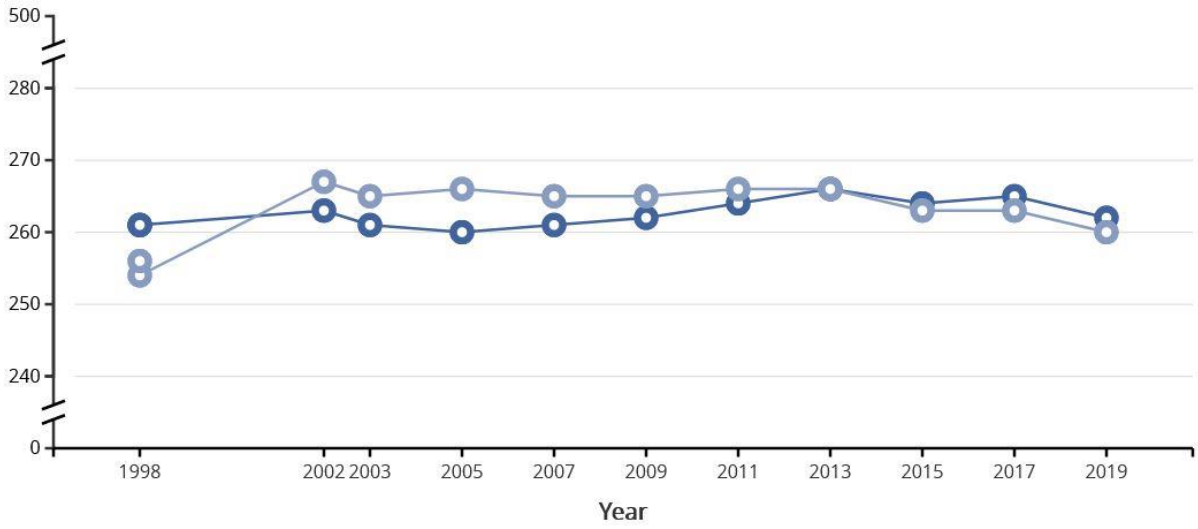


Figure 11.
NAEP Results for U.S. Public Schools and Delaware, Grade 8 Reading

Source: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore)

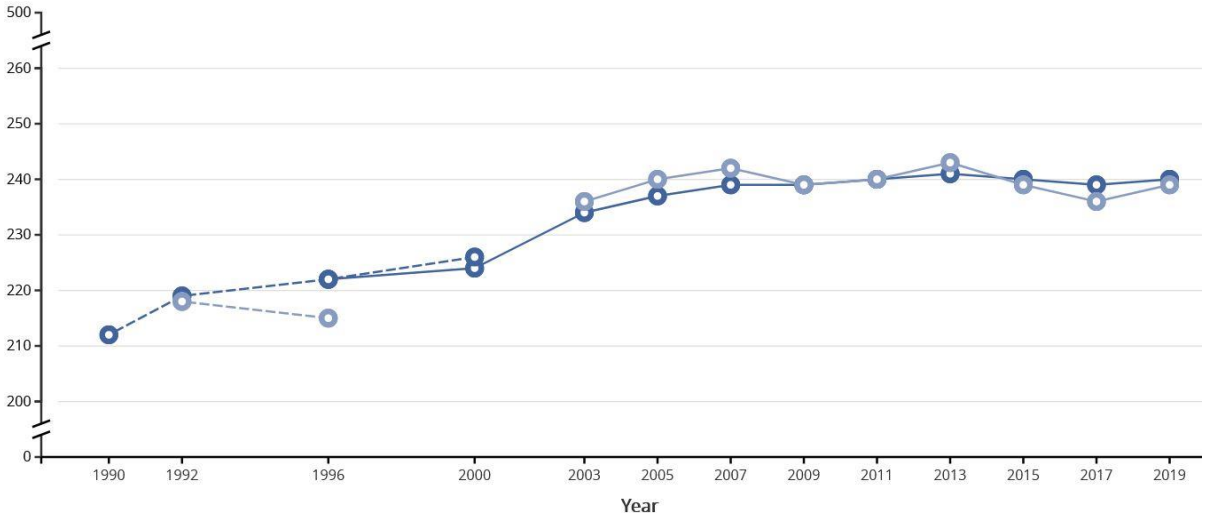


Figure 12.
NAEP Results for U.S. Public Schools and Delaware, Grade 4 Mathematics

Source: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore)

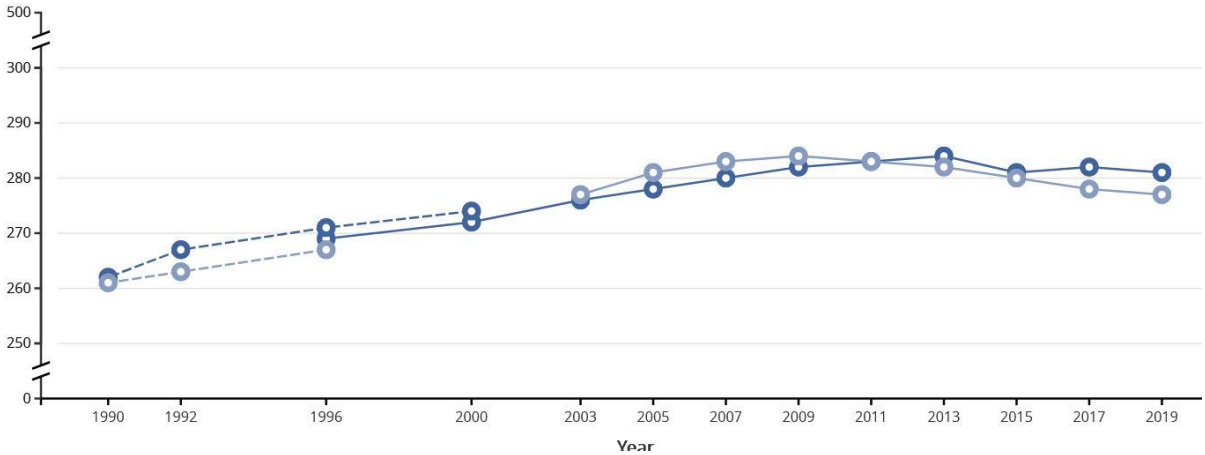


Figure 13.
NAEP Results for U.S. Public Schools and Delaware, Grade 8 Mathematics

Source: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore)

Similar results are evident in English Language Arts/Literacy and Mathematics on the Smarter Balanced assessments given in recent years (2015-2019). Figures 14 and 15 show average scaled scores for Delaware students in grades 3 through 8 across that time period in ELA/Literacy and Mathematics, respectively.

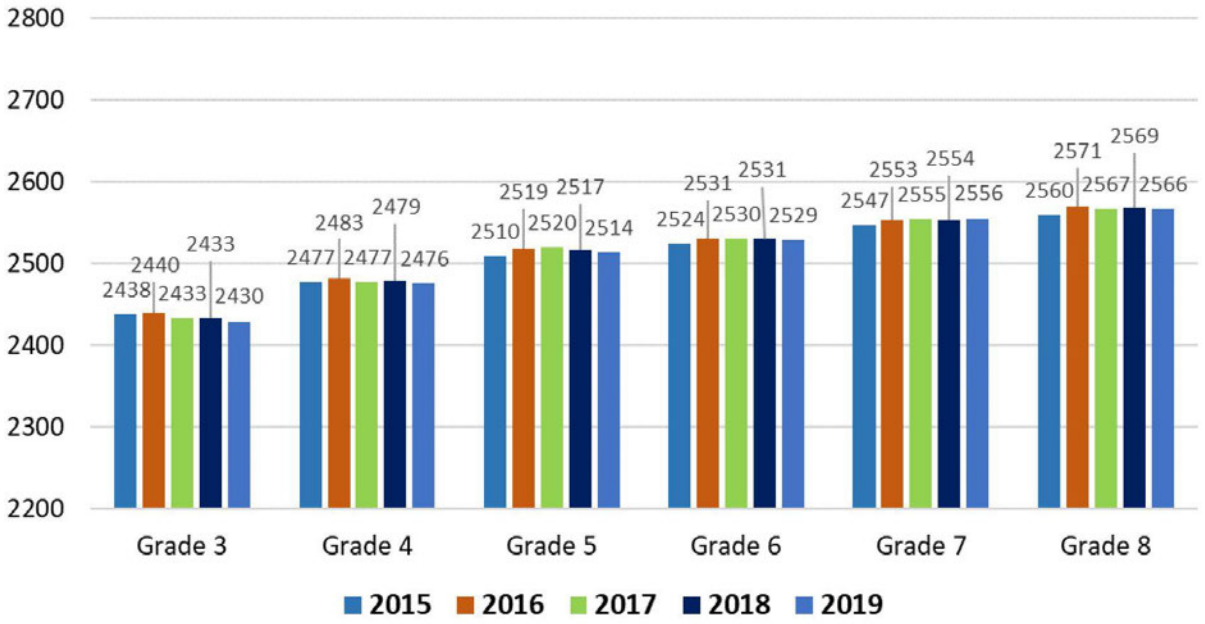


Figure 14.
Delaware Smarter Balanced Assessment Results, ELA/Literacy, 2015-2019

Source: Delaware Department of Education (2019, September) *Delaware System of Student Assessments (DeSSA) executive state summary, 2018-2019 administration* Dover, DE: Author

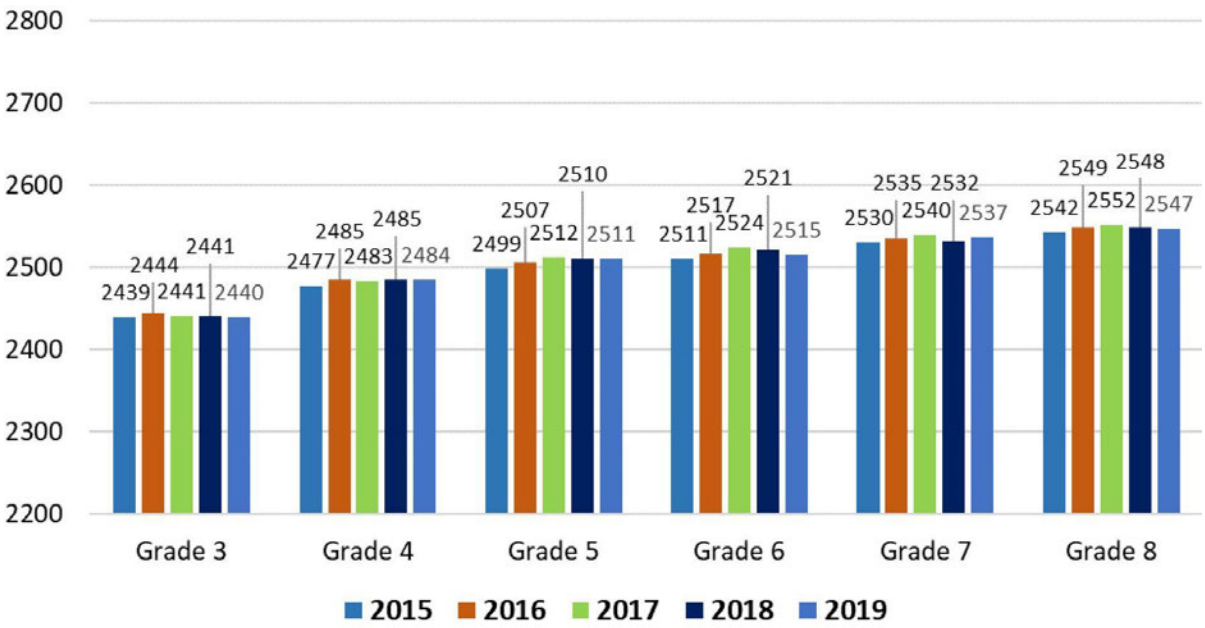


Figure 15.
Delaware Smarter Balanced Assessment Results, Mathematics, 2015-2019

Source: Delaware Department of Education (2019, September) *Delaware System of Student Assessments (DeSSA) executive state summary, 2018-2019 administration.* Dover, DE: Author

The preceding results show educational progress for all students in Delaware. However, it is often the progress of disadvantaged students that is of keen interest. Data from the National Assessment of Educational Progress (NAEP) data sets permit longitudinal examination of performance of at-risk students, identified in NAEP data collections as students eligible for participation in the National School Lunch Program (NSLP). NSLP status is often used as proxy for overall disadvantage.

Two types of analyses are relevant to gauging the progress of such students in Delaware. First, NAEP data can be used to examine the extent to which achievement gaps have been reduced in reading and mathematics across all participating jurisdictions (e.g., states) over the period 1996 (or 1998) until the present (2019). An achievement "gap" exists when differences in performance exist for student groups of interest. Achievement gap reduction is often of significant interest—that is, it is of interest to monitor the extent to which the differences in performance for the groups of students are being reduced. Figures 16-19 illustrate how well states have reduced achievement gaps between NSLP-eligible and non-NSLP-eligible students, arranged from least success in gap reduction (shown with red bars, indicating that achievement gaps have increased) to greatest success in reducing achievement gaps (shown with green bars). The height of the bars in each figure represents the size of the gap reduction, with taller red bars indicating greater gap widening, and taller green bars representing greater reduction in the performance gap between at-risk and non-at-risk students.

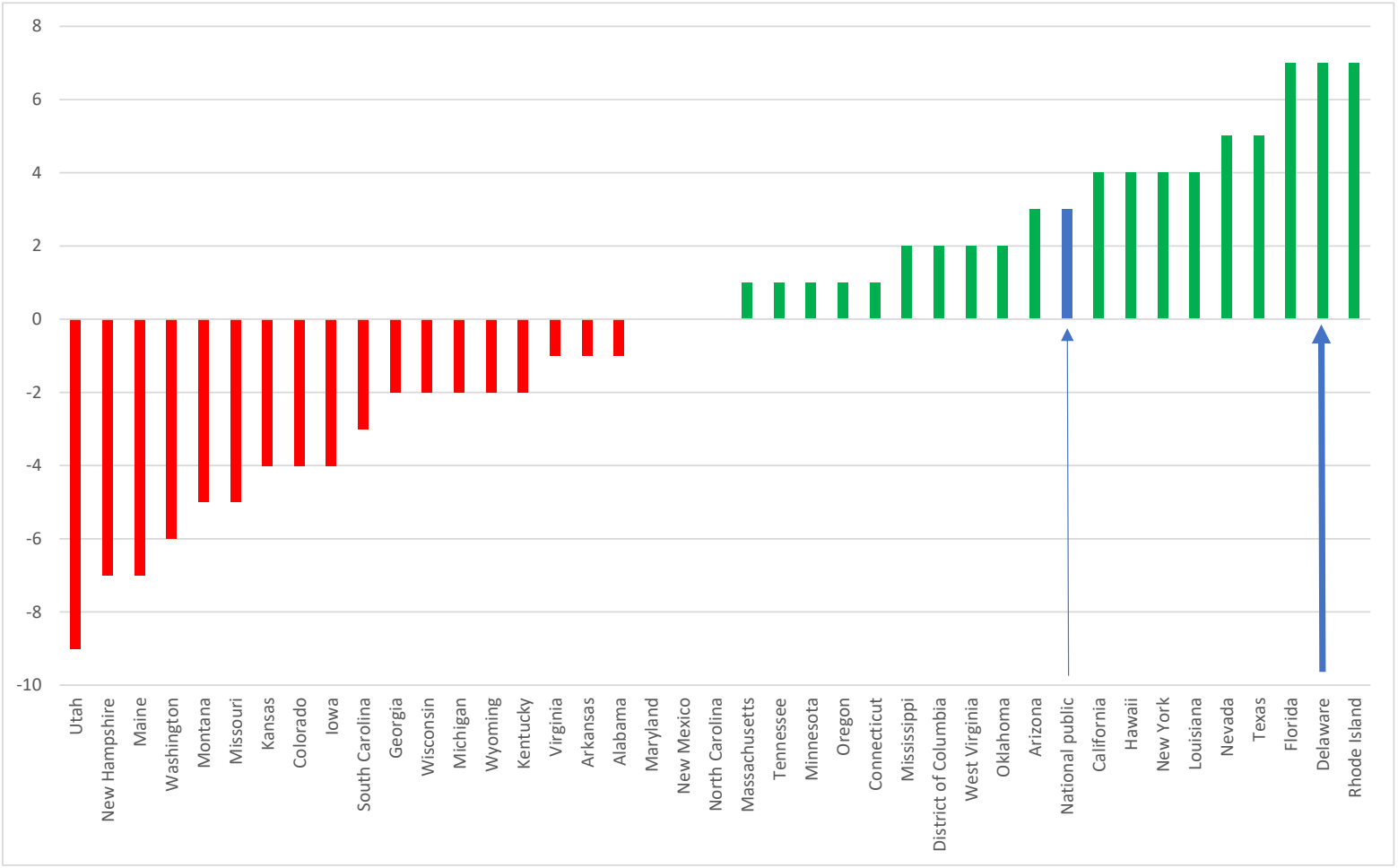
For example, Figure 16 provides gap-reduction results ordered from left to right by the size of the gap reduction in Grade 4 reading scaled scores accomplished from 1998-2019 for all jurisdictions for which data were available. The red bars at the left of the figure indicate that gaps between NSLP-eligible and non-NSLP-eligible students actually increased, with taller bars indicating greater gap widening. The green bars on the right indicate the sizes of gap narrowing, with taller bars indicating greater gap reductions. The blue bars show the U.S. national average gap reduction. Thin blue arrows indicate the national public school averages; thick blue arrows indicate Delaware's standing. As can be

seen in Figure 16, Delaware was tied (with Florida and Rhode Island) for the largest gap reduction of all participating states on the grade 4 reading assessment.

Similar results were observed for grade 8 reading, grade 4 mathematics, and grade 8 mathematics. Figure 17 shows that Delaware was tied (with California) for the second largest gap reduction in reading at grade 8; Figure 18 reveals that Delaware produced the largest gap reduction of all states in 4th grade mathematics; and Figure 19 illustrates that Delaware was tied (with Virginia) for the second largest gap reduction in mathematics at grade 8.

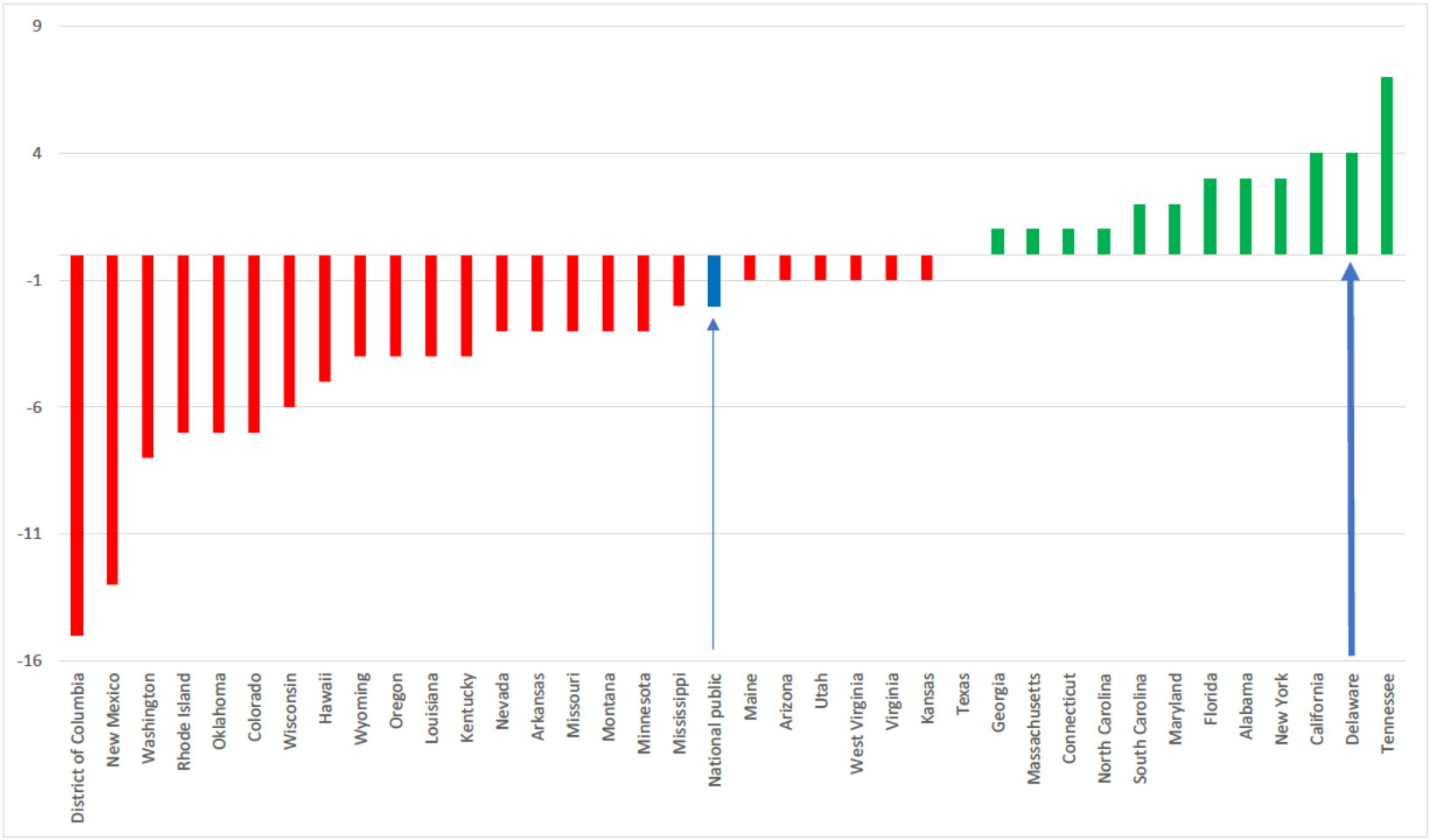
As an important aside, it should be noted that the narrowing of achievement gaps in Delaware was *not* the result of lower achievement by non-NSLP students. Rather, the NAEP data indicate that there were modest-to-large achievement gains for NSLP-eligible students compared to all other NAEP-participating jurisdictions. Figures 20-23 provide the size of scaled score increases in achievement for the same grades, subjects, and over the same periods shown in Figures 16-19. Similar to the previous set of figures, red bars at the left of the figures indicate achievement losses; green bars on the right indicate the sizes of achievement gains, with taller bars indicating greater gains. The blue bars show the U.S. national average achievement gains. Blue arrows highlight national public school average gains and gains for Delaware. As can be seen in Figure 20 for example, the performance of at-risk students in Delaware was tied for seventh largest achievement increase from 1998 to 2019 in grade 4 reading; figures 21-23 show Delaware's gains in grade 8 reading, grade 4 mathematics, and grade 8 mathematics, respectively.

Figure 16
Scaled Score Gap Reduction, NSLP Eligible/Non-Eligible, NAEP Reading, Grade 4, 1998-2019



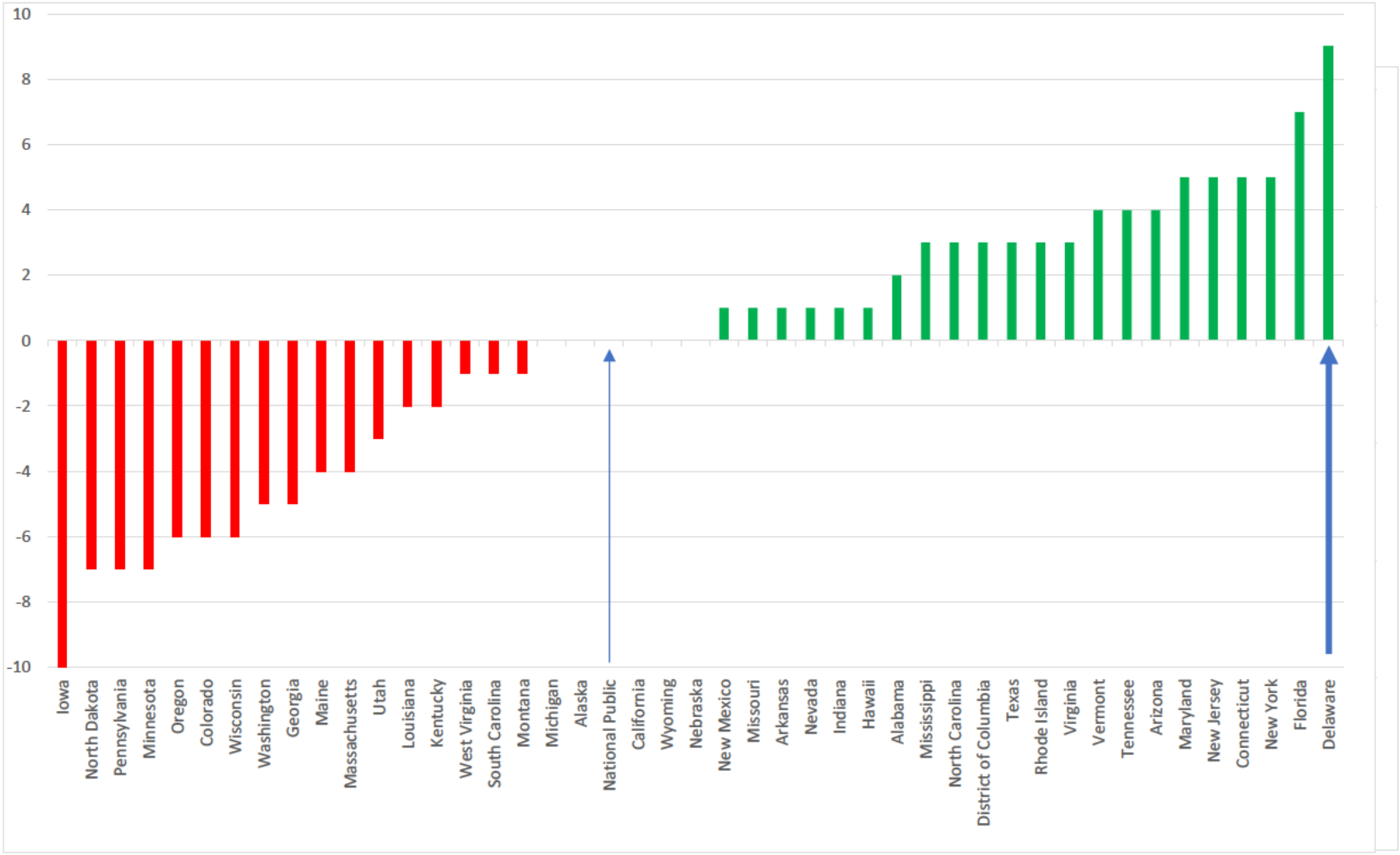
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 17
Scaled Score Gap Reduction, NSLP Eligible/Non-Eligible, NAEP Reading, Grade 8, 1998-2019



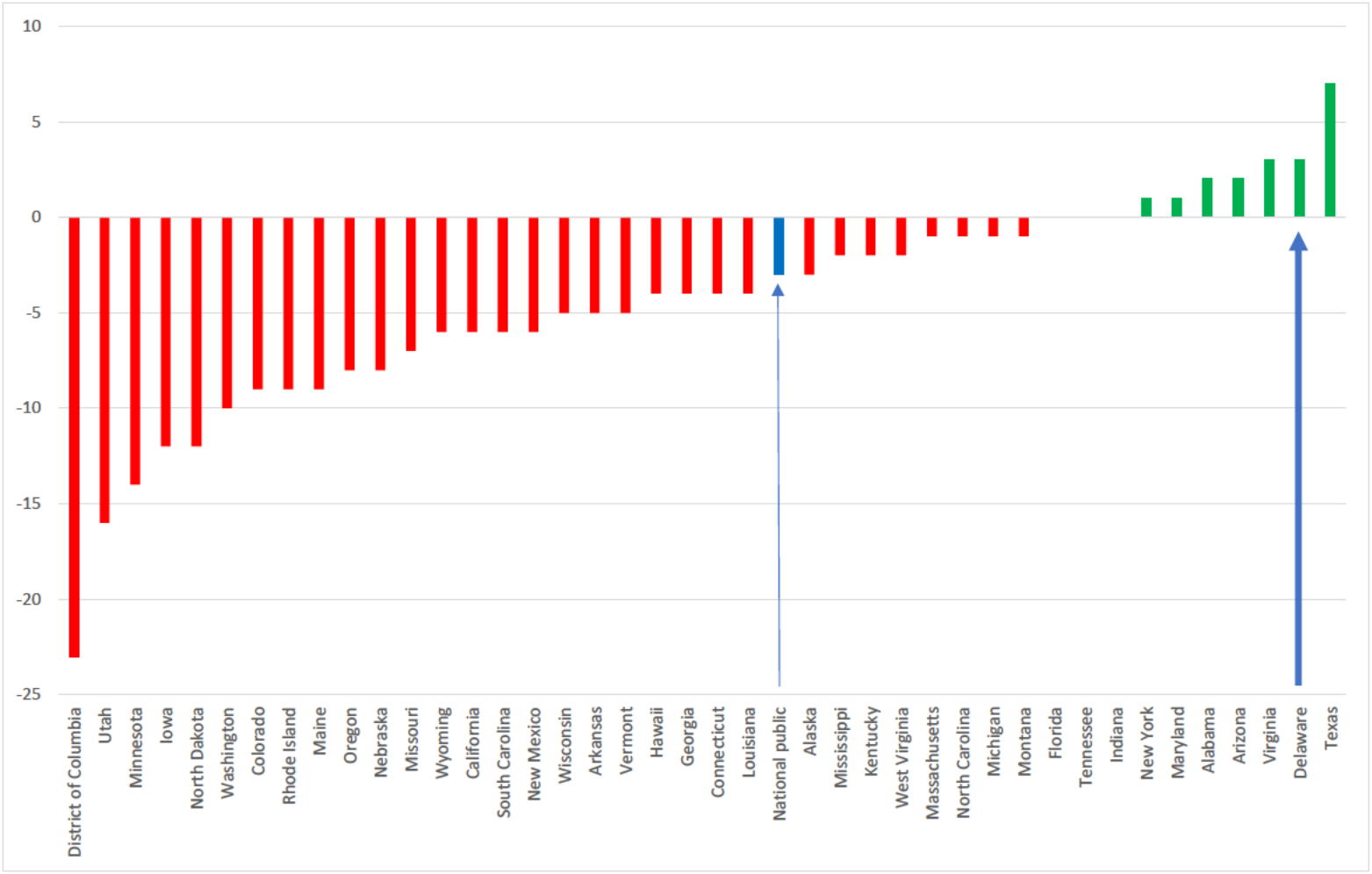
SOURCE :United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 18
Scaled Score Gap Reduction, NSLP Eligible/Non-Eligible, NAEP Mathematics, Grade 4, 1996-2019



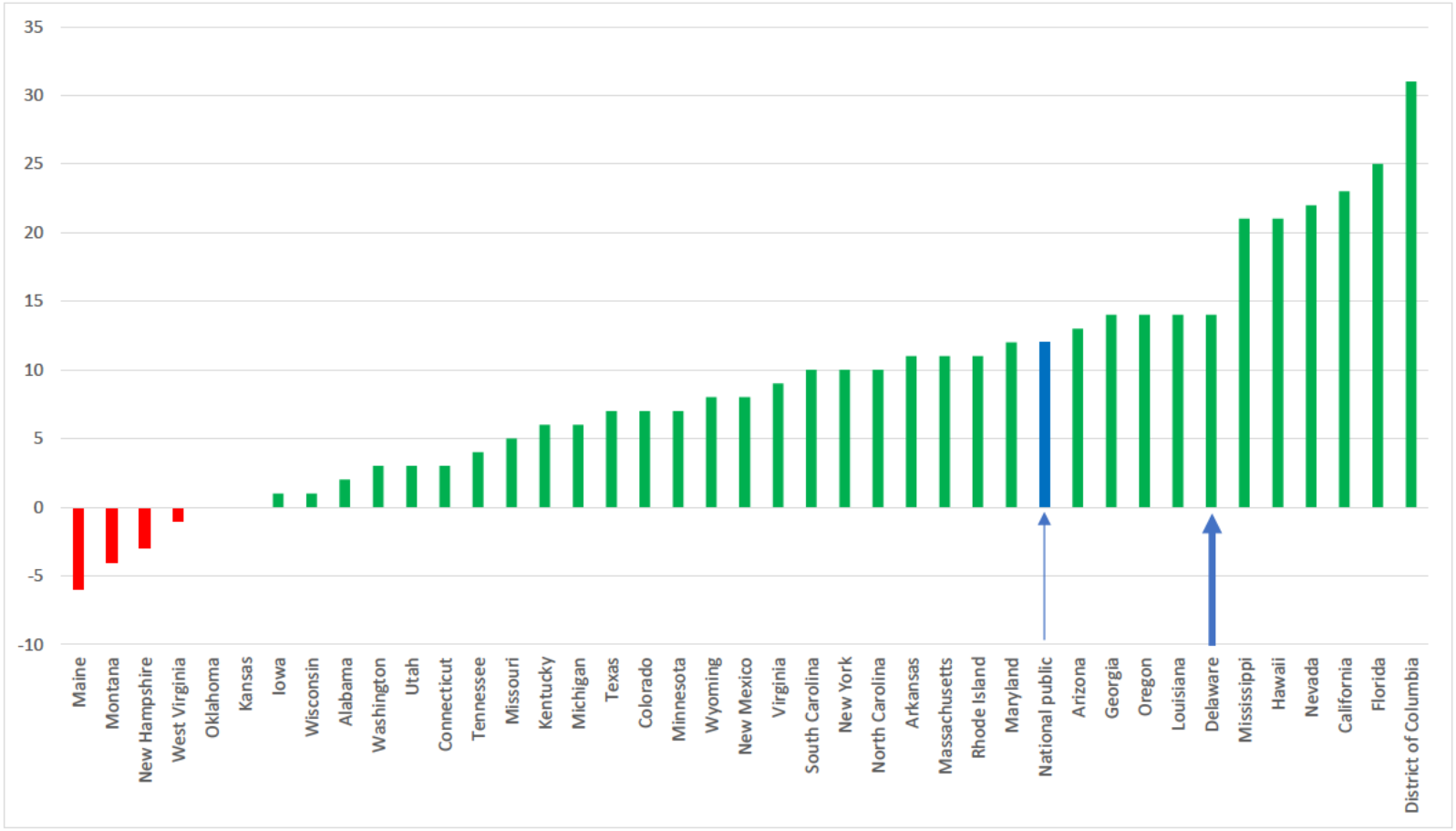
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 19
Scaled Score Gap Reduction, NSLP Eligible/Non-Eligible, NAEP Mathematics, Grade 8, 1996-2019



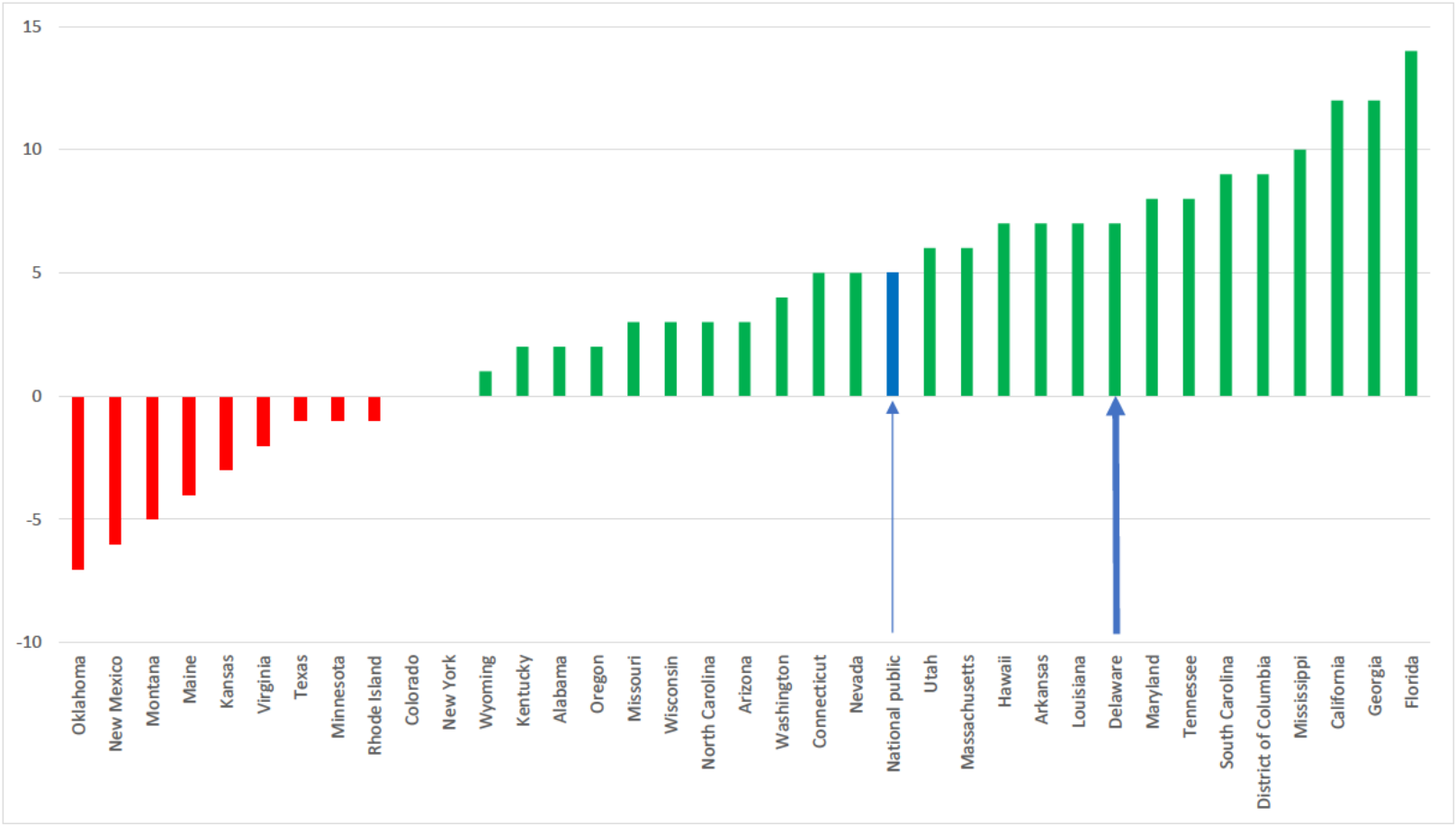
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 20
Scaled Score (Achievement) Increases, NSLP Eligible Students, NAEP Reading, Grade 4, 1998-2019



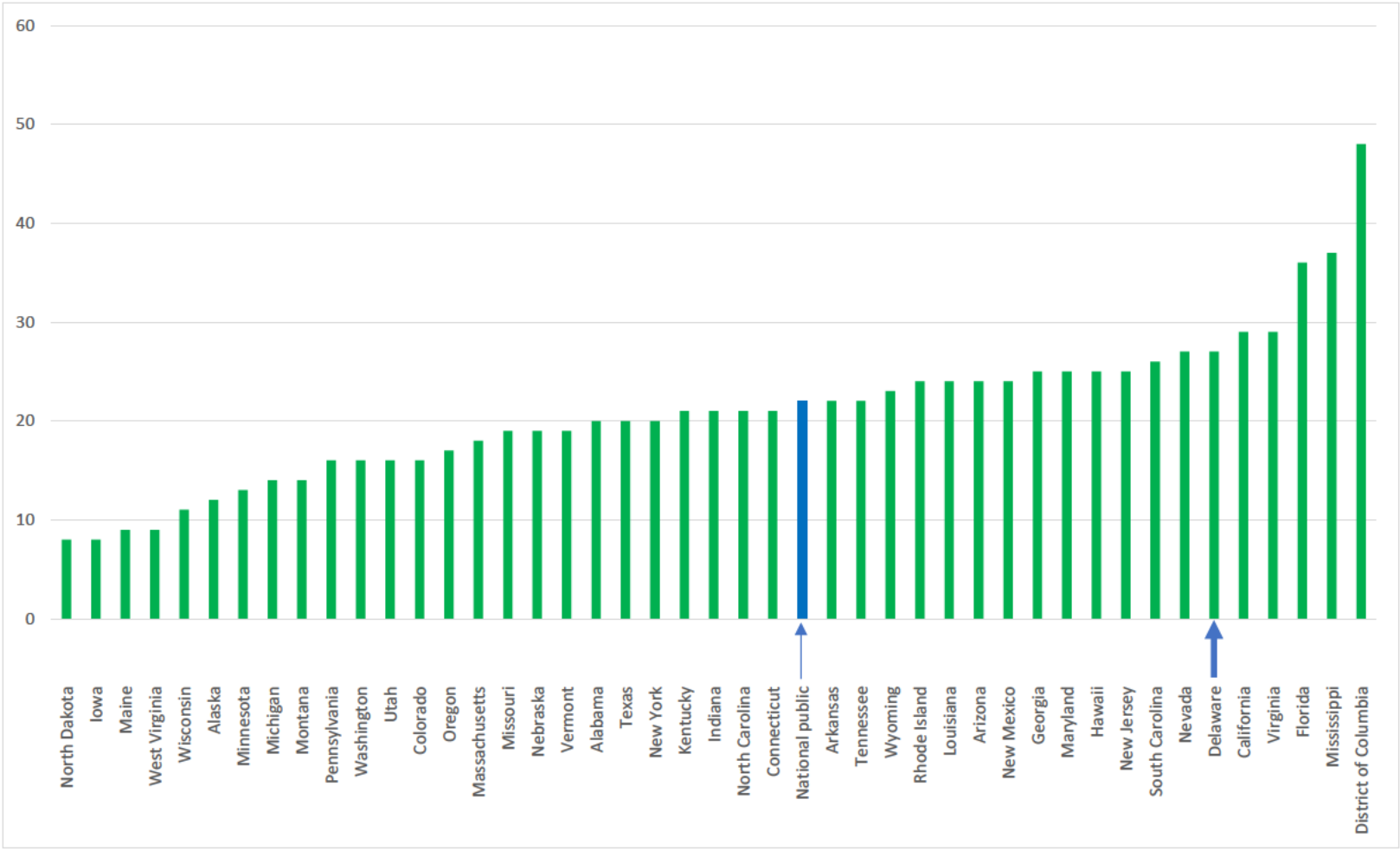
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 21
Scaled Score (Achievement) Increases, NSLP Eligible Students, NAEP Reading, Grade 8, 1998-2019



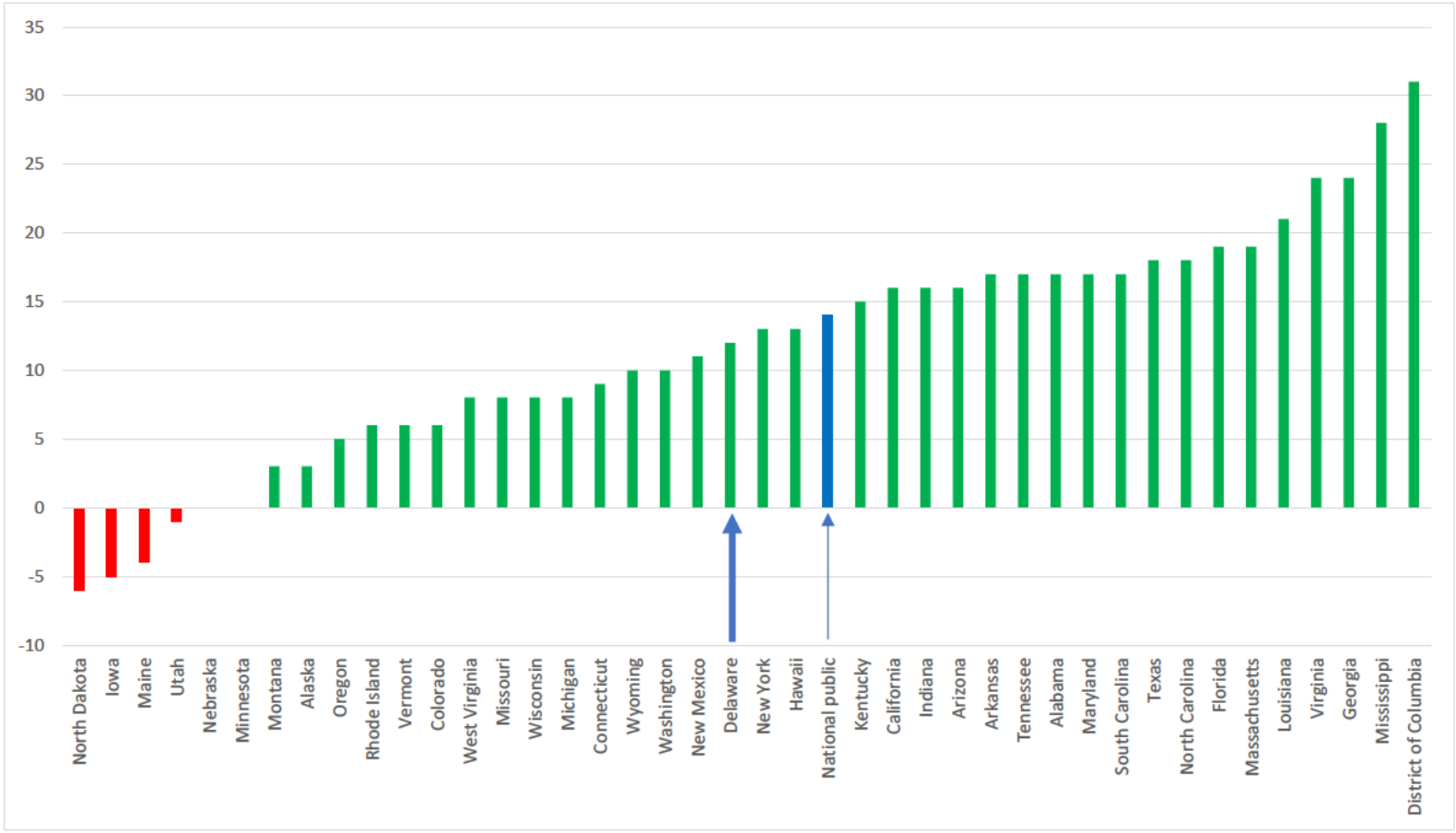
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 22
Scaled Score (Achievement) Increases, NSLP Eligible Students, NAEP Mathematics, Grade 4, 1996-2019



SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 23
Scaled Score (Achievement) Increases, NSLP Eligible Students, NAEP Mathematics, Grade 8, 1996-2019



SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

CONCLUSIONS: **Delaware has made steady educational progress in reading and mathematics since the 1990s, with trajectories showing somewhat greater increases in grade 4 and 8 mathematics compared to reading. Especially noteworthy is that achievement progress in Delaware is not limited to students overall, but has been accomplished for at-risk students as evidenced by the progress made by Delaware in both absolute achievement gains and in achievement gap reduction for at-risk students.**

VII. OVERALL SUMMARY AND CONCLUSIONS

The purpose of this report was to provide descriptive information, analyses, and expert opinions relevant to the appropriateness of drawing conclusions about educational adequacy from student performance from large-scale student achievement test data. A number of conclusions were reached.

All tests, including the Smarter Balanced and NAEP assessments, are designed for a specific purpose. The assessments used in Delaware were designed and validated as measures of student achievement. They were not designed or validated as measures of instructional quality, educator effectiveness, or educational system adequacy. Because it can result in undependable and inaccurate conclusions, it is inappropriate to use test data for unintended and unvalidated purposes.

A variety of educational reforms initiated in the 1970s and known as "the standards movement" has focused on increasing the rigor of content and performance expectations in American schools. Key aspects of the standards movement have been the introduction of mandated assessments and accountability systems for educational organizations. These requirements have been intended to spur increases in student achievement.

The assessments that are central to these reforms comprise *content standards* (sets of specific learning objectives that define the knowledge and skills students are to be taught) and *performance standards* (which specify how well students must perform in order to be classified into an achievement level). The levels of achievement represented by performance standards are established using systematic procedures to collect informed, qualified judgments of diverse stakeholders. The performance levels for Delaware state assessments have been established using contemporary best practices for standard setting.

Federal educational policy has given states wide latitude in adopting content and performance

standards that reflect greater or lesser degrees of rigor. Over three generations of standards and assessments (DSTP, DCAS, and Smarter Balanced), the standards and assessments adopted by Delaware reflect a history of increasing expectations for students and Delaware schools. The current sets of content standards and performance standards adopted by Delaware—the Common Core content standards and the Smarter Balanced assessments and performance standards—reflect a commitment to increasing levels of challenge.

Performance of Delaware students has demonstrated a predictable pattern when each generation of more rigorous content standards and/or exhortatory performance standards was adopted; namely, an initial decline in performance, followed by increases in the early years of adoption as students and systems adjust to the increased expectations, followed by more stable percentages in subsequent years. Also evident are long-term trends of increasing achievement in Delaware on both local (Smarter Balanced) and national (NAEP) assessments.

Performance standards are not typically established to reflect a status quo, but to serve a policy function. As was the case when they were established for Delaware assessments, performance standards are intended to to serve as exhortatory goals for promoting reforms, effort, greater achievement, and as mechanisms for monitoring progress. Such performance standards are not set with the belief that 100% of students will meet them. Performance standards could technically be set at any level, including at a level where all or nearly all students would be classified as Proficient. Because performance standards can be set at any level, it can be easily seen that using percentages of students at a performance level cannot be used as a measure of educational quality or adequacy; simply setting performance standards that were reached by 100% of students would not only fail to demonstrate educational adequacy, such a decision would be counterproductive to stimulating educational progress, and would misinform students, their parents, and educational personnel.

Delaware has a history of establishing challenging educational goals. Most recently, as part of

its goals required by the federal Every Student Succeeds Act, Delaware has established achievement targets for 2030 in grades 3-8 of 76% proficiency in English language arts and 70% proficiency in mathematics for general population students, and 42% and 41% proficiency in English language arts and mathematics, respectively, for students with special needs and English language learners.

Because state-level content standards and performance standards are often periodically adjusted to reflect increasing expectations for student learning, it is necessary to refer to stable, external benchmarks to gauge educational progress. One such external benchmark is the National Assessment of Educational Progress (NAEP). On this benchmark, Delaware has made steady educational progress in reading and mathematics since the 1990s, demonstrating progress not only for students overall, but also for at-risk/disadvantaged students, in many cases ranking first nationally in gap reduction.

_____ ___2020, March 11____

Gregory J. Cizek Date

VIII. REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

American Recovery and Reinvestment Act. (2009). Public Law 111–5.

Board of Regents of the University of Wisconsin System. (2014). *The English language development standards, kindergarten–grade 12.* Madison, WI. Author.

Bandeira de Mello, V., Rahman, T., Fox, M.A., & Ji, C. S. (2019). *Mapping state proficiency standards onto the NAEP scales: Results from the 2017 NAEP reading and mathematics assessments (NCES 2019-040).* U.S. Department of Education. Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved October 30, 2019 from https://nces.ed.gov/pubsearch.

Cizek, G. J. (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2012). *Setting performance standards: Foundations, methods, and innovations.* New York, NY: Routledge.

Cizek, G. J., & Bunch, M. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: SAGE.

Delaware Department of Education. (2010, July). *Delaware student testing program: State summary results of the reading and mathematics assessments spring 2010 administration, grades 2 through 10.* Dover, DE: Author.

Delaware Department of Education. (2011, July). *Delaware comprehensive assessment system: State*

summary results of the reading, mathematics, science, and social studies assessment, spring 2011 administration, Grades 3 through 10. Dover, DE: Author.

Delaware Department of Education. (2012, July). *Delaware comprehensive assessment system: State summary results of the reading, mathematics, science, and social studies assessments.* Dover, DE: Author.

Delaware Department of Education. (2014, July). *Attachment 1 – 2014 DCAS state summary statewide performance and participation summary*. Dover, DE: Author.

Delaware Department of Education. (2019, June 10). *Delaware consolidated state plan under the Every Student Succeeds Act.* Dover, DE: *Author.*

Delaware Department of Education. (2019, September). *Delaware System of Student Assessments (DeSSA) executive state summary, 2018-2019 administration*. Dover, DE: Author.

Delaware State Board of Education. (2015, January 6). *State board of education, assessment and accountability workshop.* Dover, DE: Author.

Every Student Succeeds Act. (2015). Public Law 114–95.

Goals 2000: Educate America Act. (1994). Public Law 103-227.

Leadership Conference on Civil and Human Rights. (2015, May 5). *Civil rights groups: "We oppose anti-testing efforts."* Washington, DC: Author. Retrieved from: https://civilrights.org/2015/05/05/civil-rights-groups-we-oppose-anti-testing-efforts/

Linn, R. L. (1995). The likely impact of performance standards as a function of uses: From rhetoric to sanctions. In *Proceedings of the joint conference on standard setting for large-scale assessment of the National Assessment Governing Board and the National Center for Education Statistics* (pp. 367-374). Washington, DC: U.S. Government Printing Office.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16.

Linn, R. L. (2003, September 1). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved 4 November 2019 from http://epaa.asu.edu/epaa/v11n31/.

Linn, R. L. (2004, July 28). *Rethinking the No Child Left Behind accountability system.* Paper presented at the Center of Education Policy Forum on No Child Left Behind, Washington, DC.

Lumos Learning. (2020). *Understanding the SBAC report card*. Retrieved 1/9/2020 from https://www.lumoslearning.com/llwp/teachers-speak/understanding-the-sbac-report-card.html

Maryland State Department of Education. (2008). Maryland social studies standards and framework, grade 3-8. Annapolis, MD: Author. Retrieved from http://www.marylandpublicschools.org/about/Documents/DCAA/SocialStudies/MSSSS/MarylandSocialStudiesStandardsFrameworGrades6-8.pdf

National Assessment Governing Board. (2018). *Developing student achievement levels for the National Assessment of Educational Progress.* Washington, DC: Author. Retrieved from https://www.nagb.gov/content/nagb/assets/documents/policies/ALS-revised-policy-statement-11-17-18.pdf

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

National Governors' Association (1986). *Time for results: The governors' 1991 report on education.* Washington, DC: National Governors' Association.

National Governors Association [NGA] Center for Best Practices and Council of Chief State School Officers (CCSSO]. (2010). *Common core state standards.* Washington, DC: CCSSO.

National Research Council. (2012). A framework for K-12 science education practices, crosscutting concepts, and core ideas. Washington, DC: National Academies Press.

New Jersey Department of Education (2014). *New Jersey student learning standards for comprehensive health and physical education*. Trenton, NJ. Author. Retrieved from https://www.state.nj.us/education/cccs/2014/chpe/standards.pdf

New York State Education Department. (2017). *New York state next generation learning standards.* Albany, NY: Author. Retrieved from http://www.nysed.gov/common/nysed/files/programs/curriculum-instruction/nys-next-generation-ela-standards.pdf

Next Generation Science Standards. (2013). *Next generation science standards, fifth grade*. Retrieved from https://www.nextgenscience.org/sites/default/files/5%20combined%20DCI%20standardsf.pdf

No Child Left Behind Act. (2001). Public Law 107-110.

Paley, A. R. (2007, March 14). 'No Child' target is called out of reach; Goal of 100% proficiency debated as congress weighs renewal. *Washington Post.* [Retrieved November 4, 2019].

Pennsylvania Department of Education. (2014). *PA academic standards for mathematics, grades pre-K – high school*. Harrisburg, PA: Author.

Saulny, S. (2005, January 19). Meaning of 'proficient' varies for schools across country. *New York Times*, B8.

Smarter Balanced Assessment Consortium. (2013). *Initial achievement level descriptors and college content-readiness policy*. Santa Cruz, CA: Author. Retrieved from https://portal.smarterbalanced.org/library/en/mathematics-alds-and-college-content-readiness-policy.pdf

Smarter Balanced Assessment Consortium. (2015a). Grade level tables for all claims and assessment targets and item types. Retrieved from https://portal.smarterbalanced.org/library/en/appendix-b-grade-level-tables.pdf

Smarter Balanced Assessment Consortium. (2015b). *Content specifications for the summative assessment of the common core state standards for English language arts and literacy in*

history/social studies, science, and technical subjects. Santa Cruz, CA: Author.

Smarter Balanced Assessment Consortium. (2015c). *Smarter Balanced technical report: Standard settings.* Santa Cruz, CA: Author.

State of Delaware. (2019). Delaware open data portal. Dover, DE: Author. Retrieved from: https://data.delaware.gov/Education/Student-Assessment-Performance/ms6b-mt82/data

Touchette, B. M. (2014, December 18). *Smarter achievement level setting: Presentation to the Delaware State Board of Education*. Dover, DE.

United States Department of Education, National Center for Education Statistics. (2019). NAEP data explorer. Washington, DC: Author. Available at https://nces.ed.gov/nationsreportcard/data/

University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems. (2019). *Dynamic learning maps English language arts year-end assessment model blueprint.* Lawrence, KS: Author.

The White House. (2012). *Everything you need to know: Waivers, flexibility, and reforming No Child Left Behind*. Retrieved from https://obamawhitehouse.archives.gov/blog/2012/02/09/everything-you-need-know-waivers-flexibility-and-reforming-no-child-left-behind

APPENDIX A-1

Delaware Percent of Students Level 3 or Above, ELA and Mathematics, 2009-2010 through 2018-2019

	Percent Proficient or Above	
School Year/Grade	*English Language Arts*	*Mathematics*
2018-2019/3	**51**	**53**
2018-2019/4	**54**	**51**
2018-2019/5	**57**	**44**
2018-2019/6	**52**	**38**
2018-2019/7	**55**	**41**
2018-2019/8	**52**	**38**
2018-2019/HS	**48**	**28**
2017-2018/3	**52**	**54**
2017-2018/4	**55**	**50**
2017-2018/5	**58**	**43**
2017-2018/6	**52**	**40**
2017-2018/7	**54**	**39**
2017-2018/8	**53**	**39**
2017-2018/(SAT)	**50**	**28**
2016-2017/3	**52**	**53**
2016-2017/4	**54**	**50**
2016-2017/5	**60**	**44**
2016-2017/6	**52**	**41**
2016-2017/7	**54**	**41**
2016-2017/8	**52**	**38**
2016-2017/(SAT)	**53**	**29**
2015-2016/3	**54**	**55**
2015-2016/4	**56**	**51**
2015-2016/5	**60**	**42**
2015-2016/6	**52**	**37**
2015-2016/7	**53**	**40**
2015-2016/8	**54**	**38**
2015-2016/(SAT)	**52**	**31**
2014-2015/3	**54**	**53**
2014-2015/4	**54**	**47**
2014-2015/5	**56**	**38**
2014-2015/6	**48**	**34**
2014-2015/7	**50**	**37**
2014-2015/8	**49**	**35**
2014-2015/HS	--	--
2013-2014/3	**70**	**72**
2013-2014/4	**73**	**75**
2013-2014/5	**76**	**71**
2013-2014/6	**73**	**66**
2013-2014/7	**71**	**66**
2013-2014/8	**71**	**69**
2013-2014/(10)	**75**	**69**
2012-2013/3	**71**	**73**
2012-2013/4	**74**	**74**
2012-2013/5	**77**	**71**
2012-2013/6	**74**	**66**

2012-2013/7	**72**	**67**
2012-2013/8	**73**	**71**
2012-2013/(10)	**73**	**69**
2011-2012/3	**76**	**77**
2011-2012/4	**75**	**77**
2011-2012/5	**79**	**74**
2011-2012/6	**74**	**68**
2011-2012/7	**71**	**70**
2011-2012/8	**74**	**74**
2011-2012/HS	**71**	**71**
2010-2011/3	**63**	**67**
2010-2011/4	**61**	**65**
2010-2011/5	**65**	**65**
2010-2011/6	**62**	**57**
2010-2011/7	**59**	**60**
2010-2011/8	**61**	**62**
2010-2011/(10)	**63**	**59**
2009-2010/3	**77**	**78**
2009-2010/4	**76**	**78**
2009-2010/5	**81**	**76**
2009-2010/6	**74**	**73**
2009-2010/7	**83**	**70**
2009-2010/8	**78**	**68**
2009-2010/(10)	**64**	**57**

Sources: Delaware Department of Education (2010, 2011, 2012, 2014, 2019, September); State of Delaware (2019)

APPENDIX A-2

Smarter Balanced Percentages at or Above Performance Levels by Grade for English Language Arts/Literacy (ELA-L) and Mathematics (M), 2015

	Percentages at or Above the Performance Level		
Subject/Grade	*Level 2*	*Level 3*	*Level 4*
ELA-L/3	65	38	18
ELA-L/4	63	41	18
ELA-L/5	67	44	15
ELA-L/6	70	41	11
ELA-L/7	66	38	8
ELA-L/8	72	41	9
ELA-L/11	72	41	11
M/3	68	39	12
M/4	73	37	13
M/5	65	33	15
M/6	65	33	14
M/7	64	33	13
M/8	62	32	13
M/11	60	33	11

Source: Touchette, B M (2014, December 18) Smarter achievement level setting: Presentation to the Delaware State Board of Education Dover, DE

APPENDIX B
Figures and Tables from Body of Report

Table 1
Cases in which the Witness Has Testified as an Expert in the Last 5 Years

Case Name/Citation [Court]	Year	Topic of Expert Testimony
Commonwealth of Pennsylvania v. Ary Sloane, Docket No. CP-51-CR-0009924-2014 [Philadelphia Municipal Court, Criminal Section]	2014	psychometrics, validity of test scores
Commonwealth of Pennsylvania v. Barbara McCreery, Docket No. CP-51-CR-0011071-2014 [Philadelphia Municipal Court, Criminal Section]	2014	psychometrics, validity of test scores
Commonwealth of Pennsylvania v. Arthur Melton, Docket No. CP-51-CR-0011945-2014 [Philadelphia Municipal Court, Criminal Section]	2014	psychometrics, validity of test scores
Commonwealth of Pennsylvania v. Lola Marie Davis-O'Rourke, Docket No. CP-51-CR-0000916-2015 [Philadelphia Municipal Court, Criminal Section]	2015	psychometrics, validity of test scores
State of Georgia v. Dr. Beverly Hall et al., Case No. 13-SC-1179S4 [Fulton County (GA) Superior Court]	2015	psychometrics, validity of test scores

Table 2
Examples of General Education Content Standards

State/Group	Grade	Subject	Specific Content Standard
Pennsylvania	8	Mathematics	CC.2.2.8.B.1: *Apply concepts of radicals and integer exponents to generate equivalent expressions.*
New York	4	English Language Arts	4R1: *Locate and refer to relevant details and evidence when explaining what a text says explicitly/implicitly and make logical inferences.*
Maryland	7	Social Studies/ Geography	7A1: *Use geographic tools to locate places and describe the human and physical characteristics in the contemporary world.*
New Jersey	2	Health	2.1.2.A.2: *Use correct terminology to identify body parts and explain how body parts work together to support wellness.*
SBAC	8	English Language Arts/Literacy	Target 14, Language Use: *Interpret understanding of figurative language, word relationships, nuances of words and phrases, or figures of speech (e.g., verbal irony, puns) used in context and the impact of those word choices on meaning.*
NGSS	5	Science	*5-PS1-2: Measure and graph quantities to provide evidence that regardless of the type of change that occurs when heating, cooling, or mixing substances, the total weight of matter is conserved.*

Sources: Pennsylvania Department of Education, (2014); New York State Education Department, (2017); Maryland State Department of Education, (2008); New Jersey Department of Education, 2014); Smarter Balanced Assessment Consortium, (2015a); Next Generation Science Standards, (2013)

Table 3
Delaware and NAEP Achievement Level Descriptions, Grade 8 Mathematics

	Policy ALDs	Grade/Subject Specific ALDs
Delaware (SBAC) Achievement Levels[4]		
Level 1 (*Does Not Meet/ Minimal)*	The Level 1 student demonstrates minimal understanding of and ability to apply the English language arts and literacy (mathematics) knowledge and skills needed for success in college and career, as specified in the Common Core State Standards.	
Level 2 (*Nearly Meets/ Partial*)	The Level 2 student demonstrates partial understanding of and ability to apply the English language arts and literacy (mathematics) knowledge and skills needed for success in college and career, as specified in the Common Core State Standards.	
Level 3 (*Meets/ Adequate*)	The Level 3 student demonstrates adequate understanding of and ability to apply the English language arts and literacy (mathematics) knowledge and skills needed for success in college and career, as specified in the Common Core State Standards.	
Level 4 (*Exceeds/ Thorough*)	The Level 4 student demonstrates thorough understanding of and ability to apply the English language arts and literacy (mathematics) knowledge and skills needed for success in college and career, as specified in the Common Core State Standards.	
NAEP Achievement Levels		
NAEP Basic	This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for performance at the NAEP Proficient level.	Eighth-grade students performing at the Basic level should exhibit evidence of conceptual and procedural understanding in the five NAEP content areas. This level of performance signifies an understanding of arithmetic operations—including estimation—on whole numbers, decimals, fractions, and percents.
NAEP Proficient	This level represents solid academic performance for each NAEP assessment. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.	Eighth-grade students performing at the Proficient level should apply mathematical concepts and procedures consistently to complex problems in the five NAEP content areas.
NAEP Advanced	This level signifies superior performance beyond NAEP Proficient.	Eighth-grade students performing at the Advanced level should be able to reach beyond the recognition, identification, and application of mathematical rules in order to generalize and synthesize concepts and principles in the five NAEP content areas.

[4] Sources: Smarter Balanced Assessment Consortium (2013)/Delaware Department of Education (2019 September), National Assessment Governing Board (2018). Original SBAC Policy ALDs were phrased in slightly different terms. In addition to allowing member states flexibility in naming their performance levels, some states also adopted slight revisions of the Policy ALDs.

Table 4
Delaware Performance Standards, Grades 3-8 Mathematics

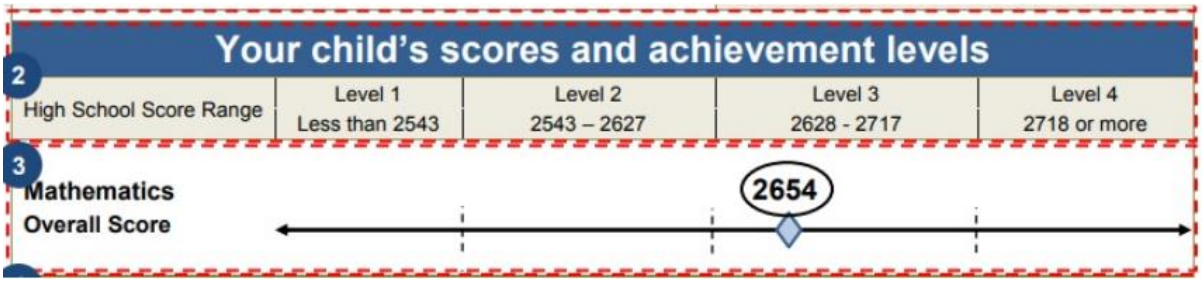
Grade	*Level 1*	*Level 2*	*Level 3*	*Level 4*
3	2380 and below	**2381** - 2435	**2436** - 2500	**2501** and above
4	2410 and below	**2411** - 2484	**2485** - 2548	**2549** and above
5	2454 and below	**2455** - 2527	**2528** - 2578	**2579** and above
6	2472 and below	**2473** - 2551	**2552** - 2609	**2610** and above
7	2483 and below	**2484** - 2566	**2567** - 2634	**2635** and above
8	2503 and below	**2504** - 2585	**2586** - 2652	**2653** and above

Source: Delaware Department of Education (2019, September)

Table 5
Delaware Long-term Proficiency Goals

Subgroups	ELA Starting Point (2015-2016)	ELA Long-Term Goal (2030)	Mathematics Starting Point (2015-2016)	Mathematics Long-Term Goal (2030)
All students	52.09%	76.05%	40.49%	70.25%
Economically disadvantaged students*	35.60%	67.80%	25.42%	62.71%
Children with disabilities*	13.48%	56.74%	10.36%	55.18%
English learners	15.14%	57.57%	18.10%	59.05%
African American	36.19%	68.10%	23.39%	61.70%

Source: Delaware Department of Education (2019, June 10) *Delaware consolidated state plan under the Every Student Succeeds Act* Dover, DE: Author



Source: Lumos Learning *Understanding the SBAC Report Card*. Retrieved 1/9/2020 from https://www lumoslearning com/llwp/teachers-speak/understanding-the-sbac-report-card html

Figure 1.
Hypothetical Illustration of Range of Performance across Performance Levels

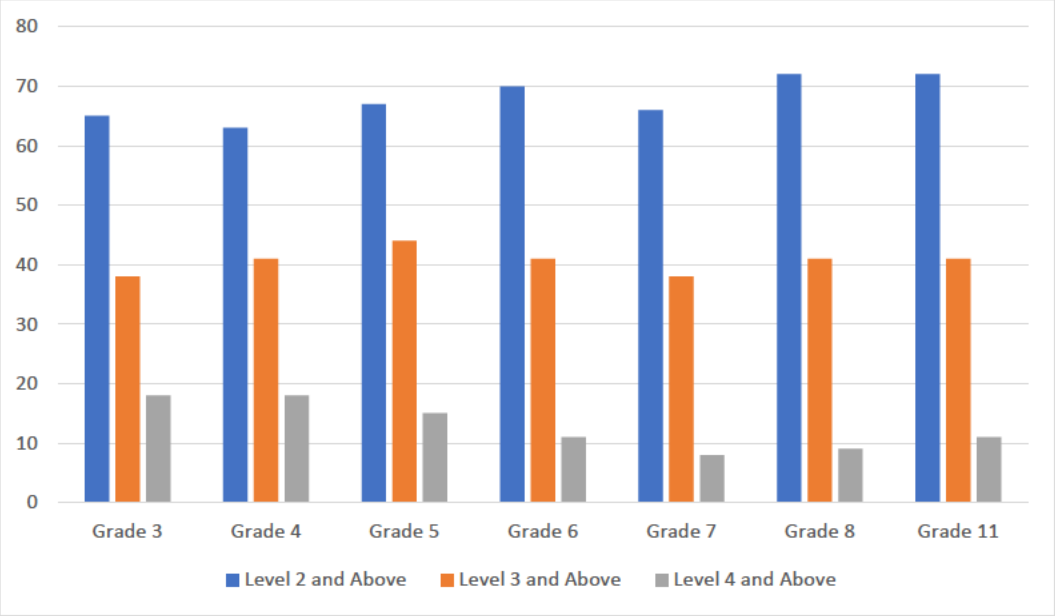


Figure 2.
Estimated Smarter Balanced ELA/L Performance by Grade, 2015

Source: Touchette, B M (2014, December 18) *Smarter achievement level setting: Presentation to the Delaware State Board of Education.* Dover, DE

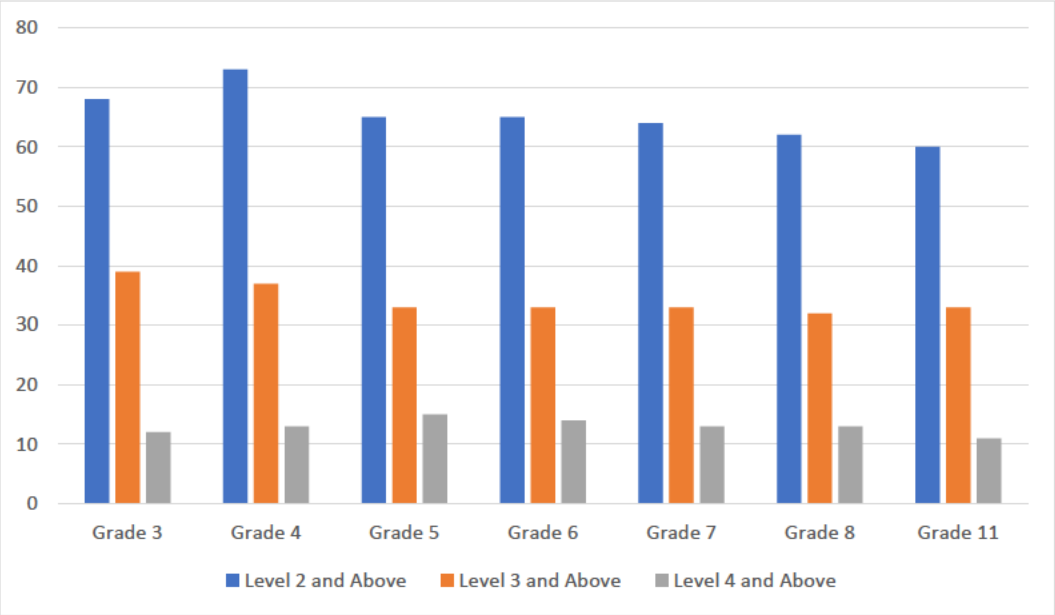


Figure 3.
Estimated Smarter Balanced Mathematics Performance by Grade, 2015

Source: Touchette, B M (2014, December 18) *Smarter achievement level setting: Presentation to the Delaware State Board of Education.* Dover, DE

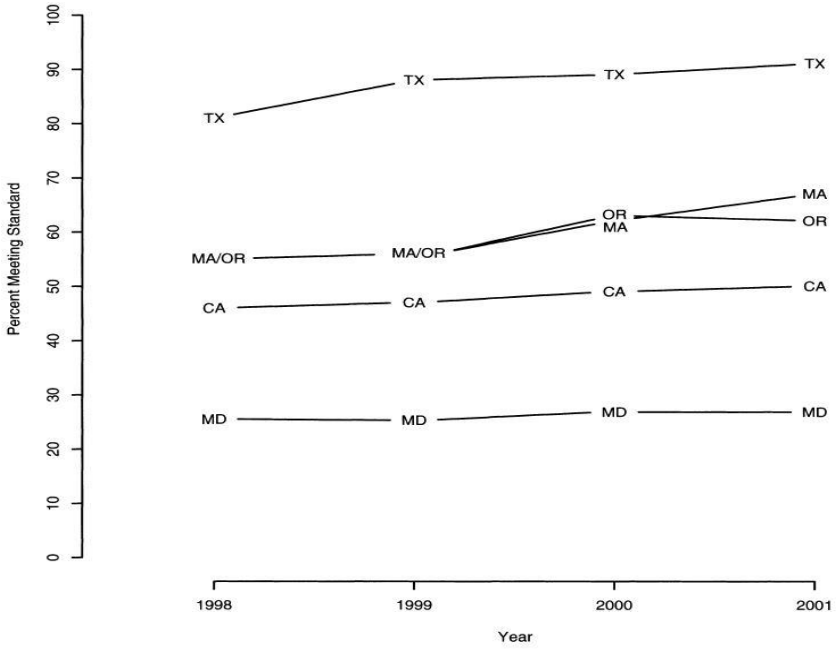


Figure 4.
Percentages Proficient in Reading in Six States, 1998-2001.

Source: Linn, Baker, & Betebenner (2002)

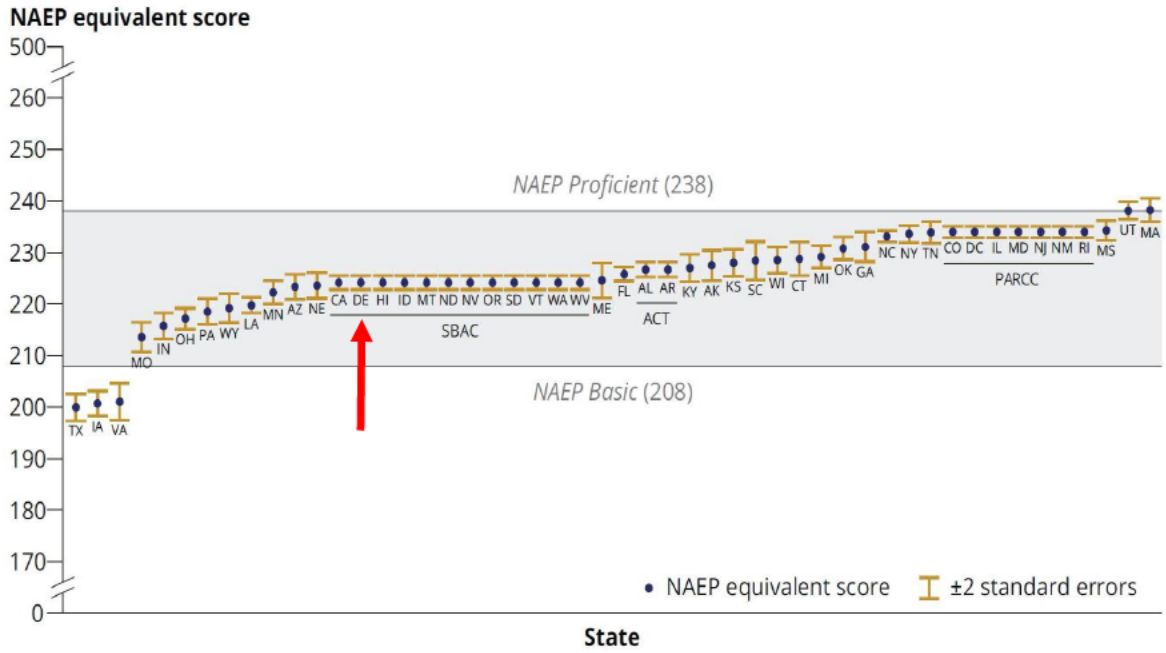


Figure 5.
NAEP Mapping Study Results, 2017, Grade 4 Reading

Source: Bandeira de Mello, V , Rahman, T , Fox, M A , & Ji, C S (2019) *Mapping state proficiency standards onto the NAEP scales: Results from the 2017 NAEP reading and mathematics assessments (NCES 2019-040)* U S Department of Education Washington, DC: Institute of Education Sciences, National Center for Education Statistics

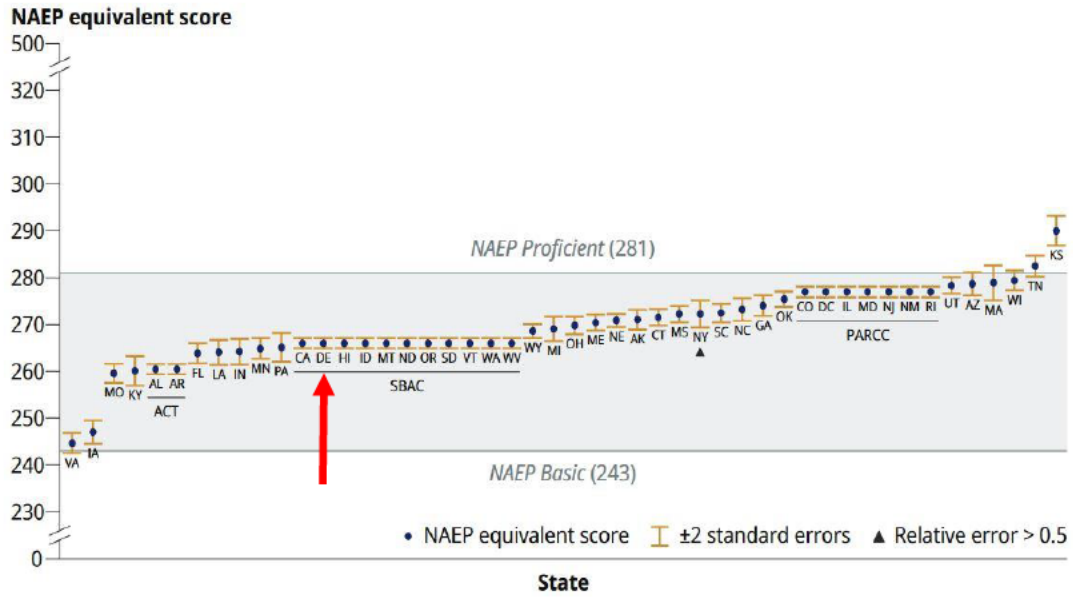


Figure 6.
NAEP Mapping Study Results, 2017, Grade 8 Reading

Source: Bandeira de Mello, V., Rahman, T., Fox, M.A., & Ji, C. S. (2019) *Mapping state proficiency standards onto the NAEP scales: Results from the 2017 NAEP reading and mathematics assessments (NCES 2019-040)* U.S. Department of Education. Washington, DC: Institute of Education Sciences, National Center for Education Statistics

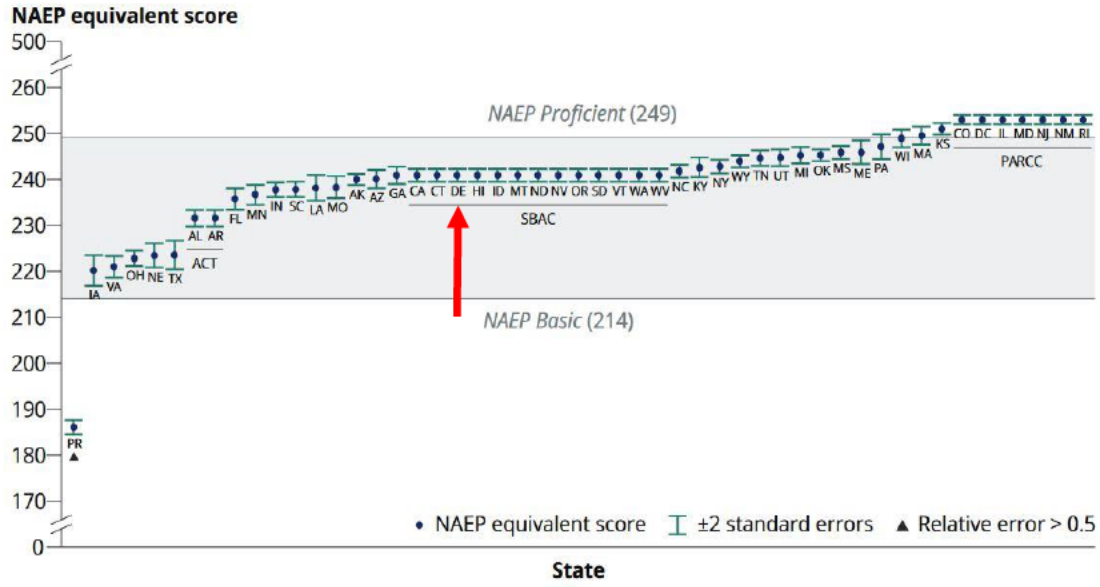


Figure 7.
NAEP Mapping Study Results, 2017, Grade 4 Mathematics

Source: Bandeira de Mello, V., Rahman, T., Fox, M.A., & Ji, C.S. (2019). *Mapping state proficiency standards onto the NAEP scales: Results from the 2017 NAEP reading and mathematics assessments (NCES 2019-040)*. U.S. Department of Education. Washington, DC: Institute of Education Sciences, National Center for Education Statistics.

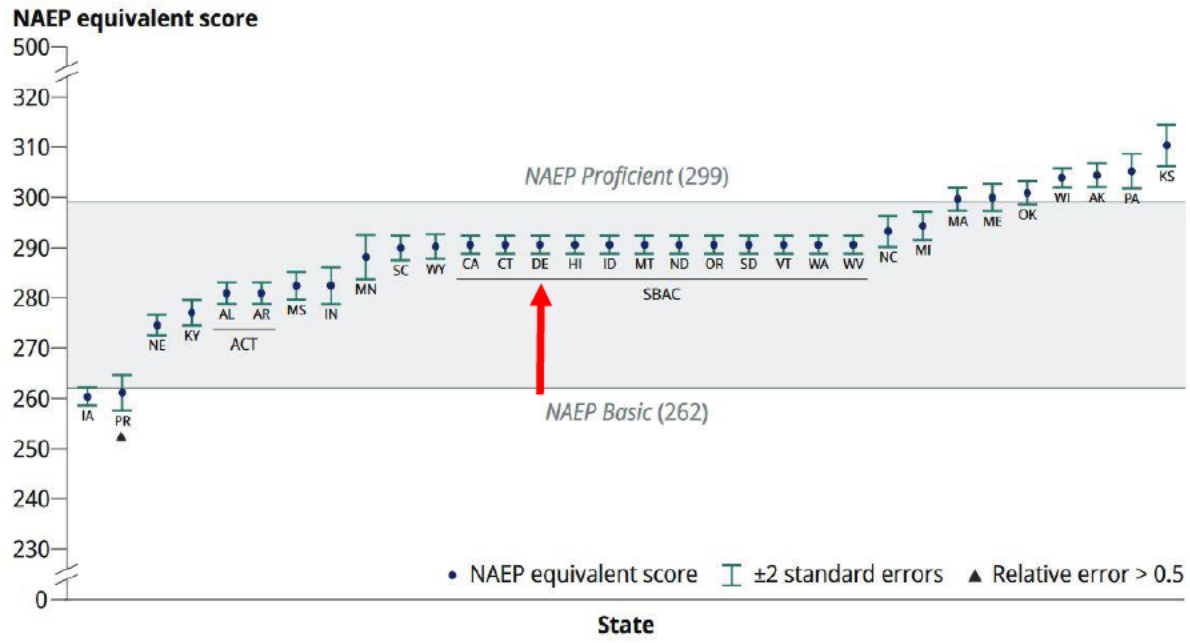


Figure 8.
NAEP Mapping Study Results, 2017, Grade 8 Mathematics

Source: Bandeira de Mello, V., Rahman, T., Fox, M.A., & Ji, C.S. (2019) *Mapping state proficiency standards onto the NAEP scales: Results from the 2017 NAEP reading and mathematics assessments (NCES 2019-040)* U.S. Department of Education. Washington, DC: Institute of Education Sciences, National Center for Education Statistics

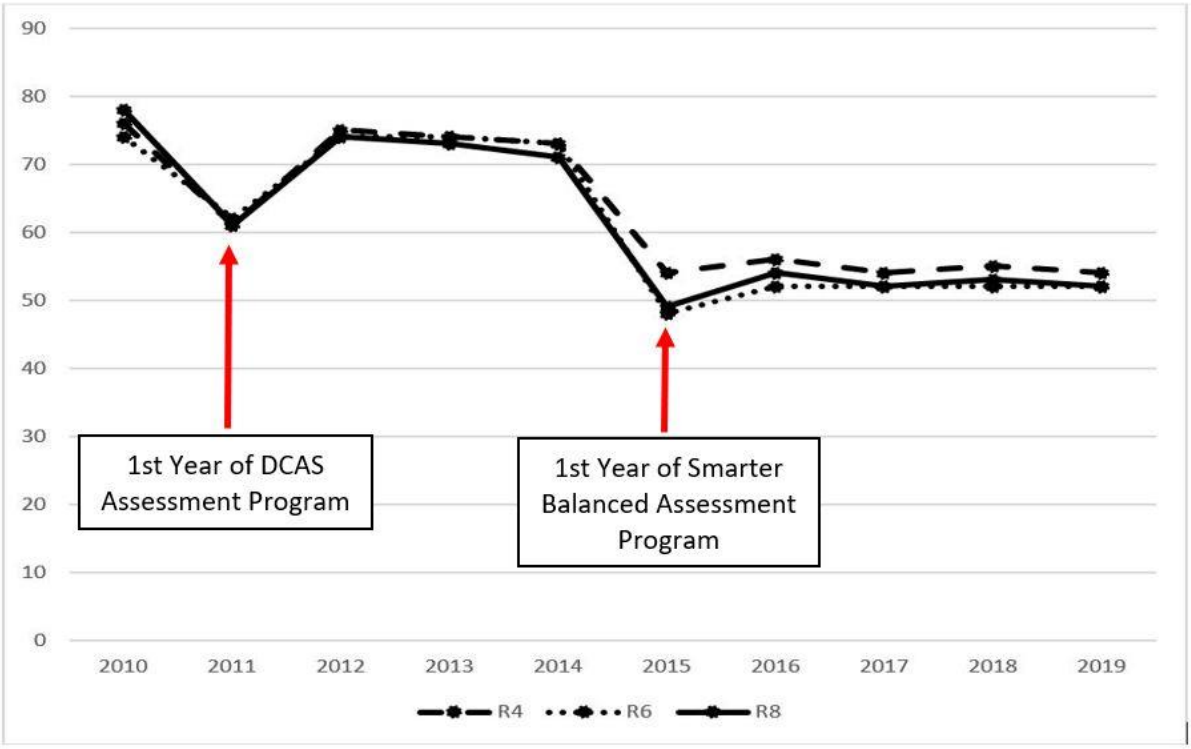


Figure 9.
Delaware Percentages Level 3 and Above, Reading, Grades 4, 6, and 8, 2009-2010 through 2018-2019

Sources: Delaware Department of Education (2010, 2011, 2012, 2014, 2019, September); State of Delaware (2019)

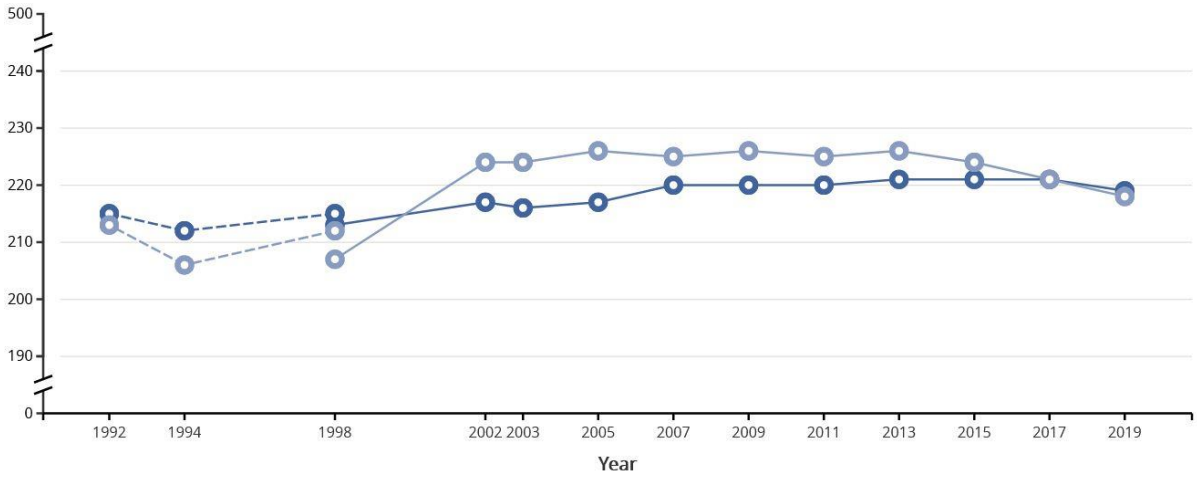


Figure 10.
NAEP Results for U.S. Public Schools and Delaware, Grade 4 Reading

Source: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore)

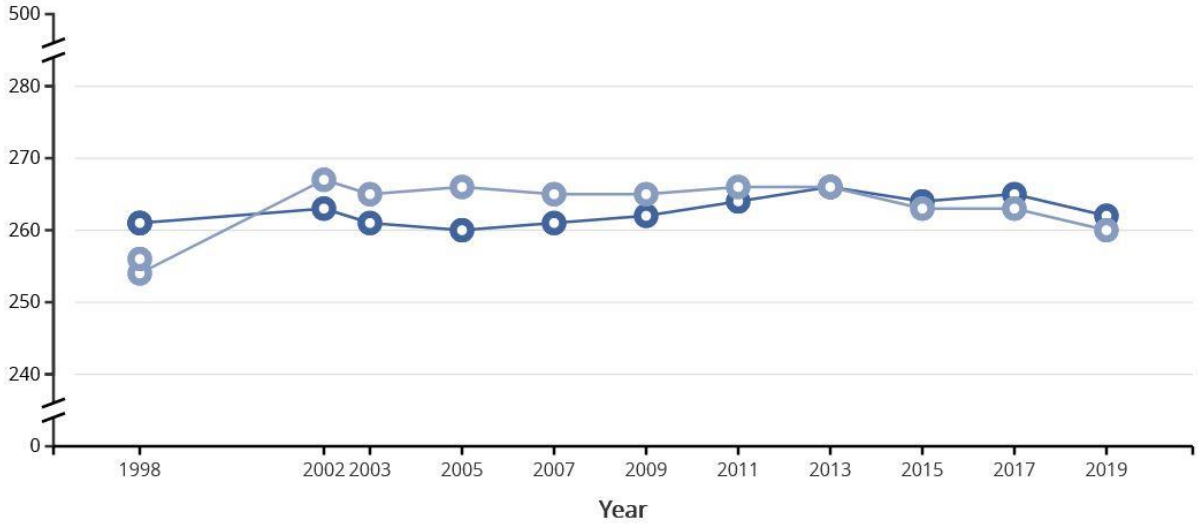


Figure 11.
NAEP Results for U.S. Public Schools and Delaware, Grade 8 Reading

Source: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore)

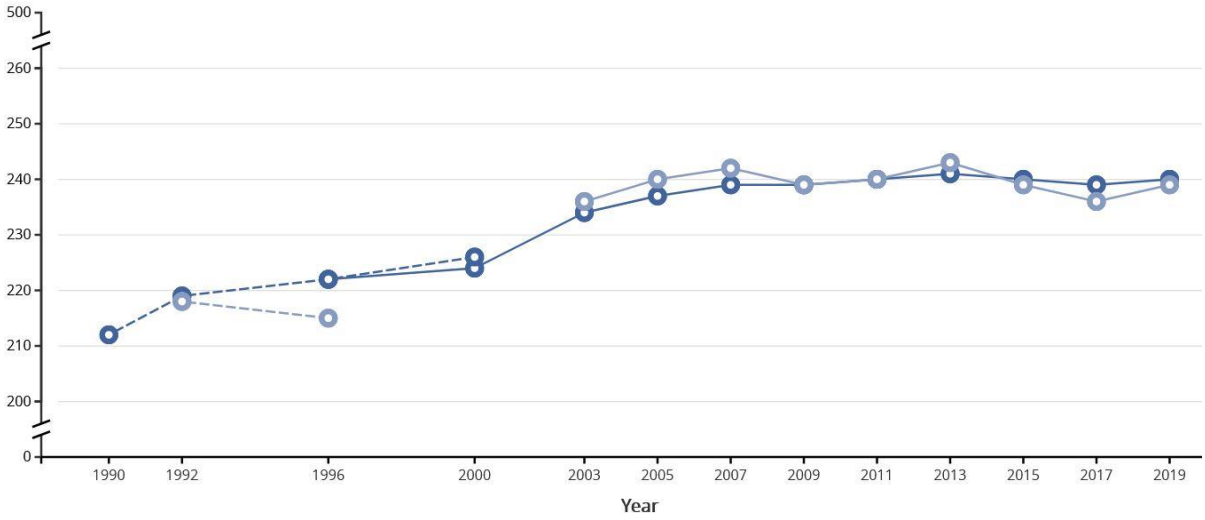


Figure 12.
NAEP Results for U.S. Public Schools and Delaware, Grade 4 Mathematics

Source: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore)

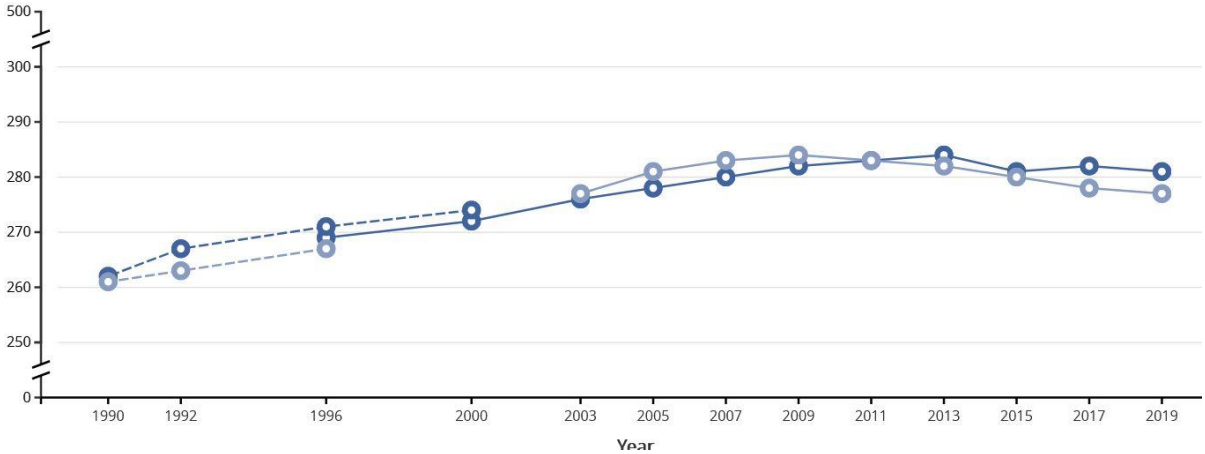


Figure 13.
NAEP Results for U.S. Public Schools and Delaware, Grade 8 Mathematics

Source: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore)

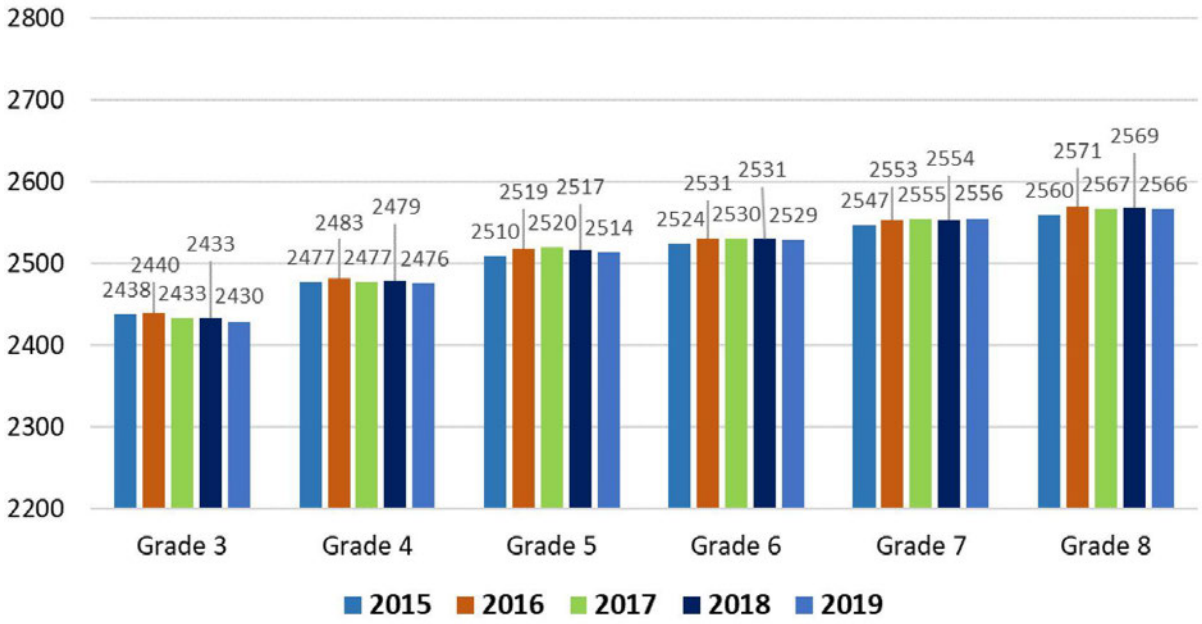


Figure 14.
Delaware Smarter Balanced Assessment Results, ELA/Literacy, 2015-2019

Source: Delaware Department of Education (2019, September) *Delaware System of Student Assessments (DeSSA) executive state summary, 2018-2019 administration.* Dover, DE: Author

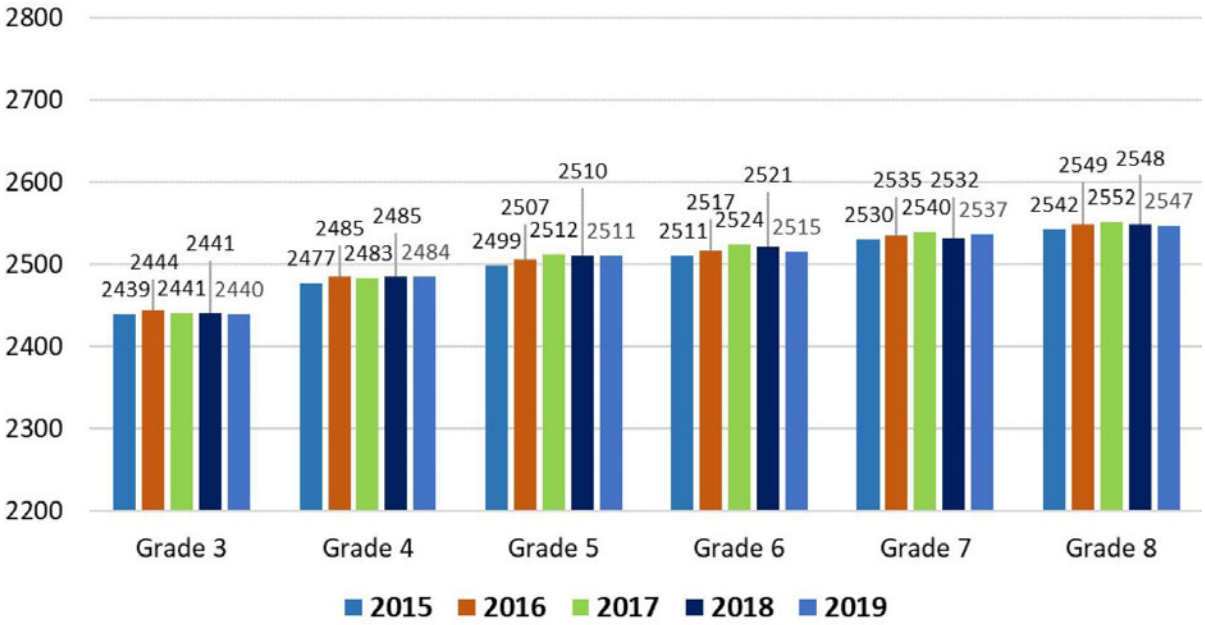
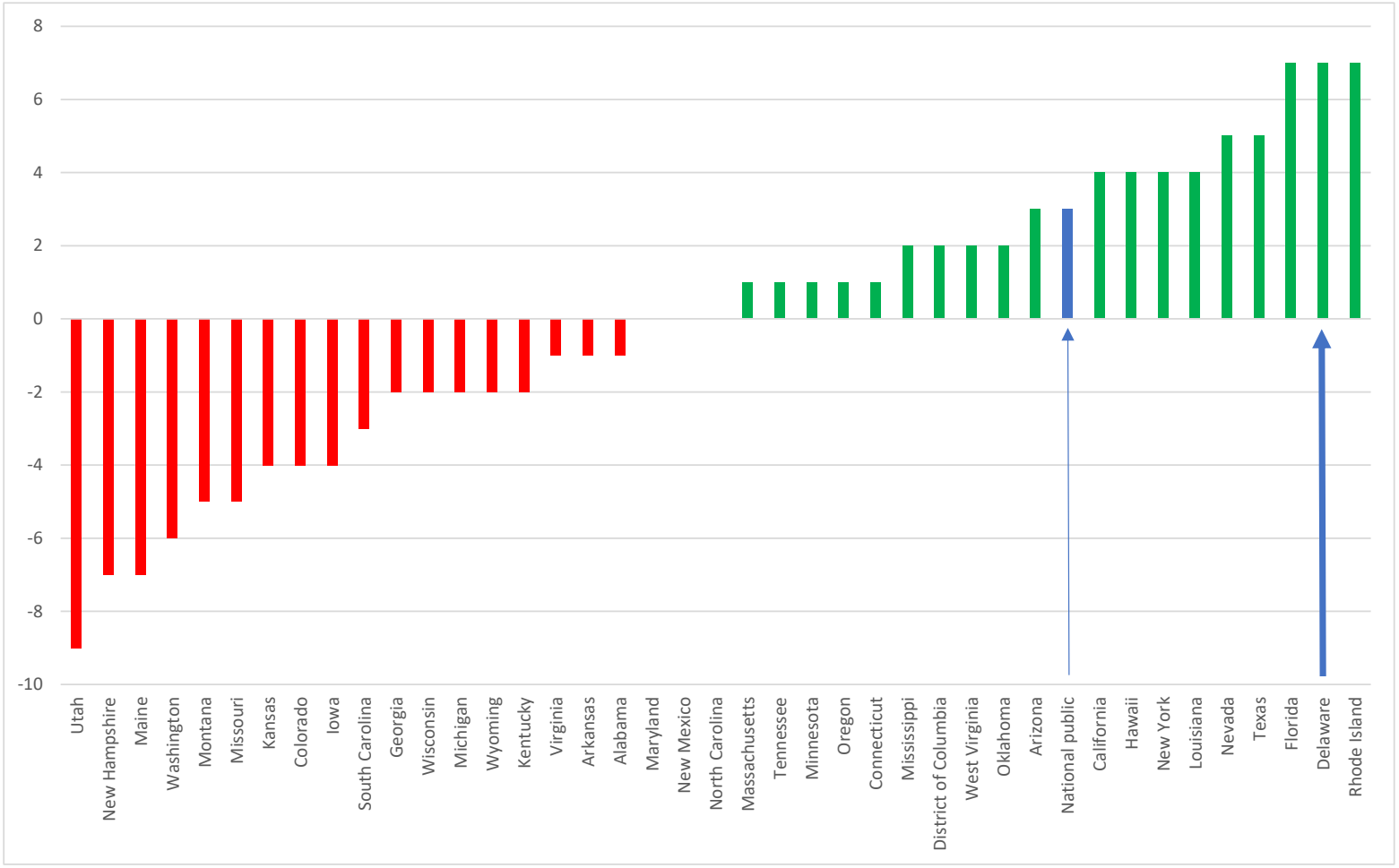


Figure 15.
Delaware Smarter Balanced Assessment Results, Mathematics, 2015-2019

Source: Delaware Department of Education (2019, September) *Delaware System of Student Assessments (DeSSA) executive state summary, 2018-2019 administration.* Dover, DE: Author

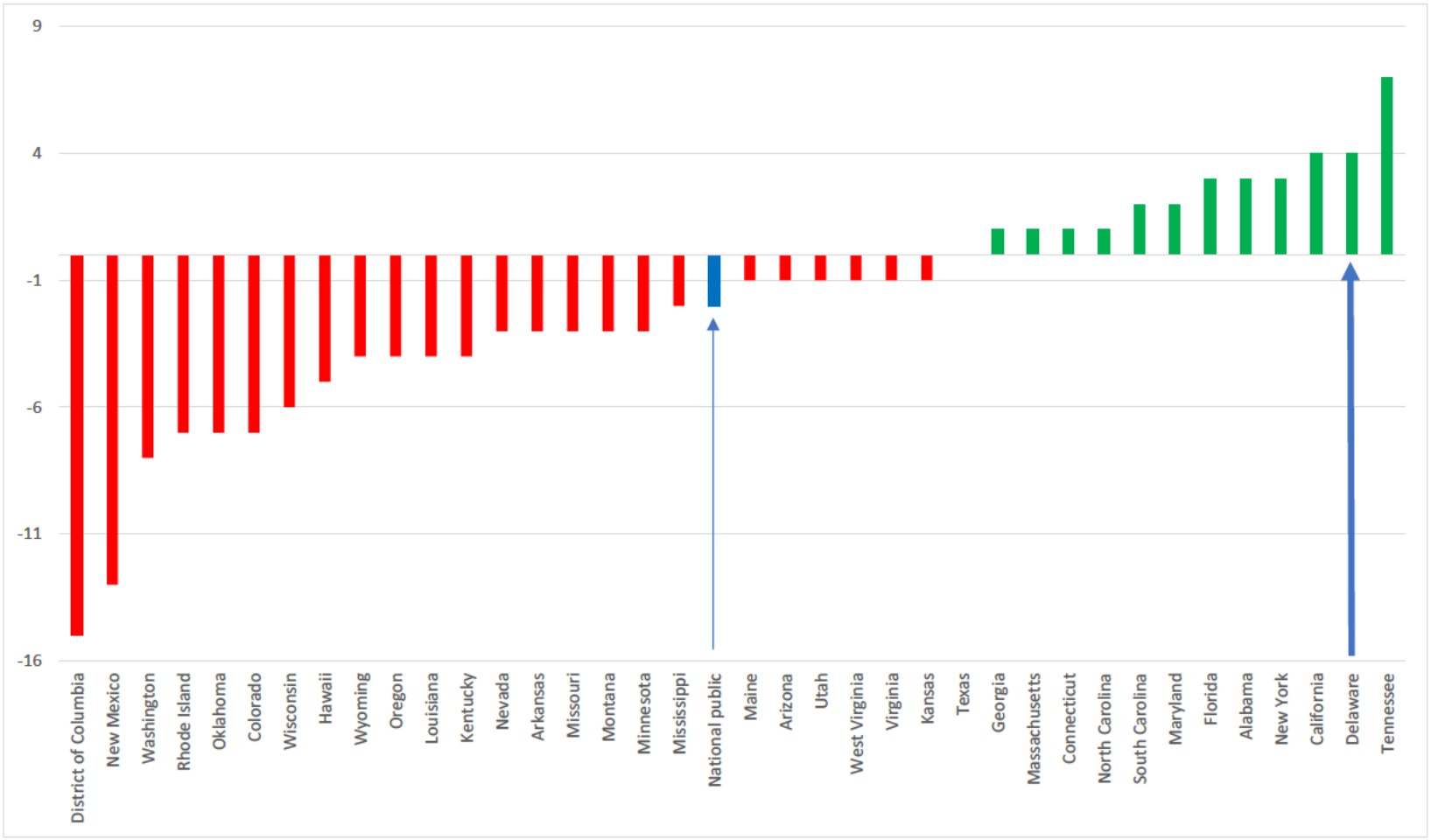
Figure 16
Scaled Score Gap Reduction, NSLP Eligible/Non-Eligible, NAEP Reading, Grade 4, 1998-2019



S

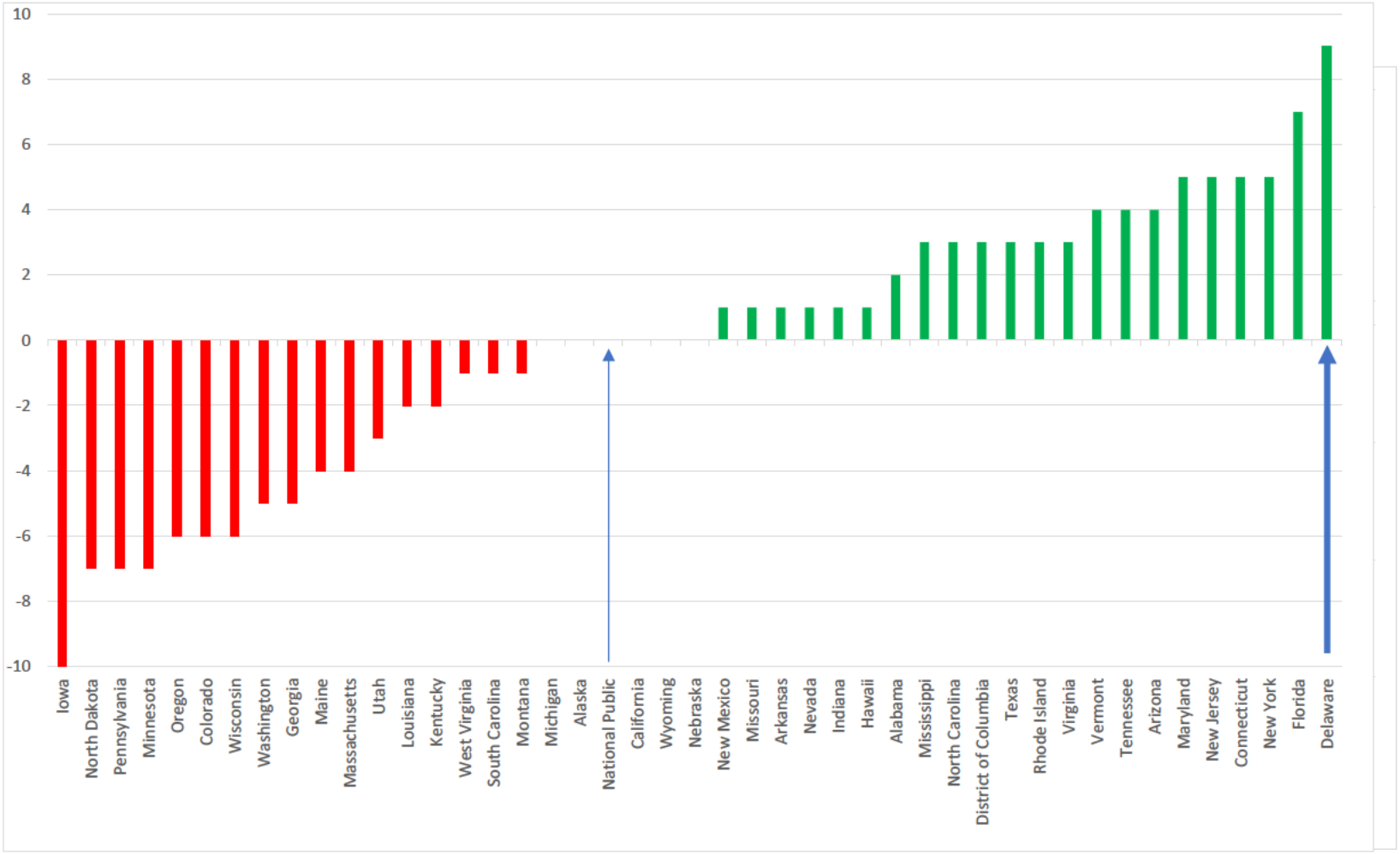
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 17
Scaled Score Gap Reduction, NSLP Eligible/Non-Eligible, NAEP Reading, Grade 8, 1998-2019



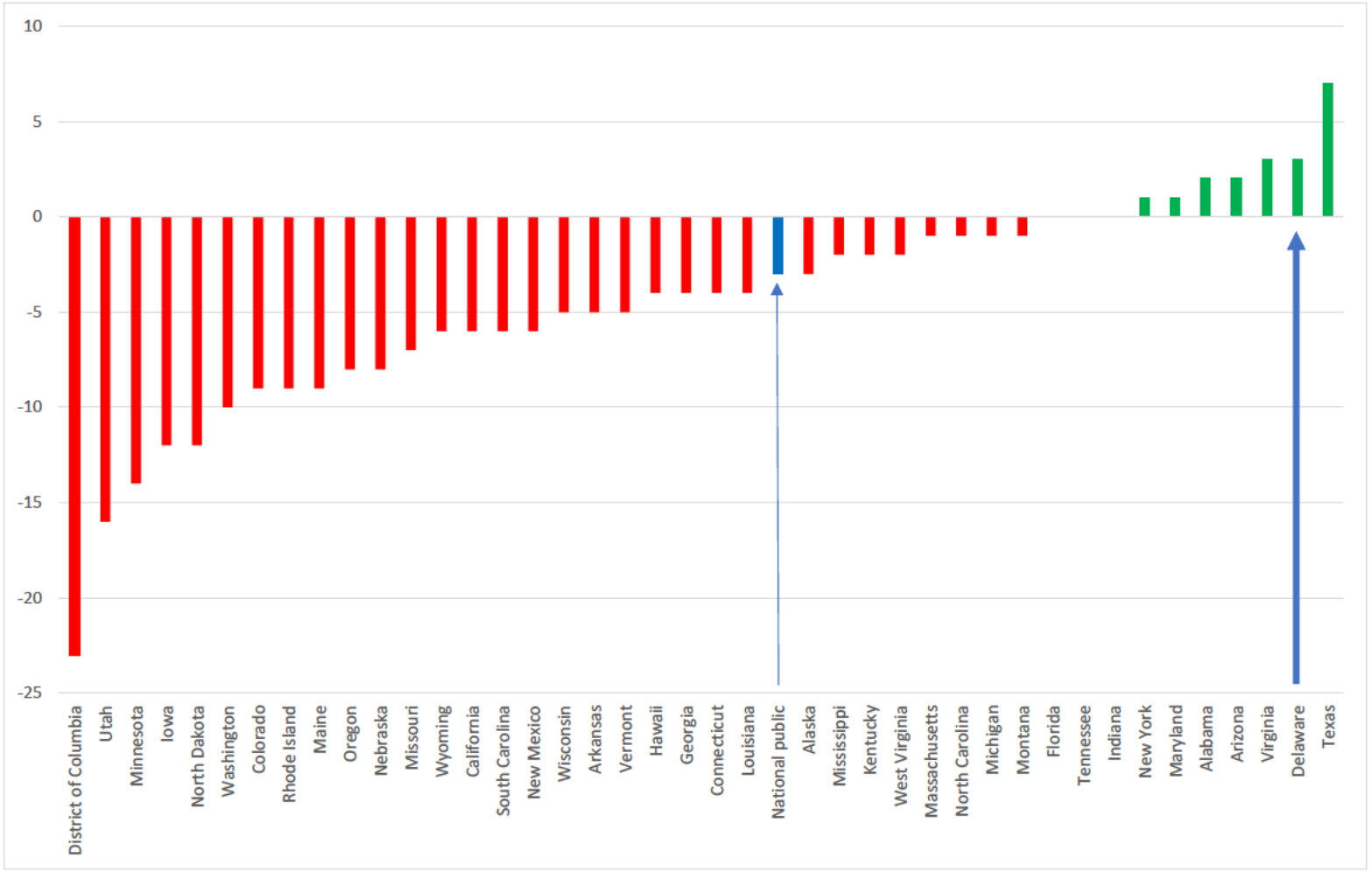
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 18
Scaled Score Gap Reduction, NSLP Eligible/Non-Eligible, NAEP Mathematics, Grade 4, 1996-2019



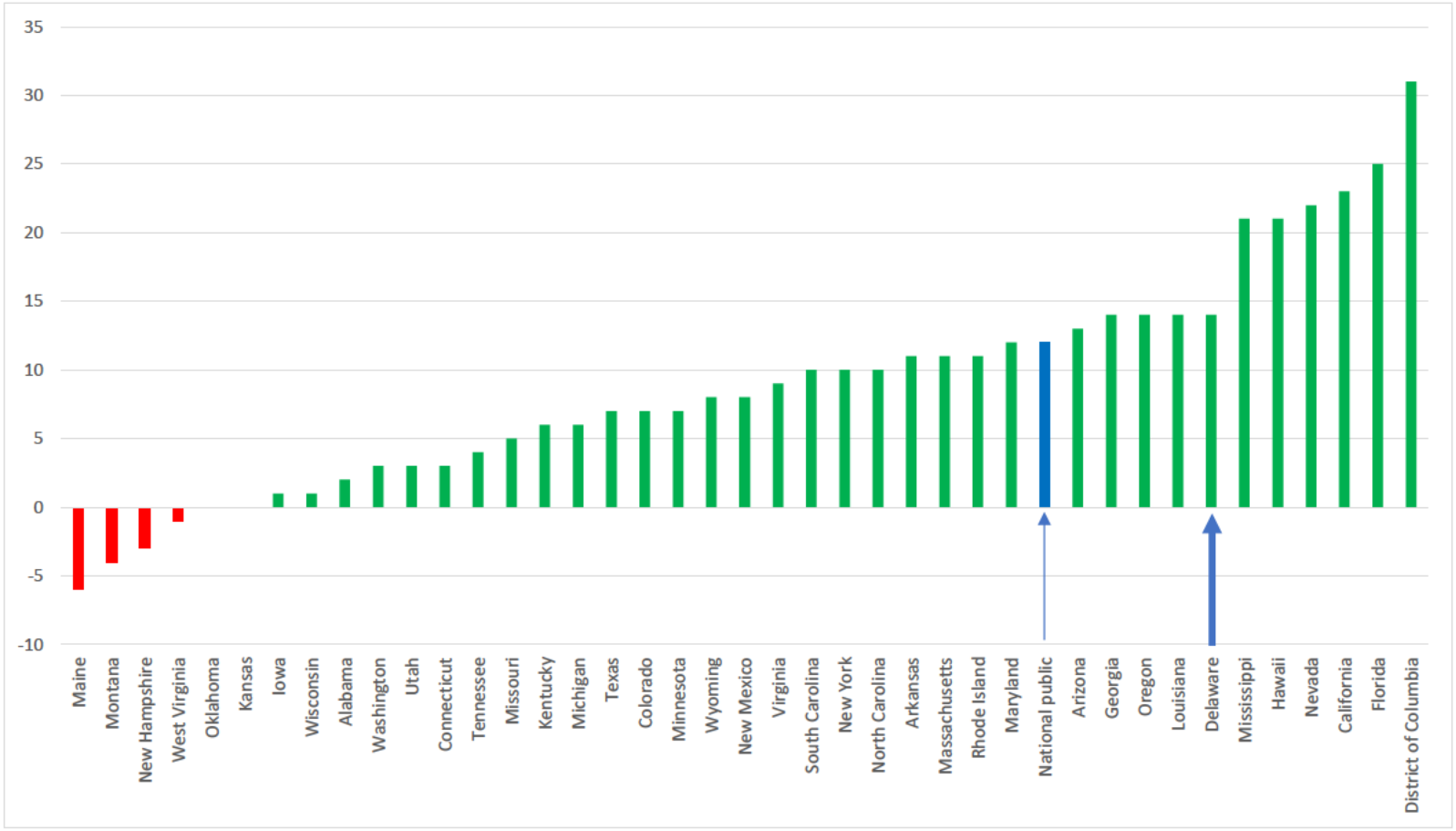
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 19
Scaled Score Gap Reduction, NSLP Eligible/Non-Eligible, NAEP Mathematics, Grade 8, 1996-2019



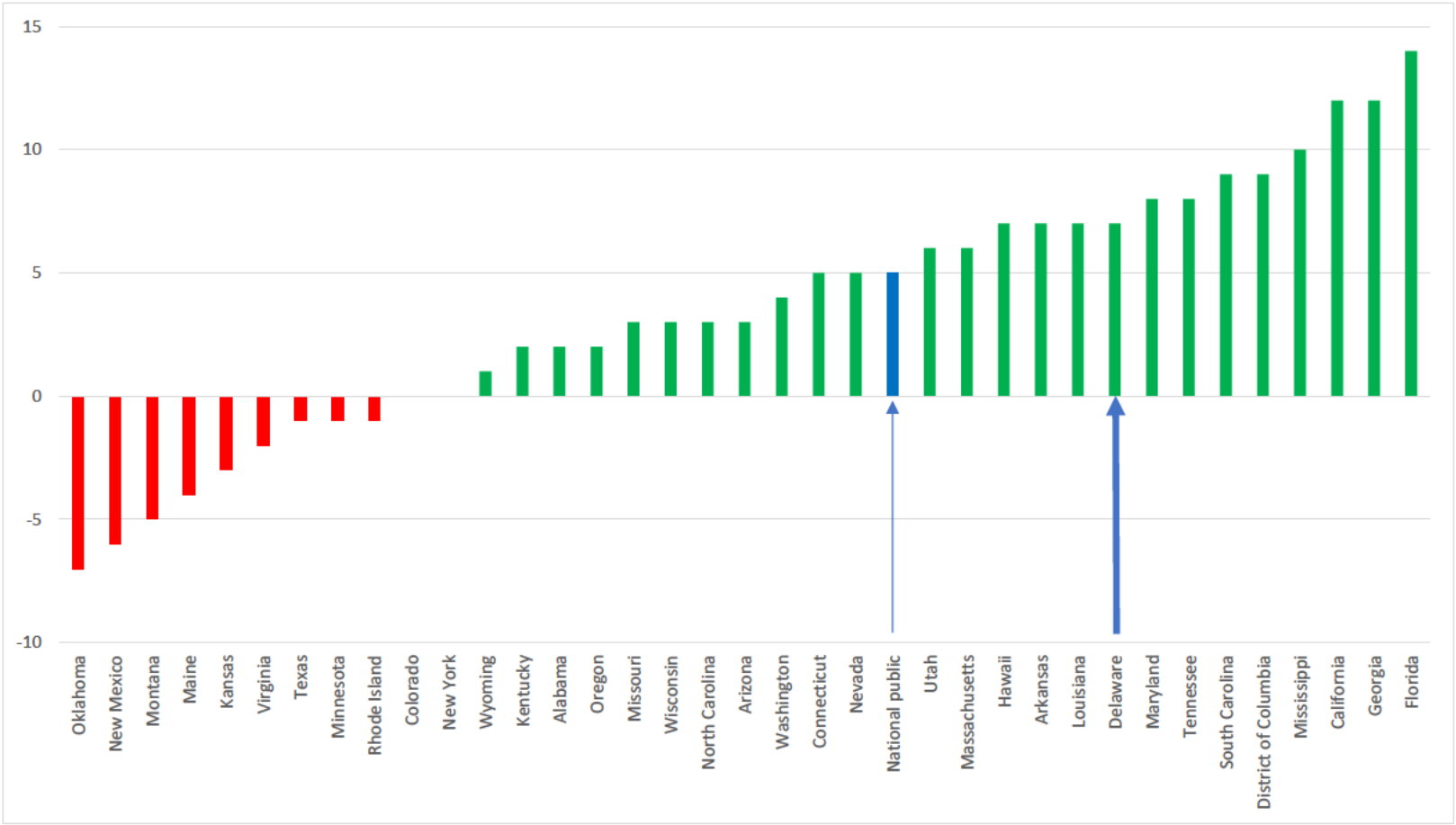
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 20
Scaled Score (Achievement) Increases, NSLP Eligible Students, NAEP Reading, Grade 4, 1998-2019



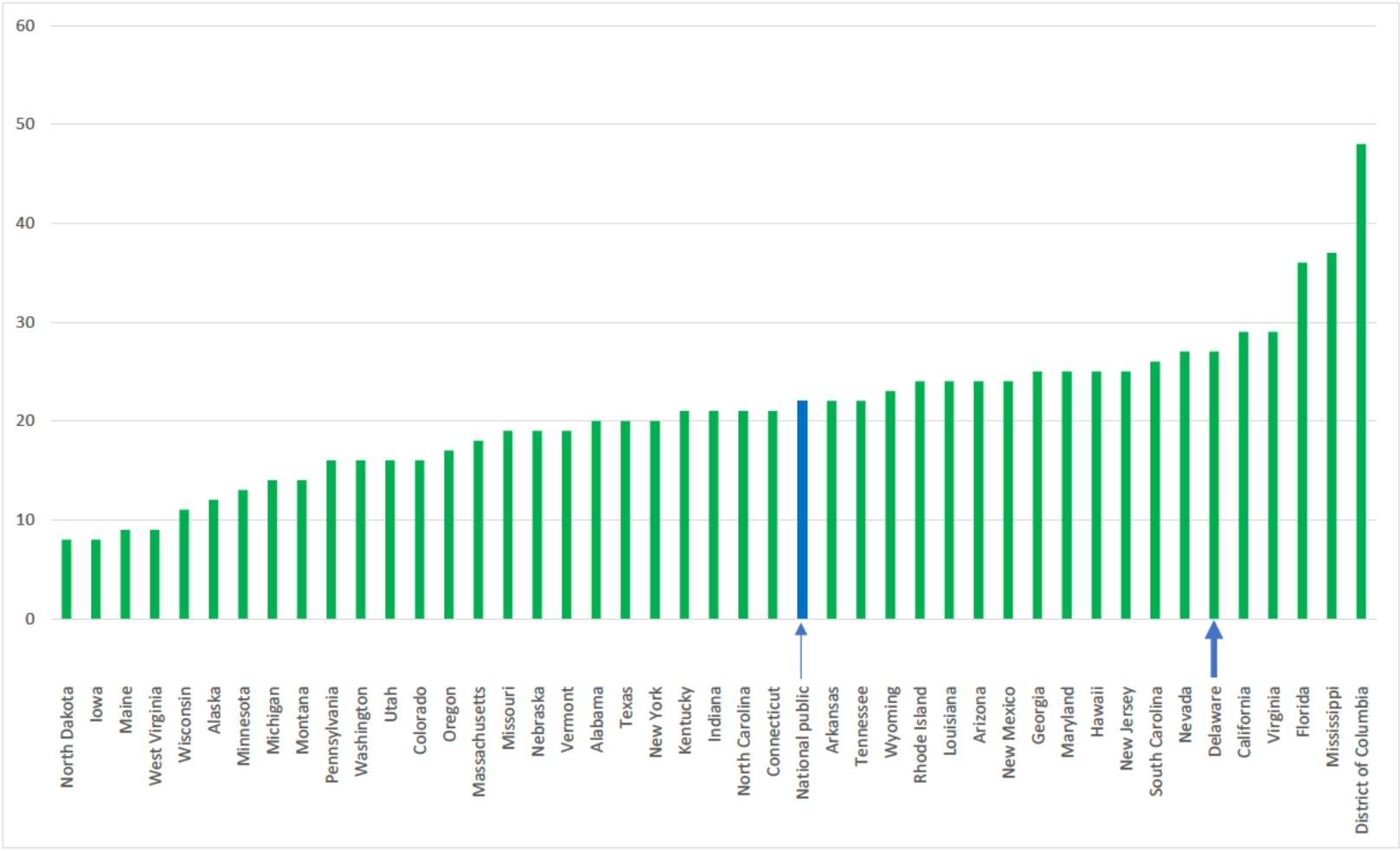
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 21
Scaled Score (Achievement) Increases, NSLP Eligible Students, NAEP Reading, Grade 8, 1998-2019



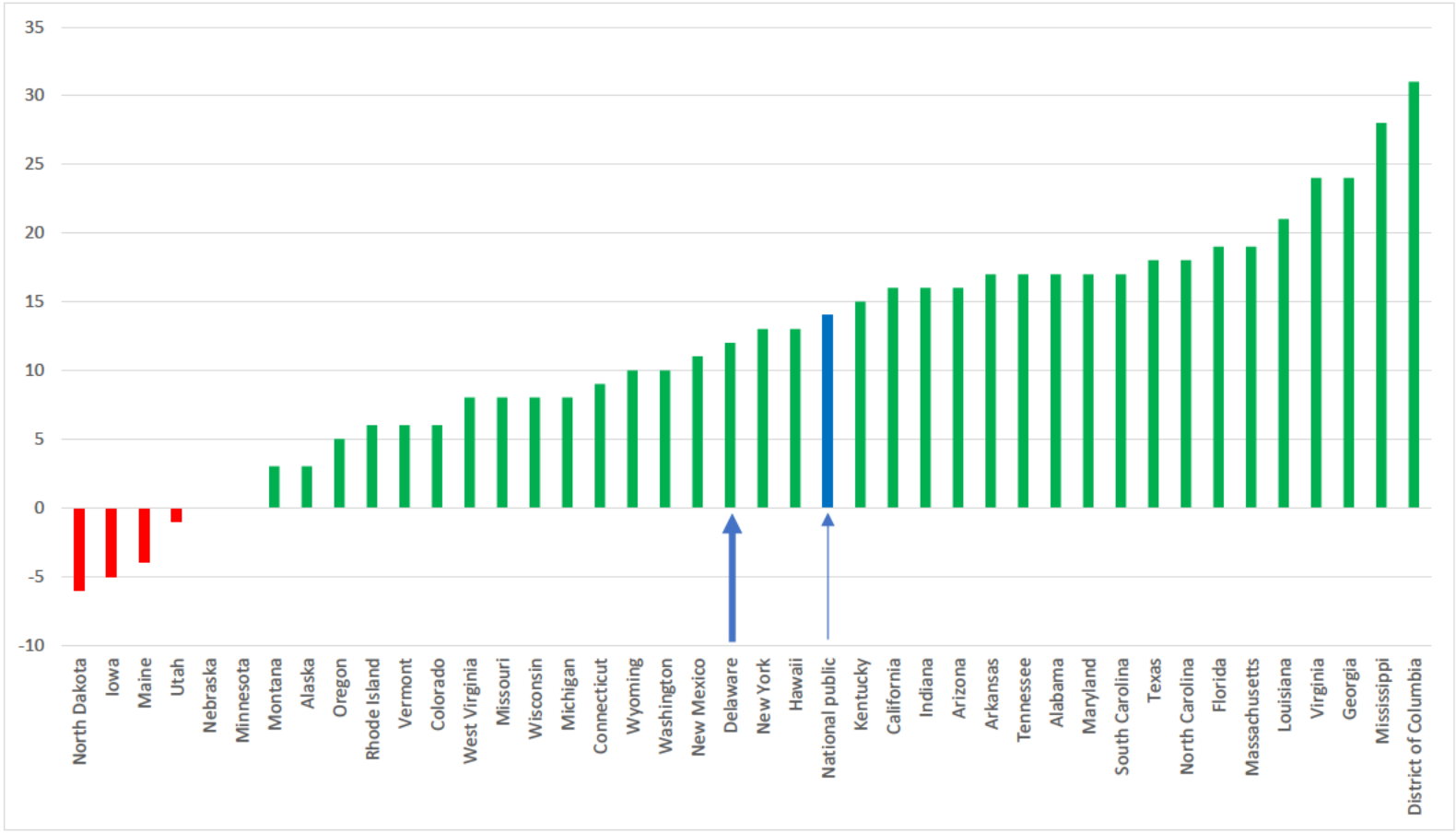
SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 22
Scaled Score (Achievement) Increases, NSLP Eligible Students, NAEP Mathematics, Grade 4, 1996-2019



SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

Figure 23
Scaled Score (Achievement) Increases, NSLP Eligible Students, NAEP Mathematics, Grade 8, 1996-2019



SOURCE: United States Department of Education, National Center for Education Statistics (2019) NAEP Data Explorer (www nationsreportcard gov/ndecore) Note: Not all jurisdictions are included due to missing data, insufficient sample sizes, or other technical reasons determined by the National Center for Education Statistics

APPENDIX C
Resume of Gregory J. Cizek

GREGORY J. CIZEK

OFFICE ADDRESS:

112 Peabody Hall
School of Education, CB 3500
University of North Carolina
Chapel Hill, NC 27599-3500

Telephone: (919) 843-7876
Fax: (919) 962-1533
Email: cizek@unc.edu

EDUCATION

Ph.D.	1991 - Michigan State University; East Lansing, MI Major: Measurement, Evaluation, and Research Design Emphasis Areas: Educational Measurement, Quantitative Research Methods, Educational Policy
M.A.	1983 - Michigan State University; East Lansing, MI Major: Curriculum and Instruction Emphasis Areas: Curriculum, Learning Theory
B.A.	1979 - Michigan State University; East Lansing, MI Major: Elementary Education Emphasis Areas: Social Studies major, Math/Science minor
A.A.	1977 - Northwestern Michigan College; Traverse City, MI Major: Journalism

CERTIFICATES AWARDED

1979 - Michigan Elementary Provisional Teaching Certificate (K-9)
1983 - Michigan Continuing Elementary Teaching Certificate (K-9)
1987 - Iowa Professional Elementary Certificate
1988 - Iowa Elementary Principal Certificate
1992 - Ohio Elementary Teacher Certificate (K-9)

PROFESSIONAL EXPERIENCES

Professor of Educational Measurement and Evaluation (tenured) - School of Education, University of North Carolina at Chapel Hill (2002-present). Guy B. Phillips Distinguished Professor (2012-present).

Associate Professor of Educational Measurement and Evaluation (tenured) - School of Education, University of North Carolina at Chapel Hill (1999-2001).

Associate Professor of Educational Research and Measurement (tenured) - College of Education, University of Toledo, Toledo, OH (1995-1999).

Assistant Professor of Educational Research and Measurement - College of Education, University of Toledo, Toledo, OH (1991-1995). Taught graduate courses in testing, psychometric theory, statistics, and research design; provided research and measurement consultation to other faculty; advised graduate students; served on departmental and college-level committees; conducted research in areas of specialization and developed service relationships with local school districts

Program Manager - American College Testing Program, Inc. (ACT); Iowa City, IA (1987-1991). Provided measurement expertise and program management for medical and allied health licensure and certification programs. Responsibilities included test development, client consultation, statistical analyses of test results, supervision of item writing, equating examination forms, conducting client meetings, preparing and delivering instructional seminars to health professions groups, conducting passing score studies, scheduling, proposal writing, budget monitoring, and consultation with other ACT departments

Test Development Consultant - Michigan Educational Assessment Program (MEAP); Lansing, MI (1985-1986). Assisted in development of objectives-based, state-wide, every-pupil testing in Michigan State Board of Education approved subjects. Responsibilities included: participation in Reading and Career Development test development projects; supervision of item writing teams; item editing; psychometric consultation; participation in scoring, formatting, and reporting decisions; composing Requests for Proposals and evaluation of responses.

Educational Policy Researcher - Michigan Senate Policy and Programs Office; Lansing, MI (1985). Performed research, analyses, policy option development and recommendations on pending/proposed legislation on education issues (emphases on student testing, dropout prevention, liability insurance, teacher shortages, home schooling, and teacher competency testing).

Statistics Instructor - Michigan State University; East Lansing, MI (1987, 1988, 1991, 1994). Taught graduate-level courses in statistics and educational research methods on the MSU campus, in Thailand, in Brazil, and in the Philippines.

Educational Psychology Instructor - Michigan State University; East Lansing, MI (1984-1987). Taught undergraduate educational psychology course.

Elementary and Middle School Teacher – Traverse City, MI (1979-1984). Taught grades 2, 4, 4/5, and 6/7. Taught all subjects in grades 2 and 4; computer science and physical education for grades 6 and 7.

COMMITTEE WORK, PROFESSIONAL MEMBERSHIPS, SERVICE, AND HONORS

* Vice-President, President, Past President, National Council on Measurement in Education (2011-2014),
* Member, American Educational Research Association, 1986-2019
* Member, National Council for Measurement in Education, 1986-present
* Member, North Carolina Association for Research in Education, 2000-2002
* Secretary, Professional Licensure and Certification-SIG, 1999-2002
* Secretary, AERA Division D (Measurement and Research Methodology) 2004-2006
* Member, program participant, Classroom Assessment-SIG (AERA), 1995-2007
* Member, program participant, Families as Educators-SIG (AERA), 1987-1992; program committee, 1988, 1990-1991
* Proposal Reviewer, NCME, AERA Divisions D, H, I, J, and various Special Interest Groups, 1991-present

* Manuscript Reviewer, *American Educational Research Journal*
* Manuscript Reviewer, *Applied Measurement in Education*
* Manuscript Reviewer, *Contemporary Educational Psychology*
* Manuscript Reviewer, *Educational Assessment*
* Manuscript Reviewer, *Educational Evaluation and Policy Analysis*
* Manuscript Reviewer, *Educational Measurement: Issues and Practice*
* Manuscript Reviewer, *Educational Policy*
* Manuscript Reviewer, *Educational Researcher*

Committee Work, Professional Memberships, Service, and Honors (cont'd)

* Manuscript Reviewer, *Home School Researcher*
* Manuscript Reviewer, *Journal of Educational Measurement*
* Manuscript Reviewer, *Journal of Educational Psychology*
* Manuscript Reviewer, *Review of Educational Research*
* Manuscript Reviewer, *Applied Psychological Measurement*

* Member, Chair, AERA E. F. Lindquist Award Committee, 2008-2010
* Member, Chair, AERA Palmer O. Johnson Award Committee, 2006-2008
* Member, AERA Outreach and Professional Liaison Committee, 2001-2004
* Member, AERA Review of Research Award Committee, 1991-1993
* Member, AERA Publications Committee, 1992-1995
* Member, NCME Nominations Committee, 2001-2002
* Member, NCME Standards and Test Use Committee, 1995-1999
* Member, NCME Career Award Committee, 1999-2000
* Member, NCME Alicia Cascallar Dissertation Award Committee, 2017-present
* Chair, NCME Standards and Test Use Committee, 2008-2011
* Co-Chair, Joint Committee on Testing Practices, 2003-2006

* Editorial Board, *Journal of Educational Measurement*, 2003-2008
* Editorial Board, *Educational Researcher*, 2007-2009
* Editorial Board, *Educational Measurement: Issues and Practice*, 2003-2008
* Editorial Board, *Applied Measurement in Education,* 2001-2014
* Editorial Board, *Educational and Psychological Measurement*, 2004-2010
* Editorial Board, *Home School Researcher*, 1990-present
* Editorial Board, *NCME Newsletter*, 1999-2003
* Editor, special issue on vertically moderated standard setting, *Applied Measurement in Education*, 2005

* Member, Smarter/Balanced Assessment Consortium Technical Advisory Panel, 2010-
* Member, National Assessment Governing Board (NAGB), 2007-2009; 2017-present
* Member, United States Department of Education, Advisory Council on Education Statistics, 2001-2003
* Member, United States Department of Education, National Technical Advisory Council, 2008-2009
* Member, National Advisory Committee, Buros Institute of Mental Measurement, 2004-present
* Contributor, *Standards for Educational Evaluation of Students* (Joint Committee on Standards for Educational Evaluation)
* Member, NCME Ad-Hoc Committee on ITEMS publication, 2000-2002.
* Elected Member, Vice-President, Sylvania (OH) Public Schools Board of Education, 1997-1999
* Advisory Board, National Home Education Research Institute, 1990-present
* Member, North American Business Research Advisory Board, University of Toledo, 1996-1999

* Awardee, Buros Institute for Mental Measurement, *Distinguished Reviewer Award,* 2005
* Awardee, AERA Division D award for *Significant Contribution to Educational Measurement and Research Methodology*, 2006
* Awardee, NCME Award for *Outstanding Dissemination of Educational Measurement Concepts*, 2007

BOOKS, BOOK CHAPTERS, and MONOGRAPHS

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use.* New York, NY: Routledge.

Andrade, H., Bennett, R., & Cizek, G. J. (Eds.) (2019). *Handbook of formative assessment in the disciplines*. New York: Routledge/Taylor and Francis.

Cizek, G. J. (2019). Formative assessment: History, definition, and progress. In H. Andrade, R. Bennett, & G. J. Cizek (eds.), *Handbook of formative assessment in the disciplines*. New York: Routledge/Taylor and Francis.

Wollack, J. A., & Cizek, G. J. (2017). Test security for licensure and certification examination programs. In S. Davis-Becker & C. Buckendahl (eds.), *Testing in the professions* (pp. 178-209). New York: Routledge/Taylor & Francis.

Cizek, G. J., & Wollack, J. A. (2017). *Quantitative methods for detecting cheating on tests.* New York: Routledge/Taylor & Francis.

Cizek, G. J. & Wollack, J. A. (2017). Exploring cheating on tests: The contexts, concerns, and challenges. In G. J. Cizek & J. A. Wollack (eds.), *Quantitative methods for detecting cheating on tests* (pp. 3-20). New York: Taylor and Francis.

Wollack, J. A., & Cizek, G. J. (2017). The future of quantitative methods for detecting cheating: Conclusions, cautions, and recommendations. In G. J. Cizek & J. A. Wollack (eds.), *Quantitative methods for detecting cheating on tests* (pp. 390-400). New York: Taylor and Francis.

Pitoniak, M. J., & Cizek, G. J. (2016). Standard setting. In C. Wells & M. Faulkner-Bond (eds.), *Educational measurement: From foundations to future* (pp. 38-61). New York: Guilford.

Cizek, G. J., & Earnest, D. (2015). Standard setting. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (pp. 212-237). New York: Routledge

Cizek, G. J. (Ed.) (2012a). *Setting performance standards: Foundations, methods, and innovations.* New York: Routledge.

Cizek, G. J. (2012b). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (pp. 3-14). New York: Routledge.

Cizek, G. J. (2012c). The forms and functions of evaluations of the standard setting process. In G. J. Cizek (Ed.) Setting performance standards: Foundations, methods, and innovations (pp. 165-178). New York: Routledge.

Mehrens, W. A., & Cizek, G. J. (2012). Standard setting for decision making: Classifications, consequences, and the common good. In G. J. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (pp. 33-46). New York: Routledge.

Cizek, G. J., & Agger, C. A. (2012). Vertically moderated standard setting. In G. J. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (pp. 467-484). New York: Routledge.

Books, Book Chapters, and Monographs (continued)

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (pp. 181-200) New York, NY: Routledge.

Cizek, G. J., Rosenberg, S. L., & Koons, H. K. (2011). Finding validity evidence: An analysis using *Mental Measurements Yearbook*. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High stakes testing in education: Science and practice in K-12 settings* (pp. 119-138). Washington, DC: APA Books.

Cizek, G. J., & Rosenberg, S. L. (2011). Psychometric methods and high stakes assessment: Contexts and methods for promoting ethics in testing. In A. T. Panter & S. Sterba (Eds.), *Handbook of ethics in quantitative methodology* (pp. 211-240). New York: Routledge.

Andrade, H., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment.* New York: Taylor and Francis.

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. Andrade & G. Cizek, (Eds.) *Handbook of formative assessment* (pp. 3-17). New York: Taylor and Francis.

Cizek, G. J. (2007). Formative classroom assessment and large-scale assessment: Implications for future research and development. In J. A. McMillan (Ed.), *Formative classroom assessment* (pp. 99-115). New York: Teachers College Press.

Cizek, G. J., & Bunch, M. (2007). Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: SAGE.

Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225-258). Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2006). Detecting and dealing with academic dishonesty. In W. Buskist & S. F. Davis (eds.), *Handbook of the teaching of psychology* (pp. 238-244). Boston: Blackwell.

Cizek, G. J. (2005). Personal and systemic influences on integrity in testing. In J. O'Reilly (ed.), *Current guidance for integrity in testing: 2005 Proceedings of the National Association of Test Directors* (pp. 1-32). Mesa, AZ: NATD.

Cizek, G. J., & Burg, S. S. (2005). *Addressing test anxiety in a high-stakes environment*. Thousand Oaks, CA: Corwin.

Cizek, G. J. (2005). High-stakes testing: Contexts, characteristics, critiques, and consequences. In R. Phelps (Ed.), *Defending standardized testing* (pp. 23-54). Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2004). Achievement tests. In C. Spielberger (Ed.), *Encyclopedia of applied psychology* (pp. 41-46). San Diego, CA: Academic.

Cizek, G. J. (2004). Cheating in academics. In C. Spielberger (Ed.), *Encyclopedia of applied psychology* (pp. 307-311). San Diego, CA: Academic.

Books, Book Chapters, and Monographs (continued)

Cizek, G. J. (2003). *Detecting and preventing classroom cheating: Promoting integrity in schools*. Thousand Oaks, CA: Corwin.

Cizek, G. J. (2003). Setting performance standards: Selected-response formats. In R. Fernandez-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 690-695) London: Sage.

Cizek, G. J. (2003). Educational testing integrity. In J. E. Wall & G. R. Walz (Eds). *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 363-386). Greensboro, NC: ERIC Counseling and Student Services Clearinghouse.

Cizek, G. J., & Page, B. A. (2003). The concept of reliability in the context of automated essay scoring. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 125-146). Mahwah, NJ: Lawrence Erlbaum.

Bergin, D. A., & Cizek, G. J. (2001). Alfred Binet. In J. A. Palmer (ed.), *100 great thinkers on education* (pp. 160-164). London, Routledge.

Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3-17). Mahwah, NJ: Lawrence Erlbaum.

Mehrens, W. A., & Cizek, G. J. (2001). Standard setting and the public good: Benefits accrued and anticipated. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 477-485). Mahwah, NJ: Lawrence Erlbaum.

Camilli, G. A., Cizek, G. J., & Lugg, C. A. (2001). Psychometric theory and the validation of performance standards: History and future perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 445-475). Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2001). An overview of issues concerning cheating on large-scale tests. In J. O'Reilly (Ed.), *2001 Proceedings of the National Association of Test Directors* (pp. 1-30). Mesa, AZ: NATD.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J., Kenney, P. A., Kolen, M. J., Peters, C. & van der Linden, W. J. (1999). *An investigation of the feasibility of linking scores on the proposed Voluntary National Tests and the National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

Cizek, G. J. (Ed.) (1999). *Handbook of educational policy*. San Diego, CA: Academic.

Cizek, G. J., & Ramaswamy, V. (1999). American educational policy: Constructing crises and crafting solutions. In G. J. Cizek (Ed.), *Handbook of educational policy*. (pp. 497-519). San Diego, CA: Academic.

Books, Book Chapters, and Monographs (continued)

Cizek, G. J. (1999). The tale wagging the dog: Narrative and neo-pragmatism in teacher education and research. In J. Raths & A. McAninch (Eds.). *Advances in teacher education, Vol. 5, What counts as knowledge in teacher education*? (pp. 47-68). Norwood, NJ: Ablex.

Cizek, G. J. (1998). *Filling in the blanks: Putting standardized tests to the test.* Washington, DC: Thomas B. Fordham Foundation.

Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (pp. 1-33). New York: Academic.

JOURNAL ARTICLES and REVIEWS

Cizek, G. J. (in press). Review of *Wide Range Achievement Test*. In J. F. Carlson, K. F. Geisinger, & J. J. Jonson (Eds.), *Twenty-first mental measurements yearbook*. Lincoln, NE: University of Nebraska Press.

Cizek, G. J., Kosh, A., & Toutkoushian, E. (2018). Gathering and evaluating validity evidence: The *Generalized Assessment Alignment Tool*. *Journal of Educational Measurement, 55*, 477-512.

Cizek, G. J. (2017). Review of *Iowa Assessments*. In J. F. Carlson, K. F. Geisinger, & J. J. Jonson (Eds.), *Twentieth mental measurements yearbook* (pp. 418-423). Lincoln, NE: University of Nebraska Press.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice, 23*(2), 212-225. DOI: 10.1080/0969594X.2015.1063479

Cizek, G. J. (2016). Progress on validity: The glass half full, the work half done. *Assessment in Education: Principles, Policy & Practice, 23*(2), 304-308, DOI: 10.1080/0969594X.2016.1156642

Agger, C., & Cizek, G. J. (2016). Measurement in education in the United States. In L. Meyer (ed.), *Oxford bibliographies in education*. New York: Oxford University Press. DOI: 10.1093/obo/9780199756810-0060

Cizek, G. J. (2014). Review of *Developing and validating test items*. *Teachers College Record.*

Agger, C., & Cizek, G. J. (2013). Measurement in education. In L. Meyer (ed.), *Oxford bibliographies in education*. New York: Oxford University Press.

Cizek, G. J., Germuth, A., & Schmid, L. (2013). *A guide to evaluating K-12 assessment programs.* Kalamazoo, MI: Evaluation Center, Western Michigan University.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use, *Psychological Methods, 17*(1)*,* 31-43.

Cizek, G. J., Schmid, L., & Germuth, A. (2012). *A guide to evaluating credentialing testing programs.* Kalamazoo, MI: Evaluation Center, Western Michigan University.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence in educational and psychological tests: A follow-up study. *Educational and Psychological Measurement, 70*, 732-743*.*

Cizek, G. J. (2010). Review of *Measures of Academic Progress*. In R. A. Spies, J. F. Carlson & K. F. Geisinger (Eds.), *Eighteenth mental measurements yearbook* (pp. 327-332). Lincoln, NE: University of Nebraska Press.

Cizek, G. J. (2009). The past of test-based educational accountability. [Review of *The future of test-based educational accountability*]. *Educational Researcher, 38,* 467-468*.*

.

Articles and Reviews (continued)

Cizek, G. J. (2009). Response to "Accountability testing: Getting situated." *Educational Researcher, 38,* 471-472.

Cizek, G. J. (2009). Reliability and validity of information about student achievement: Comparing the contexts of large-scale and classroom testing. *Theory Into Practice, 48,* 63–71.

Cizek, G. J., Rosenberg, S., & Koons, H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68,* 397-412.

Cizek, G. J. (2008). Assessing educational measurement: Ovations, omissions, opportunities. [Review of *Educational Measurement, 4th edition*]. *Educational Researcher, 37*(2), 96-100.

Cizek, G. J. (2007). Review of *Mechanical Technician B Examination*. In K. F. Geisinger, R. A. Spies, J. F. Carlson, & B. S. Plake (Eds.), *Seventeenth mental measurements yearbook*. (pp. 518-522). Lincoln, NE: University of Nebraska Press.

Burke, J. N., & Cizek, G. J. (2006). Effects of composition mode and self-perceived computer skills on essay scores of sixth graders. *Assessing Writing, 11, 148-166.*

Cizek, G. J., Crocker, L., Frisbie, D. A., Mehrens, W. A., & Stiggins, R. J. (2006). A tribute to Robert L. Ebel: Scholar, teacher, mentor, and statesman. *Educational Measurement: Issues and Practice, 25*(1), 23-32.

Cizek, G. J. (2005). Review of *TerraNova, Second Edition*. In R. A. Spies & B. S. Plake (Eds.), *Sixteenth mental measurements yearbook* (pp. 1025-1030). Lincoln, NE: Buros Institute of Mental Measurement.

Cizek, G. J. (2005). Adapting testing technology to serve accountability aims: The case of vertically-moderated standard setting. *Applied Measurement in Education, 18*(1), 1-10.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Cizek, G. J. (2003). Review of *Woodcock-Johnson III*. In J. C. Impara, B. S. Plake, and R. A. Spies (Eds.), *Fifteenth Mental Measurements Yearbook* (pp. 1019-1024). Lincoln, NE: Buros Institute of Mental Measurement.

Cizek, G. J. (2003). E. *Educational Measurement: Issues and Practice, 22*(1), 40-44.

Cizek, G. J., Hirsch, T., Trent, R., & Crandell, J. (2002). A preliminary investigation of pupil proficiency testing and state education reform initiatives. *Educational Assessment, 7*(4), 283-302.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19-27.

Cizek, G. J. (2001). Review of *Brigance Comprehensive Inventory of Basic Skills--Revised*. In B. S. Plake & J. C. Impara (Eds.), *Fourteenth mental measurements yearbook* (pp. 172-176). Lincoln, NE: Buros Institute of Mental Measurement.

Articles and Reviews (continued)

Cizek, G. J. (2001). Review of *Developmental Indicators for the Assessment of Learning--Third Edition*. In B. S. Plake & J. C. Impara (Eds.), *Fourteenth mental measurements yearbook* (pp. 394-398). Lincoln, NE: Buros Institute of Mental Measurement.

Cizek, G. J. (2001). Detecting and preventing cheating on credentialing examinations. *CLEAR Exam Review*, 12, 14-21.

Cizek, G. J. (2001). Cheating to the test. *Education Matters Journal, 1*(1), 40-47.

Cizek, G. J. (2000). Review of *Contemporary assessment for educators* [book review]. *Journal of Educational Measurement, 37,* 179-185.

Cizek, G. J. (2000). Pockets of resistance in the assessment revolution. *Educational Measurement: Issues and Practice, 19*(2), 16-23, 33.

Cizek, G. J., & Fitzgerald, S. M. (1999). An introduction to logistic regression. *Measurement and Evaluation in Counseling and Development, 31,* 223-245

Cizek, G. J., & Robinson, K. L., & O'Day, D. M. (1998). Nonfunctioning options: A closer look. *Educational and Psychological Measurement, 58*(4), 605-611.

Cizek, G. J. (1998). Review of *Office Skills Series*. In J. C. Impara & B. S. Plake (Eds.), *Thirteenth mental measurements yearbook* (pp. 706-708). Lincoln, NE: Buros Institute of Mental Measurements.

Cizek, G. J. (1998). Review of *Team Development Survey*. In J. C. Impara & B. S. Plake (Eds.), *Thirteenth mental measurements yearbook* (pp. 1002-1003). Lincoln, NE: Buros Institute of Mental Measurements.

Cizek, G. J., Rachor, R. E., & Fitzgerald, S. F. (1996). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment, 3*(2), 159-179.

Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice, 15*(2), 20-31.

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice, 15*(1), 13-21, 12.

Cizek, G. J. (1996). The hegemony of the narrative: Reflections on the contours of social science research. *Review of Higher Education, 19*(2), 227-236.

Cizek, G. J. (1995). Review of *Professional Employment Test*. In J. J. Kramer & J. C. Conoley (Eds.), *Twelfth mental measurements yearbook* (pp. 818-820). Lincoln, NE: Buros Institute of Mental Measurements.

Cizek, G. J. (1995). Review of *Expressive One-Word Picture Vocabulary Test (Revised)*. In J. J. Kramer & J. C. Conoley (Eds.), *Twelfth mental measurements yearbook* (pp. 374-376). Lincoln, NE: Buros Institute of Mental Measurements.

Cizek, G. J., & Ray, B. D. (1995). An analysis of home education research and researchers. *Home School Researcher, 11*(2), 1-9.

Articles and Reviews (continued)

Cizek, G. J., Webb, L. C. & Kalohn, J. (1995). The use of cognitive taxonomies in licensure and certification test development: Reasonable or customary? *Evaluation & the Health Professions, 18*(1), 77-91.

Cizek, G. J. (1995). Crunchy granola and the hegemony of the narrative. *Educational Researcher, 24*(2), 26-28.

Cizek, G. J. (1995). On the limited presence of African-American teachers: An assessment of research, synthesis, and policy implications. *Review of Educational Research, 65*(1), 78-92.

Cizek, G. J. & O'Day, D. (1994). Further investigation of non-functioning options in a multiple choice examination. *Educational and Psychological Measurement, 54*(4), 861-872.

Cizek, G. J. (1994). In defense of the test. *American Psychologist, 49*(6), 525-526.

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement, 54*(1), 8-20.

Cizek, G. J. (1994). Religious education in home schools: Goals/outcomes mismatch? *Religious Education, 89*(1), 43-51.

Cizek, G. J. & Rachor, R. E. (1994). The real testing bias: The role of values in educational assessment. *NASSP Bulletin, 78*(560), 83-93.

Cizek, G. J. (1993). Alternative assessment: Yes, but why? *Educational Horizons, 72*(1), 36-40.

Cizek, G. J. (1993). The mismeasure of home schooling effectiveness. *Home School Researcher, 9*(3), 1-4.

Cizek, G. J. (1993). Testing for learning: A remonstrance [book review]. *Educational Measurement: Issues and Practice, 12*(4), 40-41.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93-106.

Cizek, G. J. (1993). Some thoughts on educational testing: Measurement policy issues into the next millennium. *Educational Measurement: Issues and Practice, 12*(3), 10-16, 22.

Cizek, G. J. (1993). Rethinking psychometricians' beliefs about learning. *Educational Researcher, 22*(4), 4-9.

Cizek, G. J. (1993). The place of psychometricians' beliefs in educational reform: A rejoinder to Shepard. *Educational Researcher, 22*(4), 14-15.

Cizek, G. J. (1992). Alternative assessments: Promises and problems for home-based education policy. *Home School Researcher, 7(*4), 13-21.

Cizek, G. J. (1992). Review of *Test of Legible Handwriting*. In J. J. Kramer & J. C. Conoley (Eds.), *Eleventh mental measurements yearbook* (pp. 966-968). Lincoln, NE: Buros Institute of Mental Measurements.

Articles and Reviews (continued)

Cizek, G. J. (1992). Review of *VITAL Checklist and Curriculum Guide*. In J. J. Kramer and J. C. Conoley (Eds.), *Eleventh mental measurements yearbook* (pp. 1003-1004). Lincoln, NE: Buros Institute of Mental Measurements.

Cizek, G. J. & Bridges, J. A. (1991). Task analysis: Foundation for a valid examination. *Journal of Nuclear Medicine Technology, 19*(4), 245-246.

Cizek, G. J. (1991). Confusion effusion: A rejoinder to Wiggins. *Phi Delta Kappan, 73*(2), 150-153.

Cizek, G. J. (1991). Innovation or enervation: Performance assessment in perspective. *Phi Delta Kappan, 72*(9), 695-699. Reprinted in K. M. Cauley, F. Linder, & J. H. McMillan (Eds.) (1992), *Annual editions in educational psychology, 92-93* (pp. 232-236), Guilford, CT: Dushkin; In A. E. Woolfolk (Ed.) (1993), *Readings and cases in educational psychology* (pp. 238-243), Boston: Allyn and Bacon; In K. M. Cauley, F. Linder, & J. H. McMillan (Eds.) (1993), *Annual editions in educational psychology, 92-93* (pp. 232-236), Guilford, CT: Dushkin; In H. F. Clarizio, W. A. Mehrens, & W. G. Hapkiewicz (Eds.) (1994), *Contemporary issues in educational psychology, 6th ed.* (pp. 256-261), New York: McGraw-Hill; In W. C. Martin & G. Franklin (Eds.), *Analyzing multicultural teaching and learning styles* (pp. 236-245), Dubuque, IA: Kendall-Hunt.

Cizek, G. J. (1990). The case against the SAT: A review and comment [Review of *The Case Against the SAT*]. *Educational and Psychological Measurement, 50*(3), 701-707.

Cizek, G. J. (1990). Home education alternatives vs. accountability: A tractable problem? *Educational Policy, 4*(2), 109-125.

Cizek, G. J. (1989). Norms, standards, and nonsense: Testing policy gone bad. *Home School Researcher, 5*(4), 1-6.

Cizek, G. J. (1988). Applying standardized testing to home-based educational programs: Reasonable or customary? *Educational Measurement: Issues and Practice, 7*(3), 12-19.

OTHER PUBLICATIONS AND REPORTS

Cizek, G. J. (2019, July). *Observation and report on the North Carolina End-of Grade (3-8) , End-of-Course (NC Math 1, NC Math 3), and NCEXTEND1 (3-8) and NC Math 1 Alternate Assessment Mathematics standard setting workshops*. Raleigh, NC: North Carolina Department of Public Instruction.

Cizek, G. J. (2019, April 16). Common core-ruption: The gulf between large-scale and classroom assessment. *Education Week, 38*(29), p. 24.

Buckendahl, C. W., Cizek, G. J., & Gonthier, I. (2018). *Evaluating credentialing examination programs: Policies and practices.* Las Vegas, NV. ACS.

Cizek, G. J., Germuth, A., Kosh, A., & Schmid, L. (2016). *A guide to evaluating K-12 assessment programs (Revised edition).* Kalamazoo, MI: Evaluation Center, Western Michigan University. [Available at: https://www.wmich.edu/evaluation/checklists].

Cizek, G. J. (2016, July). *Observation and report on English language proficiency assessment for the 21st Century (ELPA21) standard setting workshops.* Monterey, CA: Pacific Metrics.

Cizek, G. J. (2016, August). *Observation and report on the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) grade 10 English/language arts, mathematics, and science standard setting.* Indianapolis, IN: Indiana Department of Education.

Cizek, G. J. (2016, September 9). Should college be free? *WalletHub.* Available at: https://wallethub.com/blog/should-college-be-free/24403/#gregory-j-cizek

Cizek, G. J. (2016, March 2.). Getting student assessment right. *The Intersection*. Durham, NC: The Hunt Institute. Available at: http://www.hunt-institute.org/resources/2016/03/getting-student-assessment-right/

Cizek, G. J. (2015, January). *Report on standard setting for the National Board Dental Examination, Part II*. Chicago, IL: Joint Commission on National Dental Examinations.

Cizek, G. J. (2015, June). *Report on standard setting for the National Board Dental Hygiene Examination.* Chicago, IL: Joint Commission on National Dental Examinations.

Kosh, A. E., & Cizek, G. J. (2015, October 14). *Checklists for evaluating K-12 and credentialing testing programs.* Available at: http://www.evalu-ate.org/category/blog/

Cizek, G. J., Schmid, L., Kosh, A., & Germuth, A. (2015). *A guide to evaluating credentialing testing programs.* Kalamazoo, MI: Evaluation Center, Western Michigan University. [Available at: https://www.wmich.edu/evaluation/checklists].

Cizek, G. J. (24 August, 2014). Testing my patience. *New York Daily News.* Available at: http://www.nydailynews.com/opinion/testing-patience-article-1.1917350

Cizek, G. J. (2014, December). *Report on standard setting for the National Board Dental Examination, Part I*. Chicago, IL: Joint Commission on National Dental Examinations.

Cizek, G. J., & Koons, H. (2014, October). *Observation and report on the Smarter Balanced standard setting*. San Diego, CA: Smarter Balanced Assessment Consortium.

Other Publications and Reports (continued)

Cizek, G. J. (2014, January 7). *Assessment strategies for evaluating instructional interventions.* Invited presentation to UNC Graduate School Instructional Leadership Program, Chapel Hill, NC.

Cizek, G. J., Germuth, A., Kosh, A., & Schmid, L. (2014). *A guide to evaluating K-12 assessment programs.* Kalamazoo, MI: Evaluation Center, Western Michigan University. [Available at: https://www.wmich.edu/evaluation/checklists].

Cizek, G. J. (2013, June 30). *Validity argument for NAEP reporting on 12th grade academic preparedness for college*. Prepared for the National Assessment Governing Board, Washington, DC.

Cizek, G. J. (2013, June). *Observation and report on the Minnesota MCA-III, MCA-Modified, and MTAS English language arts standard setting*. Roseville, MN: Minnesota Department of Education.

Cizek, G. J. (2012, August). *Observation and report on the Minnesota MCA-III 5th grade science standard setting*. Roseville, MN: Minnesota Department of Education.

Cizek, G. J. (2011, July 25). Cheating on tests and other dumb ideas. *Education Week, 30*(37). Available at: http://www.edweek.org/ew/articles/2011/07/25/37cizek.h30.html?tkn= PLVFAX0IPRIpPO3YL2UeyYaId6XxqE2Xd4h8&cmp=clp-edweek

Plake, B. S., Cizek, G. J., & Impara, J. C. (2010-2011). *Defense Language Proficiency Tests, Version 5, standard setting setting reports [Dari, Egyptian, Farsi, French, Hebrew, Iraqi, Korean, Levantine, Pashto, Russian, Serbo-Croatian, Spanish]*. Monterey, CA: Defense Language Institute.

Cizek, G. J. (2010, March). Error of measurement: Validity and the place of consequences. *NCME Newsletter, 18*(1). Available at http://www.ncme.org/pubs/pdf/vol_18_num_1.pdf

Cizek, G. J. (2010, 30 March). Eight questions for Gregory Cizek: On the role of testing in America's education system. *The Economist*. Available on-line at http://www.economist.com/blogs/democracyinamerica/2010/03/testing_and_assessment

Cizek, G. J., & Plake, B. S. (2009). *Setting performance standards for the Defense Language Proficiency Tests: Final Report.* Available from the Defense Language Institute, Monterey, CA.

Plake, B. S., Impara, J. C., Cizek, G. J., & Sireci, S. G. (2008, June). *Advanced Placement examinations standard-setting: Final report.* New York, NY: The College Board.

Plake, B. S., Sireci, S. S., Impara, J. C., & Cizek, G. J. (2008, January). *Standard setting for the College Board Advanced Placement examinations.* New York, NY: The College Board.

Cizek, G. J. (2008, June). *Review of the ASWB licensure examinations.* Culpeper, VA: Association of State Social Work Boards.

Cizek, G. J. (2007, June). *Observation and report on the Minnesota MCA-II standard setting*. Roseville, MN: Minnesota Department of Education.

Other Publications and Reports (continued)

Cizek, G. J. (2007, June). *Observation and report on the New Jersey High School Proficiency Assessment in science*. Trenton, NJ: New Jersey Department of Education.

Cizek, G. J. (2007). Fostering healthy views about testing. *The Score* [APA Division 5 Newsletter], *29*(2), 1, 12.

Linn, R. L., Cizek, G. J., Kolen, M. J., & Swaminathan, H. (2007). *Recommendations to the Delaware Department of Education on the Next Generation of Statewide Assessment System*. Available from Delaware Department of Education, Dover, DE.

Cizek, G. J. (2006, January). Evaluation of College Board Advanced Placement best practices study methodology. Available from The College Board, New York, NY.

Cizek, G. J. (2005. July). *Final report: Review and recommendations related to test security.* Report prepared for the Texas Education Agency, Austin, TX. Available from www.tea.state.tx.us/student.assessment/admin/texasreport.pdf

Cizek, G. J., Engelhard, G. Jr., & Moody, M. (2005, January). *Final report: Review of Delaware Student Testing Program assessments and performance standards.* Available from Office of the Governor, Dover, DE.

Cizek, G. J. (2004, May). *Review and recommendations related to standard setting procedures applied to certain examinations of the American Nurse Credentialing Center.* Washington, DC: Knapp & Associates International, Inc.

Federation of State Boards of Physical Therapy. (2003). *National physical therapy examination (NPTE) commission report*. Alexandria, VA: Author.

Brosnan, W., Chapman, S., Cizek, G. J., et al. (2003). *Final report to the New York State Board of Regents and the New York State Commissioner of Education, Independent Panel on Math A.* Albany, NY: New York State Education Department.

Cizek, G. J. (2003, September 21). High-stakes testing must pass the integrity test. M*emphis Commercial Appeal*, p. A-7.

Cizek, G. J. (2002/2003). When educators cheat. *NAESP Streamlined Seminar, 21*(2), 1-3. [Reprinted as Cizek, G. J. (2003). When teachers cheat. *Education Digest*, 68(6), 28-31.]

Cizek, G. J. (2002). *A report to the Dayton area chamber of commerce on pupil achievement in Dayton area charter schools, 2001-2002 academic year*. Chapel Hill, NC: Author.

Cizek, G. J. (2000, December 6). School politics 101: It's not really about the children. *Education Week*, pp. 35-36.

Cizek, G. J. (2000, January/February). Academic Notebook: Standards and Testing, *Michigan Learning, 11*(2), 7-8.

Cizek, G. J. (2000, January/February). A user-friendly guided to standards and testing terms, *Michigan Learning, 11*(2), 9-10.

Other Publications and Reports (continued)

Cizek, G. J. (1999, December 8). How cartoons and calculators resolved the class-size debate. *Education Week*, pp. 26, 30.

Cizek, G. J. (1999). Give us this day our daily dread: Manufacturing crises in American education. *Phi Delta Kappan, 80*(10), 737-743.

Cizek, G. J. (1998). The assessment revolution's unfinished business. *Kappa Delta Pi Record, 34*(4), 138-143.

Cizek, G. J. (1996). Grades: The final frontier in assessment reform. *NASSP Bulletin, 80*(584), 103-110.

Donmoyer, R., Eisner, E., Gardner, H., Stotsky, S., Wasley, P., Tillman, L., Cizek, G., & Gough, N. (1996). Viewpoints: Should novels count as dissertations in education? *Research in the Teaching of English, 30*(4), 403-427.

Cizek, G. J. (1996, April 17). There's no such thing as grade inflation. *Education Week, 15*(30), 32, 22.

Cizek, G. J. (1996). Voices in education. *Midwestern Educational Researcher, 9*(2), 43.

Cizek, G. J. (1996). Voices in education. *Midwestern Educational Researcher, 9*(1), 34-35.

Cizek, G. J. (1995). The big picture in assessment and who ought to have it. *Phi Delta Kappan, 77*(3), 246-249.

Cizek, G. J. (1995, February). *An evaluation of "An Integrated Curricular Approach to Teaching about the Great Lakes Region, 1993-1994*." Available from Lucas County Office of Education, Toledo, OH.

Cizek, G. J. (1994, December). An evaluation of nature education/ environmental workshops, 1993-1994. Available from Lourdes College, Sylvania, OH.

Cizek, G. J. (1994). Voices in education. *Mid-Western Educational Researcher*, 7(4), 35.

Cizek, G. J. (1994, September 21). SAT scores recentered: Baby boomers get a break. *Education Week, 14*(3), pp. 40, 34.

Cizek, G. J. (1994, December 7). S.A.T. recentering redux. [Response] *Education Week, 14*(14), 38.

Cizek, G. J. (1993, November 10). On the disappearance of standards. *Education Week, 13*(10), pp. 32, 24.

Cizek, G. J. (1993, March). *Evaluation of General Accounting Office report on national assessment of educational progress achievement levels*. Available from National Assessment Governing Board, Washington, DC.

Cizek, G. J. (1993, May). *Evaluation of Ohio ninth-grade proficiency test technical characteristics*. Columbus, OH: Ohio Legislative Office of Education Oversight.

Other Publications and Reports (continued)

Cizek, G. J. (1993, September). *Evaluation and comment on National Academy of Education report on National Assessment of Educational Progress achievement levels setting*. Available from National Assessment Governing Board, Washington, DC.

Cizek, G. J. and others (1993, September). *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing: A technical report on reliability and validity*. Iowa City, IA: American College Testing.

Cizek, G. J., Jurs, S. G. & Maynard, J. (1993, April). *Report on content analysis of responses to "Teacher Education and Certification Discussion Guide*." Columbus, OH: Ohio Department of Education, Standards Revision Committee for Teacher Education.

Cizek, G. J. (1992). *Evaluation of the Davenport (IA) Community School District Performance Assessment Program*. Toledo, OH: Author.

Cizek, G. J. & Butman, A. M. (1992). Essentials for teaching EMTs: An instructor's guide to better teaching. In J. D. Heckman (Ed.), *Emergency care and transportation of the sick and injured*, 5th ed., (pp. 247-271). Park Ridge, IL: American Academy of Orthopaedic Surgeons.

Cizek, G. J. (Speaker). (1992). *Conversations about authentic assessment* (Instructional Cassette; Project RP91002002). Charleston, WV: Appalachian Educational Laboratory.

Cizek, G. J. (1992, April 8). From a 'Card-carrying Psychometrician'. *Education Week, 11*(29), p. 27.

Cizek, G. J. (1992). Standardized tests should not be eliminated from schools. In C. P. Cozic (Ed.), *Education in America: Opposing viewpoints* (pp. 51-54). San Diego, CA: Greenhaven.

Cizek, G. J. & Hartnett, S. (1991). Nuclear Medicine Technology ready for task analysis revalidation. *NMTCB News, 4*(1), pp. 2, 4.

Cizek, G. J. (1991). Reasoning about testing. *Educational Digest, 56*(5), 56-58. Reprinted in (1996) *Educated in the USA*, J. Nelson, C. Hass, & S. Greene (Eds.), Dubuque, IA: Kendall/Hunt.

Cizek, G. J., Webb, L. C., & White, A. S. (1990). *Criterion-referenced standard setting: A User's Guide*. Iowa City, IA: ACT Publications.

Cizek, G. J. (1990, April 4). The 'sloppy' logic of test abolitionists. *Education Week*, p. 64. Reprinted in (1996) *Issues in literacy*, J. Nelson (Ed.), Dubuque, IA: Kendall Hunt.

Cizek, G. J. (1990). Sloppy reasoning about testing. *California School Boards Journal, 49*(2), 9-11.

Cizek, G. J. (1990). Using standardized tests to evaluate educational quality. *The Teaching Home, 8*(1), 35-36.

Beechick, R., Cizek, G. J., & Bumcrot, C. (1990). Glossary of testing terms. *The Teaching Home, 8*(1), 30.

Other Publications and Reports (continued)

Beechick, R., Karman, D., & Cizek, G. (1990). Test-taking tips. *The Teaching Home, 8*(1), 31.

Cizek, G. J. (1989). Planning and presenting a lesson. *The Teaching Home, 7*(5), 25-28.

Cizek, G. J. (1989). *GSCORE/EQANAL User's Manual*. Iowa City, IA: American College Testing Program.

Numerous other expert testimony reports, technical reports, evaluation reports, and research reports.

CONFERENCE PAPERS AND PROFESSIONAL PRESENTATIONS

Cizek, G. J. (2020, February). *Prevention, detection, and responding to cheating on tests.* Invited presentation to the North Carolina State University College of Veterinary Medicine, Academy of Educators, Raleigh, NC.

Cizek, G. J. (2019, June b). *Effective components of working with a technical advisory committee.* Presentation to the National Conference on Student Assessment, Orlando, FL.

Cizek, G. J. (2019, June a). Formative assessment: General and discipline-specific aspects and practices. Presentation to the National Conference on Student Assessment, Orlando, FL.

Cizek, G. J. (2019, April c). *Autoethnography of the depressed: Narratives from a failed academic career.* Presentation to the American Educational Research Association annual meeting, Toronto, Ontario, Canada.

Cizek, G. J. (2019, April b). *Detecting and managing test irregularities.* Presentation at the annual meeting of the National Council on Measurement in Education, Toronto, Ontario, Canada.

Cizek, G. J. (2019, April a). *Formative assessment in the disciplines: Advances in theory and practice.* Presentation at the annual meeting of the National Council on Measurement in Education, Toronto, Ontario, Canada.

Cizek, G. J. (2019, March b). *Communicating technical testing concepts to non-technical audiences: Usefulness and examples of analogies.* Presentation to the annual meeting of the Association of Test Publishers, Orlando, FL.

Cizek, G. J. (2019, March a). *The future of assessment: Technical challenges, opportunities, and enabling conditions for advancing assessment in 2040.* Invited closing plenary session presentation, annual meeting of the Association of Test Publishers, Orlando, FL.

Cizek, G. J., & Kosh, A. (2019, February). *Innovation in alignment: The generalized assessment alignment tool.* Invited presentation to the Council of Chief State School Officers, Technical Issues in Large-Scale Assessment, Los Angeles, CA

Cizek, G. J. (2018, May). *Automated scoring of constructed responses*. Invited presentation to the Ohio State Board of Education, Columbus, OH.

Cizek, G. J. (2018, April). *Innovations in standard setting.* Presentation at the annual meeting of the National Council on Measurement in Education, New York, NY.

Cizek, G. J. (2018, February). *Test score validity, validation, and use: A comprehensive model of defensible testing practice*. Invited presentation to the University of North Carolina Eshelman School of Pharmacy Education Research Community (PERC). Chapel Hill, NC

Cizek, G. J. (2017i, June). *Transporting best assessment practices across credentialing and K-12 contexts.* Presentation at the annual meeting of the National Conference on Student Assessment, Austin, TX.

Cizek, G. J. (2017h, June). *Interacting with media representatives on assessment issues.* Presentation at the annual meeting of the National Conference on Student Assessment, Austin, TX.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (2017g, June). *NextGen test security for on-line and paper-based testing.* Invited presentation at the annual meeting of the Technical Issues in Large Scale Assessment (TILSA/SCASS), Austin, TX.

Cizek, G. J., Kosh, A. E., & Toutkoushian, E. K. (2017, April). *Essential content validity evidence and innovation in alignment methodology.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Kosh, A. E., Cizek, G. J., & Toutkoushian, E. K. (2017, April). *Gathering and evaluating validity evidence: The generalized assessment alignment tool.* Presentation at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Cizek, G. J. (2017f, April). *How can philosophy of measurement illuminate what we mean when we "measure" in education?* Invited presentation at the annual meeting of the American Educational Research Association, San Antonio, TX.

Cizek, G. J. (2017e, April). *Advances in standard setting: The ELPA 21 process.* Presentation at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Cizek, G. J. (2017d, April). *Engineered cut scores: Promises and practicalities.* Presentation at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Cizek, G. J. (2017c, April). Balancing psychometrics and policy. Presentation at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Cizek, G. J. & Kosh, A. (2017b, March). Recommendations for licensure and certification agencies on assessment best practices from the K-12 context. Presentation at the annual meeting of the Association of Test Publishers, Scottsdale, AZ.

Cizek, G. J. (2017a, February). *Validation of score meaning and justification of a score use: A comprehensive model of defensible testing practice*. Invited presentation to quarterly meeting of the Triangle Assessment Networking Group, Raleigh, NC.

Cizek, G. J. (2016, November). *NextGen assessment and accountability: Innovations in policy and practice*. Keynote address, annual meeting of the Michigan Educational Research Association, Frankenmuth, MI.

Cizek, G. J. (2106, October). *Validity, quantitative methods, and test security: Advances in a maturing field*. Opening plenary presentation, 5th Annual Conference on Test Security, Cedar Rapids, IA.

Kosh, A., & Cizek, G. J. (2016, September). *Assessment alignment: Background, need, and a methodological innovation in gathering and evaluating validity evidence.* Invited presentation, Quantitative Psychology Colloquium, University of North Carolina, Chapel Hill, NC.

Cizek, G. J. (2016, June a). *Engineered cut scores: Context, technical, and policy considerations*. Presentation at the National Conference on Student Assessment, Philadelphia, PA.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (2016, June b). *Opting-out of accountability: Background and implications.* Presentation at the National Conference on Student Assessment, Philadelphia, PA.

Cizek, G. J. (2016, June c). *Incorporating contextual information in standard setting: Five preliminaries.* Presentation at the National Conference on Student Assessment, Philadelphia, PA.

Cizek, G. J. (2016a, April). *Opting-out: Psychometric and policy implications.* Presentation at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Cizek, G. J. (2016b, April). *Epic fails in the history of the Uhmairikun Edjoocayshunul Reesirch Ahsosheeayshun*. Presentation at the annual meeting of the American Educational Research Association, Washington, DC.

Cizek, G. J. (2015, December). *Quality assessments and effective state policy.* Invited presenter, Hunt Institute Holshouser Legislators Retreat, Greensboro, NC.

Cizek, G. J. (2015, September). *Measuring achievement: Trends and issues in ensuring high-quality student assessment.* Invited presentation to Indiana House and Senate Joint Study Group on Student Assessment. Indianapolis, IN.

Cizek, G. J. (2015, September 16). *The concept of validity in a comprehensive model of defensible testing practice*. UNC School of Education Research Brown Bag presentation, Chapel Hill, NC.

Cizek, G. J. (2015, June 24). *Next generation assessment systems for achievement and accountability: Where do we go from here?* Presentation at the National Conference on Student Assessment, San Diego, CA.

Cizek, G. J. (2015, June 22). *Innovations in achievement level setting for Smarter Balanced assessments.* Presentation at the National Conference on Student Assessment, San Diego, CA.

Cizek, G. J. (2015, May 2). Fundamentals of setting performance standards. Invited presentation to the National Council on Architectural Registration Boards, Washington, DC.

Cizek, G. J. (2015, March 13). *Emerging research and practice in standard setting.* Invited research colloquium, International Credentialing Associates, Philadelphia, PA.

Cizek, G. J. (2015, March 4). *Four critical issues in educational assessment for North Carolina.* Invited presentation to NC Senate and House Education Committee members, Raleigh, NC.

Cizek, G. J. (2015, January 12). *Making sense of the evolving Common Core assessment landscape.* Invited presentation to the Education Writers Association seminar on Covering Standards & Testing in the Common Core Era. Chapel Hill, NC.

Cizek, G. J. (2014g, April). *Modern validity theory: Consensus, concerns, and a course for the future.* Presentation to the University of North Carolina Department of Psychology Quantitative Methods program. Chapel Hill, NC.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (2014f, April). *Validation of score inferences is different from justification of test use.* Presentation at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Cizek, G. J. (2014e, April). *The UnCommon core: A problematized procrustean primer.* Presentation at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Cizek, G. J. (2014d, March). *Ten years of test security: The road traveled and the road ahead.* Presentation at the Association of Test Publishers annual conference, Scottsdale, AZ.

Cizek, G. J. (2014c, February). *A problem with the canon: An overview of validity fundamentals and a proposed remedy.* Invited address to the Research Triangle Program Evaluators Group, Durham, NC.

Cizek, G. J. (2014b, February). *Assessment consorti-yum! A menu of measurement mysteries.* Invited keynote address to the 25th Annual Conference of the South Carolina Educators for the Practical Use of Research, Columbia, SC.

Cizek, G. J. (2014a, January). *Assessment and evaluation in education.* Invited workshop to faculty of Shenzhen Polytechnic University, Chapel Hill, NC.

Cizek, G. J. (2013b, October). *Three eras of cheating detection: An illustrated history and critique of policy and practice.* Invited keynote address to the Second Annual Conference on Statistical Detection of Test Fraud, Madison, WI.

Cizek, G. J. (2013a, August). *Contemporary trends in large-scale, K-12 assessment: A survey and critique of current and emerging practices.* Invited keynote address to the Minnesota Assessment Conference, St. Paul, MN.

Cizek, G. J. (2013b, August). *Modern validity theory: Consensus, concerns, and a course for the future.* Presentation to the Minnesota Assessment Conference, St. Paul, MN.

Cizek, G. J. (2013, June.) *Best practices in test security*: *Next steps*. Presentation to the National Conference on Student Assessment, National Harbor, MD.

Cizek, G. J. (2013, May). *Reconceptualizing a fundamental: Validation of score meaning and justification of test use.* Invited research colloquium, CTB/McGraw-Hill, Monterey, CA.

Cizek, G. J. (2013, April). *An unpublishable address.* Presidential address to the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Cizek, G. J. (2013, April). *Score meaning as a validity concern.* Invited presentation to the Association of Social Work Boards spring meeting, Austin, TX.

Cizek, G. J. (2013, February). Current contexts and challenges in ensuring test score integrity. Invited presentation to the CCSSO State Collaborative on Assessment and Student Standards/Technical Issue in Large Scale Assessment Conference, Atlanta, GA

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (2012f, November). *How and what he teaches: A different perspective.* Invited address at the University of Massachusetts Conference honoring Ronald Hambleton. Amherst, MA.

Cizek, G. J. (2012e, September). *Setting performance standards: Foundations, features, and the future*. Invited address to the American Institutes for Research and Washington (DC) Statistical Society. American Institutes for Research, Washington, DC.

Cizek, G. J. (2012d, April). *Validation and justification: Recommendations for conceptual clarity in test score meaning and use.* Paper presented at the annual meeting of the American Educational Research Association, Vancouver, WA.

Cizek, G. J. (2012c, April). *Ensuring integrity in large-scale assessment: Shared responsibilities.* Paper presented at the annual meeting of the American Educational Research Association, Vancouver, WA.

Cizek, G. J. (2012b, February). *Prevention strategies to ensure test score integrity: shared responsibilities*. Invited presentation, United States Department of Education, National Center for Education Statistics, Washington, DC.

Cizek, G. J. (2012a, February). *Ensuring integrity in testing: Context, responsibilities, and recommendations*. Keynote Address, Michigan School Testing Conference, Ann Arbor, MI.

Cizek, G. J. (2011c, October). *Preventing, detecting, and responding to cheating on exams and assignments: Background, context, and recommendations.* Invited address, UNC-CH School of Nursing, Chapel Hill, NC

Cizek, G. J. (2011b, September). *Ensuring test score integrity: Background, contexts, and recommendations*. Invited presentation, United States Department of Education, Washington, DC.

Cizek, G. J. (2011a, July). *Setting performance standards: Foundations, trends, and issues. Invited* presentation, Pearson professional development seminar, Iowa City, IA.

Cizek, G. J. (2011b, April). *Rethinking the concept of validity, the place of consequences, and the practice of validation.* Colloquium presentation, University of Michigan School of Education, Ann Arbor, MI.

Cizek, G. J. (2011c, April). *Ins and outs of educational assessment: From value-added measures to common core.* Invited presentation American Educational Research Association and Education Writers Association Seminar for Journalists presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Cizek, G. J. (2011d, April). *Validity: What it is, and isn't.* Invited presentation, Validity Research SIG and Professional Licensure and Certification SIG at the annual meeting of the American Educational Research Association, New Orleans, LA.

Cizek, G. J. (2011e, April). *Secrets for early career professionals: The real AERA pre-session.* Invited presentation at the annual meeting of the American Educational Research Association, New Orleans, LA.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (2011f, April). *Test security concerns for K-12 testing programs: What has changed, and what hasn't.* Presentation at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Cizek, G. J. (2011g, April). *Error of measurement: Reconsidering validity theory and the place of consequences.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Cizek, G. J. (2010, November). *Toward a reconceptualization of validity theory: Validation of score inferences and justification of test use.* Invited presentation for inaugural distinguished lecture series, College of Education, University of Texas—Austin.

Cizek, G. J. (2010, October). *Translating standards into assessments: The opportunities and challenges of a common core.* Invited paper presentation to a symposium on Common Core Assessments. Washington, DC: Brookings Institution.

Cizek, G. J. (2010, October). *Reconceptualizing validity and the place of consequences.* Invited keynote address to the ASI Annual Research Conference, Ottawa, Ontario, Canada.

Cizek, G. J., Church, K., & Bowen, D (2010, September). *Sources of validity evidence for educational and psychological tests: A follow-up study.* Poster presented at the UNC School of Education 125th anniversary research symposium, Chapel Hill, NC.

Cizek, G. J., Bowen, D., & Church, K. (2010, May). *Sources of validity evidence for educational and psychological tests: A follow-up study.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Cizek, G. J. (2009e, November). *Error of measurement: Validity and the place of consequences.* Invited lecture, University of Minnesota, College of Education, Minneapolis, MN.

Cizek, G. J. (2009d, October). *Error of measurement: Reconsidering validity theory and the place of consequences.* Invited address, National Institute of Testing and Evaluation, Jerusalem, Israel.

Cizek, G. J. (2009c, October). *Fundamentals of psychometrics.* Invited presentation to the Federation of State Boards of Physical Therapy, San Diego, CA.

Cizek, G. J. (2009b, October). *Setting performance standards on licensure examinations*. Invited presentation to the Federation of State Boards of Physical Therapy, San Diego, CA.

Cizek, G. J. (2009a, April). *Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Cizek, G. J. (2008, October b). *An introduction to setting performance standards. Invited presentation*, University of Iowa, Iowa City, IA.

Cizek, G. J. (2008, September). *Test security: A primer and current issues*. Keynote address for the annual meeting of the National College Testing Association, Baltimore, MD.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (2008, June). *State assessment programs and the problem of cheating: How to fail.* Presentation at the National Conference on Student Assessment, Orlando, FL.

Cizek, G. J. (2008, March c). *AERA, unaffiliated organizations, and remora: An introduction to NCME.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Cizek, G. J. (2008, March b). *Standard setting challenges in the context of augmented achievement testing.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Cizek, G. J. (2008, March a). *The future of educational measurement.* Invited presentation at the annual meeting of the National Council on Measurement in Education, New York, NY.

Cizek, G. J. (2008, February). *Reconsidering the place of consequences in validity theory and practice.* Invited presentation to MetaMetrics, Durham, NC.

Cizek, G. J. (2007f, October). *Fundamentals of test item writing and analysis.* Invited presentation to the University of North Carolina School of Medicine, Chapel Hill, NC.

Cizek, G. J. (2007e, August). *Introduction to modern validity theory and practice.* Invited presentation to the National Assessment Governing Board, McLean, VA.

Cizek, G. J. (2007d, May). *Chronicling and questioning validity: Mental Measurements Yearbook as a context for investigating sources of evidence for high-stakes tests.* Paper presented at the Invitational Conference of the Buros Institute of Mental Measurements, Lincoln, NE.

Cizek, G. J., Rosenberg, S., & Koons, H. (2007c, April). *Sources of validity evidence for educational and psychological tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Cizek, G. J. (2007b, March). *Fundamentals of psychometrics.* Invited presentation to the Federation of State Boards of Physical Therapy, Milwaukee, WI.

Cizek, G. J. (2007a, March). *Setting performance standards on licensure examinations*. Invited presentation to the Federation of State Boards of Physical Therapy, Milwaukee, WI.

Cizek, G. J. (2006e, July). *Possibly intended consequences of high-stakes testing.* Keynote address, annual meeting of the New York Schools Data Analysis Group, Saratoga Springs, NY.

Cizek, G. J. (2006f, July). *Setting standards, including vertically-moderated standard setting.* Invited presentation at the annual meeting of the New York Schools Data Analysis Group, Saratoga Springs, NY.

Cizek, G. J. (2006e, April). *Test consequences, test validity, and testing policy.* Symposium presentation at the annual meeting of the American Educational Research Association, San Francisco, CA.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (2008, October a). *Error of measurement: Reconsidering validity theory and the place of consequences.* Invited address, 12th annual William E. Coffman Invited Lecture, University of Iowa, Iowa City, IA.

Cizek, G. J. (2006d, April). *Tricks of the trade: Decisions made, risks taken, and opportunities seized by accomplished researchers throughout their careers.* Inhibited presentation at the annual meeting of the American Educational Research Association, San Francisco, CA.

Cizek, G. J. (2006c, April). *A systemic approach to test security.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Cizek, G. J. (2006b, February). *Enhancing test security: A comprehensive approach* Invited workshop, Association of Test Publishers annual meeting, Orlando, FL.

Cizek, G. J. (2006a, January). *Enhancing and monitoring Test Security: Individual and systemic aspects.* Invited presentation to the National Council on Education Statistics/Council of Chief State School Officers Joint Conference, LaJolla, CA.

Cizek, G. J. (2005a, August). *Cheating on tests: A systemic perspective.* presentation at the 2nd annual Lexile National Reading Conference, Durham, NC.

Cizek, G. J. (2005b, August). *Testing myths.* Keynote address, 2nd Annual Lexile National Reading Conference, Durham, NC.

Cizek, G. J. (2005c, September). *Psychometrics of the National Physical Therapy Examination.* Invited presentation, annual meeting of the Federation of State Boards of Physical Therapy, Austin, TX.

Cizek, G. J. (2005d, April). *Setting and reviewing the passing standard.* Keynote address, annual meeting Federation of State Boards of Physical Therapy, Austin, TX.

Cizek, G. J. (2005e, April). *Testing myths.* Invited presentation, annual meeting of the American Educational Research Association, Montreal, PQ, Canada.

Cizek, G. J. (2005f, April). *Portrait of the artist as a young psychometrician.* Presentation to the annual meeting of the American Educational Research Association, Montreal, PQ, Canada.

Cizek, G. J. (2005g, April). *Personal and systemic influences on integrity in testing.* Paper presented to the annual meeting of the National Council on Measurement in Education, Montreal, PQ, Canada.

Cizek, G. J. (2005h, April). *A look the other way: From measurement practice to theory.* Presentation to the annual meeting of the National Council on Measurement in Education, Montreal, PQ, Canada.

Cizek, G. J. (2005i, April). *Formative classroom assessment and large-scale testing: The state of the union.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, PQ, Canada.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (2005j, April). *High-stakes testing: Contexts, characteristics, critiques, and consequences.* Paper presented to the annual meeting of the National Council on Measurement in Education, Montreal, PQ, Canada.

Cizek, G. J. (2005k, February). *Individual vs. systemic aspects of cheating: Test score corruption in context*. Invited address, 19th annual Texas Assessment Conference, Austin, TX.

Cizek, G. J. (2004a, November). *High-stakes testing: Myths and consequences*. Keynote address, Florida Educational Research Association annual meeting, Tampa, FL.

Cizek, G. J. (2004b, November). *Setting performance standards: Concepts and methods*. Invited presentation, Florida Educational Research Association annual meeting, Tampa, FL.

Cizek, G. J. (2004c, July). *More unintended consequences of high-stakes testing*. Invited presentation to the Council of Chief State School Officers annual curriculum and assessment conference, Baltimore, MD.

Cizek, G. J. (2004d, April). *Protecting the integrity of computer-adaptive tests: Results of a legal challenge.* Presentation to the annual meeting of the American Educational Research Association, San Diego, CA.

Cizek, G. J. (2004e, April). *Robert Ebel: Educational statesman.* Presentation to the annual conference of the National Council on Measurement in Education, San Diego, CA.

Cizek, G. J. (2004f, February). *Test cheating: Problems and solutions.* Presentation to the annual conference of the Association of Test Publishers, Palm Springs, CA.

Cizek, G. J. (2003, November). *PhDs say the darndest things (about testing).* Invited address to the annual meeting of the Virginia Association of Test Directors, Richmond, VA.

Cizek, G. J. (2003a, October). *Three critical issues in assessment.* Invited Keynote Address to the annual meeting of the Arizona Educational Research Organization, Phoenix, AZ.

Cizek, G. J. (2003b, October). *Contemporary methods and issues in setting performance standards.* Invited workshop presentation for the annual meeting of the Arizona Educational Research Organization, Phoenix, AZ.

Cizek, G. J. (2003a, February). *More unintended consequences of high-stakes testing.* Invited address to the Joint Meeting of the Texas Association of Collegiate Testing Personnel and the Texas Annual Assessment Conference, Austin, TX.

Cizek, G. J. (2003b, February). *Consequences of Testing: There's the rub.* Invited presentation to the Joint Meeting of the Texas Association of Collegiate Testing Personnel and the Texas Annual Assessment Conference, Austin, TX.

Cizek, G. J. (2002, November). *High-stakes testing consequences: There's the rub.* Invited plenary presentation to the Virginia Association of Test Directors, Richmond, VA.

Cizek, G. J. (2002a, May). *Accountability for what? Rounding out the accountability picture.* Presentation at the Ohio Charter Schools Annual Conference, Columbus, OH.

Conference Papers and Professional Presentations (continued)

Cizek, G. J., & Pinkerton, T. (2002b, May). *The Dayton assessment project.* Presentation at the Ohio Charter Schools Annual Conference, Columbus, OH.

Cizek, G. J., & Pammer, M. (2002c, May). *Choosing valid instruments for effective assessment.* Presentation at the Ohio Charter Schools Annual Conference, Columbus, OH.

Cizek, G. J. (2002a, April). *Standard setting using the item mapping approach.* Invited presentation to the Texas State Board of Education, Austin, TX.

Cizek, G. J. (2002b, April). (Re)forming the triennial travesties. Presentation at the annual meeting of the American Educational Research Association, New Orleans, LA.

Cizek, G. J. (2002, February). *Unintended consequences of high-stakes testing.* Invited presentation to the Buffalo area Phi Delta Kappa chapter, Williamsville, NY.

Cizek, G. J. (2001a, April). *Testing accommodations: Raising a white flag or waving a checkered one?* Symposium presentation at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Cizek, G. J. (2001b, April). *Disseminating stories in education: Power and the physicist's plea.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Cizek, G. J. (2001c, April). *An overview of issues concerning cheating on large-scale tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Cizek, G. J. (2001d, April). *Unintended consequences of high-stakes pupil testing programs.* Invited presentation, Virginia Commonwealth University College of Education, Richmond, VA.

Cizek, G. J. (2001a, March). *High-stakes testing and accountability systems: Unintended consequences, unrecognized benefits.* Invited presentation to the New York State Education Department Conference on Validity in Testing, Albany, NY.

Cizek, G. J. (2001b, March). *Accountability and assessment for charter school operators*. Invited presentation to the annual meeting of the New York Charter Schools Association, New York, NY.

Cizek, G. J. (2001c, March). *Elements of an effective assessment component for charter schools*. Invited presentation to the New York Charter Schools Resource Center Conference, New York, NY.

Cizek, G. J. (2001, January). *Unanticipated consequences of high-stakes testing.* Paper presented at the annual North Carolina Department of Public Instruction Accountability Conference, Greensboro, NC.

Cizek, G. J. (2000, November a). *When the alarms should go off when test results are reported.* Presentation to the Education Writers Association, Cleveland, OH.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (2000, November b). *The problem of cheating on tests in education*. Invited keynote address to the Dutch Testing Society, Zuthphen, Netherlands.

Cizek, G. J. (2000, September). *Cheating on credentialing examinations: Who, why, how, detecting, and preventing*. Presentation at the annual meeting of the National Council on Licensure, Enforcement, and Regulation (CLEAR), Key Biscayne, FL.

Cizek, G. J., Trent, E. R., Crandell, J., Hirsch, T., & Keene, J. (2000, April). *Research to inform policy: An investigation of pupil proficiency testing requirements and state education reform initiatives.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ED 443 873)

Cizek, G. J. (2000, April). *Factors affecting linkage of the Voluntary National Tests and the National Assessment of Educational Progress.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.(ED 447 196)

Cizek, G. J. (2000, April). *You do your work and you do my work: Bearing one another's burdens in classroom assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Cizek, G. J. (2000, February). *Cheating on tests and its threat to school accountability programs.* Invited presentation to the North Carolina Department of Public Instruction 2000 Accountability Conference, Greensboro, NC.

Cizek, G. J. (1999, April). *The role and uses of assessment in charter schools*. Invited presentation to the New York State Charter Schools Conference, New York, NY.

Cizek, G. J. (1999, April). *(Re)Forming the AERA Annual Meeting.* Address presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Cizek, G. J. & Husband, T. H. (1997, March). *A Monte Carlo investigation of the contrasting groups standard setting method.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Cizek, G. J. (1996a, June). *Developing and evaluating tests for nurse educators.* Invited presentation, Medical College of Ohio, Toledo, OH.

Cizek, G. J., & Fitzgerald, S. M. (1996b, April). *A comparison of group and independent standard setting*. Paper presented at the annual meeting of the American Educational Research Association, New York. [ERIC Document Reproduction Service No. TM025679]

Rachor, R. E., & Cizek, G. J. (1996c, April). Reliability of raw gain, residual gain, and estimated true gain scores: A simulation study. Paper presented at the annual meeting of the American Educational Research Association, New York.

Cizek, G. J. (1996d, April). Statistical detection of answer copying: Getting a focus on the big picture. Presented at the annual meeting of the American Educational Research Association, New York.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (1996e, April). Comment on proposed revisions to the AERA/APA/NCME Standards for Educational and Psychological Testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Cizek, G. J. (1996f, April). *Yes, but is it research? Should a novel count as a dissertation in education?* Symposium presentation at the annual meeting of the American Educational Research Association, New York.

Cizek, G. J. (1996g, April). *Can we talk? An attempted conversation across research paradigms, purposes and perspectives.* Symposium presentation at the annual meeting of the American Educational Research Association, New York.

Stiggins, R., & Cizek, G. J. (1996h, February). *Assessment: The key to high quality student learning.* Invited presentation to the John P. Rusel Center for Educational Leadership, University of Toledo, Toledo, OH.

Cizek, G. J. (1995, May). *Future directions for the National Assessment of Educational Progress (NAEP).* Invited paper prepared for the National Assessment Governing Board, Washington, DC.

Cizek, G. J. (1995, April). *Standard setting as psychometric due process: Going further down an uncertain road.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA (ED 384 614).

Cizek, G. J., Rachor, R. E., & Fitzgerald, S. M. (1995, April). *Further investigation of teachers' grading practices*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA (ED 384 613).

Cizek, G. J. (1995, April). *On the relevance of intelligence: Theory and practice in education.* Remarks presented at symposium presentation, annual meeting of the American Educational Research Association, San Francisco, CA.

Cizek, G. J. & Rachor, R. E. (1995, April). *Nonfunctioning options in multiple-choice tests: A closer look.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Cizek, G. J. (1995, May). *Home school assessment: Obligations, alternatives, and interpretations.* Invited address to the annual meeting of Christian Home Educators of Ohio, Columbus, OH.

Cizek, G. J. (1995, January). *Preparing good tests and using them well.* Invited presentation, Davis College Faculty In-Service, Toledo, OH.

Cizek, G. J. & Rachor, R. E. (1994, October a). *Non-functioning options in multiple-choice tests: Another look.* Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago, IL.

Rachor, R. E. & Cizek, G. J. (1994, October b). *An empirical investigation of the reliability of gain scores and modified gain scores.* Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago, IL.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. & Rachor, R. E. (1994, October c). *Teachers' grading practices: Who's doing what, and why?* Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago, IL.

Cizek, G. J. (1994, May). *What is standard setting? Current conceptualizations and future issues*. Invited presentation, Educational Testing Service, Princeton, NJ.

Cizek, G. J. (1994, April). *Whatever happened to the measurement of intelligence?* Symposium presentation at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Sun, A., & Cizek, G. J. (1994, April). *Development of an instrument for measuring high school student resistance to schooling*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Cizek, G. J. (1994, February). *Issues in establishing standards of performance for a credentialing program*. Invited presentation to the PES Annual Invitational Conference on Licensure and Certification, Seattle, WA.

Cizek, G. J. (1993, November). *A critical look at the 1992 NAEP achievement levels setting process*. Invited presentation to the National Assessment Governing Board, San Francisco, CA.

Cizek, G. J. (1993, October). *Setting levels: Those little devils*. Invited presentation to the Council of Chief State School Officers, Education Information Advisory Committee Meeting, Alexandria, VA.

Cizek, G. J. (1993, April). *Home education research: On the right road?* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA (ED 360 331).

Cizek, G. J., Webb, L. C. & Kalohn, J. (1993, April). *The use of cognitive taxonomies in licensure and certification test development: Reasonable or customary?* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Cizek, G. J. (1993, April). *Cheating on educational assessments: An introduction to frame investigations.* Introductory remarks, session chair, presented at the annual meetings of the American Educational Research Association & National Council on Measurement in Education, Atlanta, GA.

Cizek, G. J. (1993, January). *Constructing and evaluating tests for nurse educators*. Faculty presentation, Mercy College of Nursing, Toledo, OH.

Cizek, G. J. (1992, September). *Evaluating the quality of test items: The good, the bad, and the ugly.* Invited seminar for the University of Toledo Department of Health Promotion and Human Performance, Toledo, OH.

Conference Papers and Professional Presentations (continued)

Cizek, G. J. (1992, August). *Utilizing testing data in regular and special education.* Invited seminar presentation for the Ottawa County (Ohio) schools Annual Administrators' Symposium, Avon Lake, OH.

Cizek, G. J. (1992, March). *Issues in educational testing*. Invited address to the University of Toledo Annual Conference on Testing, Toledo, OH.

Cizek, G. J. (1992, January). *Performance assessment: Uses, abuses, excuses.* Keynote address presented to the annual meeting of the Michigan Educational Research Association, Novi, MI.

Cizek, G. J. (1992, January). *Performance assessment: Questions and answers.* Symposium presentation at the annual meeting of the Michigan Educational Research Association, Novi, MI.

Cizek, G. J. (1991, April). *The effect of altering the position of options in a multiple-choice examination.* Paper presented to the National Council on Measurement in Education, Chicago, IL. (ERIC Document Reproduction Service No. ED 333 024).

BOOK/PUBLICATION PROPOSAL REVIEWS

Cizek, G. J. (2014). Assessment in education [book series]. Routledge.

Cizek, G. J. (2011). *Handbook of test development*. Routledge.

Cizek, G. J. (2011). *Automated essay evaluation.* Taylor & Francis.

Cizek, G. J. (2009). *Equating groups: Modern matching and other methods.* Routledge.

Cizek, G. J. (2004). *Validity and Accommodations.* Lawrence Erlbaum Associates.

Cizek, G. J. (2002). *Classroom assessment: Enhancing the quality of teacher decision making*. Lawrence Erlbaum Associates.

Cizek, G. J. (2001). *Statistical methods in education and psychology (3rd ed.)*. Allyn and Bacon.

Cizek, G. J. (2000). *The charter school landscape: Politics, policies, and prospects*. Teachers College Press.

Cizek, G. J. (1999). *Issues, research, and recommendations for large-scale Assessment programs*. Lawrence Erlbaum Associates.

Cizek, G. J. (1998). *Testing in American schools: Getting the right answers*. Lawrence Erlbaum Associates.

Cizek, G. J. (1997). *Educational testing and measurement (5th ed.).* Harper-Collins.

Cizek, G. J. (1995). *Computer-based tutorials on statistical concepts.* Longman.

Cizek, G. J. (1993). *Authentic testing in the classroom.* Harper-Collins.

Cizek, G. J. (1992). *Practical statistics for educators.* Longman.