NAEP Validity Studies Panel Responses to the Reanalysis of TUDA Mathematics Scores

Gerunda Hughes
Howard University

Peter Behuniak
Criterion Consulting LLC

Scott Norton
Council of Chief State School Officers

Sami Kitmitto
American Institutes for Research

Jack Buckley
American Institutes for Research

October 2019
Commissioned by the NAEP Validity Studies Panel

The NAEP Validity Studies Panel was formed in 1995 to provide a technical review of NAEP plans and products and identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members

Peter Behuniak
Criterion Consulting LLC

Jack Buckley
American Institutes for Research

James R. Chromy
Research Triangle Institute (retired)

Phil Daro
Strategic Education Research Partnership Institute

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Ina V. S. Mullis
Boston College

Scott Norton
Council of Chief State School Officers

James Pellegrino
University of Illinois at Chicago

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado, Boulder

David Thissen
University of North Carolina, Chapel Hill

Gerald Tindal
University of Oregon

Sheila Valencia
University of Washington

Denny Way
College Board

Project Director

Frances B. Stancavage
American Institutes for Research

Project Officer

Grady Wilburn
National Center for Education Statistics

For Information

NAEP Validity Studies Panel
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
E-mail: fstancavage@air.org

ACKNOWLEDGMENTS

The authors would like to thank Fran Stancavage at the American Institutes for Research, whose assistance was instrumental in the successful completion of this work. In addition, the authors would like to thank the other members of the NAEP Validity Studies Panel for their helpful comments and discussion on numerous occasions.

CONTENTS

INTRODUCTION **1**

BACKGROUND **4**

Alignment Between Standards and Assessments 5
NAEP in the Context of Common Core and College and Career Ready Standards 6
The Value of Alignment Studies to Investigate Threats to the Validity of the NAEP Results 7
Criteria for Alignment of Expectations and Assessments 9
Implications for the Dogan 2019 Analysis 9

TWO METHODOLOGICAL CONSIDERATIONS **11**

Overweighting Versus Underweighting 11
State Assessments Are Assumed to Be Proxies for the Opportunity to Learn 12

CONSIDERATIONS FOR REPORTING REANALYSIS FOR STATES AND DISTRICTS **14**

One NAEP, or More? 14
Communications Efforts 14
Other Practical Considerations 15

CONCLUSION **16**

REFERENCES **18**

APPENDIX: ANALYSIS OF RECENT NAEP TUDA MATHEMATICS RESULTS BASED ON ALIGNMENT TO STATE ASSESSMENT CONTENT **21**

INTRODUCTION

During the past decade, the NAEP Validity Studies (NVS) Panel has been monitoring, studying, and commenting on potential issues with the validity of the National Assessment of Educational Progress (NAEP) arising from changes that have been brought about by the adoption of rigorous state college and career readiness standards, such as the Common Core State Standards (CCSS). NAEP is meant to be reflective of the entirety of what is taught in the United States, and the many changes to standards in the past 10 years have led to questions about the extent to which NAEP continues to meet this objective. The NVS Panel has conducted several studies to investigate this issue (Behuniak, 2015; Daro, Hughes, & Stancavage, 2015; Daro et al., in press; Hughes, Daro, Holtzman, & Middleton, 2013; Valencia, Wixson, Kitmitto, & Doorey, in press) and has found some variations in the alignment between state and NAEP standards across different NAEP grades and subjects.

Given the prominent attention that the NAEP results receive, states and participating districts are especially interested in learning how well those results relate to what is taught and tested in schools. States and districts have informally posited that, if the alignment between NAEP frameworks and their own content standards were closer, then their NAEP scores might be higher. Stated another way, there are concerns that NAEP may be underreporting the actual abilities of their students—and trends in achievement—because of some degree of misalignment.

When the 2017 NAEP Mathematics TUDA (Trial Urban District Assessment) results were reported and it appeared that student performance trends on NAEP were not similar to student performance trends on the state assessments that were aligned to college and career ready standards, several leaders in the affected TUDAs called for what amounted to a "recount." The results of student performance on the state assessments from 2013 to 2017 were showing more positive trends than the results of student performance on NAEP during the same period. In the reanalysis discussed here (Dogan, 2019; reproduced in the appendix), the study author noted that most of the negative NAEP trends observed in districts in recent years (i.e., 2015 to 2019) coincide with major changes in states' learning standards and assessments, raising the legitimate question: Can these trends be a function of the differences between NAEP assessment content and states' transition to new college and career readiness learning standards, such as the CCSS, and the corresponding shift in the content of state assessments to be aligned to these newer standards?

This issue of mismatch trends or the misalignment of results should be examined for at least two reasons: (a) Urban districts are held accountable for students' performance on state assessments that are aligned to state-mandated content standards; however, (b) urban districts are not held directly accountable for performance on NAEP, which historically was regarded by many as the standard against which the adequacy of state assessments is judged.

The results of the recent NVS Panel study by Daro et al. (in press) document the extent of alignment between NAEP and state mathematics assessments according to several important dimensions, one of which is content distribution. In the 2019 study, Daro et al. estimated the content distributions in NAEP and four assessments used in states for their respective 2017 mathematics assessments in Grades 4 and 8 by operationalizing content emphasis as the

percentage of total test score points for each content domain (e.g., Algebra, Geometry, Measurement, Data); then they compared the content percentage for each domain on each state assessment in the study with the corresponding content domain percentage in the 2017 NAEP. These results provide an opportunity for further analysis, as performed by Dogan (2019). Dogan's TUDA reanalysis study was designed to explore whether content misalignment might be a possible reason for the mismatched results for the TUDAs on NAEP and the respective state assessments. The following research questions were posited:

1. How would the 2017 mathematics Grades 4 and 8 TUDA mean scores change if the NAEP subscales were weighted according to the content distribution of selected state assessments?
2. How would the mathematics Grades 4 and 8 TUDA mean scores change in 2013, 2015, and 2019 if the NAEP subscales were weighted according to the content emphasis of selected state assessments, assuming the content emphasis of those assessments and NAEP were similar in these years compared with 2017?

This report serves as a response from the NVS Panel to the analysis conducted by Dogan (2019). Selected panel members were asked to provide comment, and their responses have been edited together into this report. With Dogan's analysis in the appendix of this report, the main body has organized the comments from the selected panel members into three areas. First, an extensive background section provides context for the motivation behind conducting such an analysis. This section covers important historical background on the alignment of standards and assessments, the implications of the college and career ready standards for NAEP, and the value of alignment studies to investigate the validity of NAEP. The second section provides comments on and caveats for the methods used in Dogan's analysis. The final section considers the implications of Dogan's results for NAEP and the reporting of results.

The conclusion of the report is that the secondary analysis done by Dogan for the NAEP TUDA scores is important and worthy of further exploration as part of ongoing efforts to monitor the validity of NAEP.[1] However, such analyses should not be used in the reporting of any official statistics or even as a recurring set of ancillary results or appendix material. To the extent that there is a real and educationally significant mismatch between the content covered on NAEP and that in the states, the best way to ameliorate this is by modifying the NAEP frameworks, not through post hoc reweighting of the NAEP results. In the case of mathematics, the National Assessment Governing Board (NAGB) has nearly completed an update of the framework for implementation in 2025 that will hopefully address the issue of alignment with newer state content frameworks comprehensively.

Summary of Dogan's Results

For the first research question, Dogan investigated if and how the 2017 NAEP mathematics Grades 4 and 8 TUDA means would change if the subscales were weighted according to the content distribution of the selected state assessments. The results show that for Grade 4, for example, the 10% weight assigned to the content domain Data in the NAEP framework is

[1] Issues discussed in this paper apply to NAEP state assessment results as well. However, the focus of this paper is solely on the TUDA assessments because the Dogan (2019) study analyzed TUDA results exclusively, in response to concerns raised by TUDA stakeholders.

reweighted to 0% or 1% when the weights for the state assessments are applied. Similarly, the results show that for Grade 8, the 30% weight assigned to Algebra in the NAEP framework is reweighted to 45% when the weight for one of the state assessments is applied. Reweighted composite mean scores were computed for nine TUDAs that take either SA2, SA3, or SA4 (names withheld for confidentiality)[2] as their state assessment. All nine TUDAs showed positive changes in their mean scores.

For the second research question, for Grades 4 and 8, the results showed positive changes in the TUDA means when the subscale weights were adjusted in a way that they mirror the content emphasis of the state assessment associated with each TUDA when the same 2017 weights were applied to the 2015 and 2019 assessments. The pattern of results was not the same when these weights were applied to the 2013 assessments. Dogan concluded that this difference in the pattern of results might be explained either (a) by changes in content emphasis within state assessments from 2013 to 2017 or (b) because any differences in content emphasis between states and NAEP mattered less in earlier years as a result of the newness of the standards and assessments.

[2] SA = state assessment.

BACKGROUND

For more than 50 years, NAEP, often called the Nation's Report Card, has been in the unique position of providing periodic measures of student achievement in a variety of subjects based on nationally representative probability samples. NAEP has two separate components: long-term trend and main NAEP. Long-term trend and main NAEP both assess mathematics and reading; however, there are several differences, particularly in the content assessed. Content in the long-term trend assessment has remained essentially consistent across time; whereas, content in the main NAEP is expected to be updated periodically to reflect changes in educational objectives and curricula in the nation's schools. Furthermore, as its name suggests, NAEP provides reports at the national level on the educational progress and status of student groups defined by gender, race/ethnicity, disability status, and levels of English proficiency. NAEP also provides progress and status reports for states and selected urban districts that participate in the TUDA.

A primary goal of the TUDA program is to support the improvement of student achievement in the nation's large urban districts and focus attention on the specific challenges of groups that often are underserved in America's educational systems because of their race, ethnicity, language background, culture, or socioeconomic status. For participating TUDAs, NAEP is administered to a sample size large enough to support the official reporting of scores for the districts in the same manner as scores are reported for states and the nation. In 2002, six districts participated in the TUDA program; by the 2017 NAEP administration, the number of participating districts had grown to 27.[3] To become eligible to participate in the TUDA, an urban district must meet the following criteria: (a) have a population of 250,000 or more; (b) have a student enrollment large enough to support NAEP in three subjects in each grade assessed (i.e., a minimum of 1,500 students per subject per grade level assessed); and (c) meet at least one of the following criteria:

- At least 50% of the students are from minority backgrounds (i.e., African American, American Indian/Alaskan Native, Asian, Hispanic, Native Hawaiian/Other Pacific Islander, and/or multiracial.
- At least 50% of the students are eligible for participation in the free or reduced-price lunch program (or other appropriate indicator of poverty status; NAGB, 2012).

Urban districts often face numerous challenges in their efforts to educate all of their students—many of whom fall into the categories just described. Stakeholders in these districts, such as teachers, parents, policymakers, and administrators, often view their local educational systems as being more test centered rather than student centered because they perceive that a large portion of time and financial resources that could be used to improve instruction and achievement are used instead for the development and administration of tests and assessments. Furthermore, concerns often are raised about the reliability or validity of using test results or items from large-scale standardized tests for instructional purposes. Yet, the same test results may be used, for example, to make high-stakes decisions about retention or graduation. In addition, many educators are concerned that the performance of students on external tests and assessments, such as NAEP, does not accurately reflect what

[3] For a history of district participation in the TUDA program, see https://nces.ed.gov/nationsreportcard/tuda/.

the students in their districts know and can do because the external tests are not adequately aligned with what students are expected to know and do or with what is being taught.

Alignment Between Standards and Assessments

Across the decades, it was necessary for NAEP to have a measure of sensitivity to what students were learning in school, even as a plethora of educational reform movements driven by different educational policies defined and changed educational priorities. Because NAEP is the Nation's Report Card, it could not operate in a vacuum with respect to the designs and demands of the American standards-based educational system and still provide accurate and meaningful assessment information about the educational progress and status of student achievement.

When Congress mandated the establishment of NAEP in 1969, mathematics education reform had just exited the decade of the 1960s and the "new math" movement and entered the decade of the 1970s, which was focused on "back to basic" skills. This era was followed by the "standards" movement in the 1980s that produced reports such as *A Nation at Risk* (National Commission on Excellence in Education, 1983) and the *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 1989).

The decade of the 1990s built on what was accomplished in the 1980s regarding content standards and witnessed a proliferation of standards related to pedagogy, assessment, and professional development, just to name a few. The newly developed content standards, for example, provided guidance for what should be taught. In addition, affiliated education professionals helped teachers and other school-based personnel think deeply about ways in which pedagogy and assessment could be consistent with or aligned to the NCTM standards. The NAEP framework used to build the NAEP mathematics assessments from 1990 to 2003 reflected elements of the NCTM (1989) content standards, including emphasis on "mathematical power" defined by reasoning, connections, and communication. These components of mathematics learning, along with three types of "mathematical abilities" (e.g., problem solving, conceptual understanding, and procedural knowledge), were the forerunners to current mathematical practices (CCSS; National Governors Association Center for Best Practices, & Council of Chief State School Officers, 2010; Reese, Miller, Mazzeo, & Dorsey, 1997). By 1996, 40 states had adopted the NCTM standards, and many were aligning their state assessment programs with those standards (Council of Chief State School Officers, 1996; Webb, 1997).

Then, early in the first decade of the 21st century, the passage of the No Child Left Behind Act (2001) signaled strong support for the standards movement and attached to it a layer of testing and accountability that brought with it rewards as well as sanctions for teachers, students, administrators, schools, districts, and states. The intent of the educational policy statement, Public Law 107-110, was to close the achievement gap between children of different racial and ethnic groups in the United States and between American children and children from other countries on international assessments. To accomplish this policy goal, the federal government would (a) hold states, districts, and schools accountable for developing educational delivery systems that will ensure that all students, including those who are disadvantaged, meet high academic standards; (b) require states to assess students annually in specific grades in reading and mathematics and share that information with parents and other

stakeholders; and (c) implement a system of rewards and sanctions for schools based on the performance of students and the progress that they make yearly. Under this policy, NAEP was expected to serve, in part, as a common yardstick to monitor these new, heterogeneous state assessments. The move to have NAEP play this role raised concerns among some stakeholders about the potential encroachment of the federal government on states' rights and responsibilities for setting their own educational policy and practices. Furthermore, stakeholders were fearful that using NAEP in this role could lead to unfair state comparisons that would advantage states whose assessments were more aligned with NAEP.

Though not explicitly stated, the educational policy statements of the No Child Left Behind Act (2001) and its successor, the Every Student Succeeds Act (2015), assume that the assessment instruments or procedures used at the state or national levels, including NAEP, to carry out the aforementioned educational policy goals are fair and valid representations of what was intended to be taught (as defined by high standards and curriculum frameworks); what was actually taught (as defined by, though not limited to, students' opportunity to learn [OTL]); and what was actually learned (which may or may not be represented by students' test scores). Specifically, if the content domains and emphases sampled by NAEP are different from the content domains and emphases on state assessments, which are, in turn, purportedly aligned with state-mandated content standards, then NAEP must be prepared to address issues related to levels of alignment required for successful student performance on NAEP.

Therefore, given the aforementioned policy and stakeholder concerns, it became necessary for the NVS Panel to explore and recommend a validity research agenda on issues related to potential threats to the validity of NAEP results that are used for reporting, including but not limited to content representation and emphasis, uses, sampling, trends, and the analysis of data (U.S. Department of Education, 2003).

NAEP in the Context of Common Core and College and Career Ready Standards

In 1989, as the nation was about to enter the last decade of the 20th century, it was a time for the renewal of standards (e.g., NCTM standards)—that is, the bold expectations about what children in American schools should know and be able to do with that knowledge. These expectations were the result of a lot of "think-tank" collaborations in the 1980s that produced reports such as *A Nation at Risk* (National Commission on Excellence in Education, 1983) and the founding of the Mathematical Sciences Education Board in 1985. Similarly, in 2009, as the nation was about to enter the second decade of the 21st century, another set of bold expectations about what children in American schools should know and be able to do was launched by the National Governors Association and the Council of Chief State School Officers in the form of the CCSS (National Governors Association Center for Best Practices, & Council of Chief State School Officers, 2010). These bold, new expectations were developed, in part, because of concerns about international competitiveness and the need for a workforce with technological and analytical thinking skills, which had their basis in the educational policy goals of the No Child Left Behind Act (2001).

Second, just as the NCTM content standards were accompanied by mathematical process standards (e.g., problem solving, reasoning and proof, communications, connections, and representation) in the publication of NCTM's *Principles and Standards for School Mathematics*

(2000), the CCSS-M content standards were accompanied by mathematical practices (e.g., problem solving and perseverance, reasoning abstractly and quantitatively, constructing arguments and critiquing the reasoning of others, modeling with mathematics) in their release in 2009. It also is important to report that in this context, students were expected to engage in deeper learning by integrating content with "practices" in ways that are authentic in real-life situations.

Third, within 7 years of the release of the NCTM content standards in 1989, the overwhelming majority of states had adopted them; and within about 3 years of the release of the CCSS-M in 2009, most states, the District of Columbia, and the territories had adopted CCSS-M or adopted their own college and career ready standards. Lastly, in both time frames, emphasis is placed on aligning standards, assessments, and instruction.

Two notable differences in the two decades are as follows: (a) There was no explicit, high-stakes expectation that every student was required to demonstrate proficiency on mathematics assessments purportedly aligned with district/state mathematics content standards, but all students are explicitly expected to develop the mathematical competencies and knowledge bases as set forth in the college and career ready standards of the 21st century; and (b) NAEP aligned its frameworks with the NCTM content standards in 1990, a year after the standards were released; however, NAEP has not officially aligned its frameworks with CCSS-M or any set of college and career ready standards since the CCSS-M were released in 2009. In fact, according to the *2017 Mathematics Framework* document, "mathematics content objectives for grades 4 and 8 have not changed. Therefore, main NAEP trend lines from the early 1990s can continue for fourth and eighth grades for the 2017 assessment" (NAGB, 2017, p. 1). Although the decision to maintain trend lines is based on an NAEP mathematics framework from the early 1990s, the good news is that NAEP has supported the conduct of a series of alignment studies to provide qualitative descriptions and quantitative estimates of the extent to which the NAEP frameworks and item pools are correspondingly aligned to state-adopted college and career ready standards, such as CCSS-M and their associated state assessments (Daro et al., 2015; Daro et al., in press; Hughes et al., 2013).

The Value of Alignment Studies to Investigate Threats to the Validity of the NAEP Results

In spring 2011, the NVS Panel began a series of studies to examine the validity and utility of NAEP in the context of the CCSS-M. The purpose of these studies was twofold: (a) to compare the content of the NAEP mathematics frameworks and items in the NAEP item pool for Grades 4 and 8 with the content standards of the CCSS-M and the items on a sample of state assessments designed for accountability purposes and purportedly aligned to the CCSS-M or states' respective college and career ready standards; and (b) to make recommendations to the National Center for Education Statistics (NCES) regarding issues related to the content comparison of NAEP and state assessments, including the extent of alignment that is appropriate to support NAEP's continuing role as an independent monitor of student achievement in the United States.

The examination of the validity and utility of NAEP mathematics in the context of college and career ready mathematics standards and state assessments was organized into a series of

three successive studies. The first study in the series examined the alignment of content objectives in the 2011 NAEP Mathematics Framework and the CCSS-M for Grades 4 and 8 (Hughes et al., 2013); the second study examined the alignment of the 2015 NAEP mathematics item pools for Grades 4 and 8 to the CCSS-M (Daro et al., 2015); and the third study compared the 2017 NAEP mathematics item pools for Grades 4 and 8 with items on a sample of 2017 state assessments built specifically to align with the respective state's college and career ready standards for content balance, construct centrality, and complexity (Daro et al., in press).

Each study used a different alignment approach—one that was appropriate for the research questions that were posed for that study. In the first study, where the NAEP Mathematics Framework was compared with the CCSS-M, the approach was to examine the match and mismatch between NAEP and CCSS-M content; that is, to provide a description of what content is in the NAEP Mathematics Framework but not in CCSS-M and, conversely, what content is in CCSS-M but not in the NAEP Mathematics Framework. In the second study, the 2015 NAEP item pools were compared with the content domains that are targeted by the CCSS-M for instruction at or below grades tested by NAEP for Grades 4 and 8, respectively. The approach was to express agreement as a percentage of NAEP items that were clearly matched to content standards that appear in CCSS-M at or below Grade 4 or Grade 8, and, conversely, agreement was expressed as the percentage of CCSS-M standards at Grade 4 or Grade 8 assessed by at least one NAEP item in the respective grade item pool. In the third study, in which the 2017 NAEP item pools were compared with items on a sample of 2017 state assessments, the approach was to develop a consolidated content framework to classify each item (NAEP and state assessment items) into one of several content domains and subdomains. In addition, rubrics were used to rate each item along levels of content centrality and the four dimensions of complexity. After all items were classified and rated, NAEP and state assessment profiles were developed and compared. Because NAEP and state assessments differed by the number of items as well as the number of score points assigned to individual items by their own assessment program, comparisons of assessment profiles were based, in part, on the percentage of total score points that a particular content domain or subdomain contributes to the total score points for that assessment.

The different approaches used to examine the nature and extent of alignment between the various combinations of standards to assessments assist educators and policymakers in determining whether and how much these two components of an educational system are working together toward the same goal of providing valid and reliable information about student achievement. Although comparisons between standards and assessments yield useful information, more is needed. Martone and Sireci (2009) observed the following: "Beyond just the alignment of standards and assessments, the instructional content delivered to students also needs to be in agreement" (p. 1333). Put another way, what students really know and can do is most likely a reflection of what content they have been taught and what they have learned in the classroom. Hence, an alignment study that takes into account the agreement between the curriculum (the content standards), the instruction (how and what content standards are taught), and the assessment (how the content standards are operationalized into test items) will provide evidence about the validity and reliability of the inferences made about student achievement and also address some issues related to equity and fairness.

Criteria for Alignment of Expectations and Assessments

In the comprehensive, seminal monograph in which he addressed the criteria for aligning expectations and assessments in mathematics and science education, Webb (1997) noted that "[*a*]*lignment* is the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (p. 4). Expectations are defined as the major elements of educational policy that express what students should know and what they should be able to do with that knowledge. Assessments are major elements of educational policy for measuring student achievement, given the stated expectations. These two major elements must work together or be closely aligned to create a coherent education system that serves to benefit students and help them maximize their learning and achievement. To the extent that expectations and assessments are not aligned, Webb noted that the content validity of test results is threatened, and concerns about the consequential validity of the test results have become more acute.

Webb (1997) presented 12 criteria for judging the alignment of expectations and assessments. The criteria were organized into five general categories: content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability. Six subcategories are subsumed under content focus: categorical concurrence, depth of knowledge consistency, range of knowledge correspondence, structure of knowledge comparability, balance of representation, and dispositional consonance. In this context, the analysis by Dogan (2019; see the appendix) can be seen as an examination of the degree of alignment of expectations (Common Core State Standards for Mathematics [CCSS-M] or college and career ready standards) and assessments (NAEP and state assessments) by conducting a reanalysis of NAEP Mathematics TUDA results with respect to (a) categorical concurrence, which means that comparable content domains and subdomains appear in both, and (b) balance of representation, which means that the degree of importance or emphasis of different content topics is the same. Closely associated with content focus are issues related to pedagogical implications (i.e., OTL), equity, and fairness.

Implications for the Dogan 2019 Analysis

Historically, NAEP has aspired to represent the union of all the various state curricula while also reaching beyond these curricula to lead as well as reflect what they measure. What is very clear is that the NAEP Mathematics Framework does not attempt to answer the question: "What (or how) mathematics should be taught?" The introduction of the college and career ready standards provides both new opportunities and challenges for NAEP. Furthermore, as the nation moves toward widespread implementation of instruction and assessment based on the CCSS-M or other college and career ready standards, NAEP must balance the goals of comparability across time (i.e., maintaining trend) with current relevance in a dynamic educational policy environment where daily concerns about state-based standards and state accountability assessments really carry more weight. Dogan's analysis should help NAEP balance these two by providing information on the impact of potential misalignment between NAEP and state frameworks and can be considered a form of validity study.

Webb (1999) and Martone and Sireci (2009) helped us appreciate the limitations inherent in making inferences and drawing conclusions from different types of alignment study

methodologies. They also stressed the importance of developing a systemic process and analytic tools for judging the alignment among all components of a standards-based education system. To stress that point, Webb (1997) provided a comprehensive list of 12 criteria that are important to consider in conducting a thorough, comprehensive alignment study for the purpose of creating and nurturing a coherent standards-based educational system. Webb (1997) noted the complexities inherent in conducting comprehensive alignment studies; the researcher must identify which criteria are relevant to a particular case and how each criterion is operationalized. In light of this, Dogan's analysis can be considered the "tip of the tip of the iceberg" in examining the differences in content emphasis in NAEP and a sample of state assessments by approaching it from one angle: the effects of the application of different weights on student performance means.

Educators at all levels and in all aspects of the American standards-based education system must recognize that each version of the reauthorization of the Elementary and Secondary Act of 1965, which was passed as part of President Lyndon Johnson's War on Poverty, carries with it the pledge to continuously design, implement, and improve equitable educational systems in which all students have the opportunity to reach their maximum potential, are afforded the opportunity to demonstrate what they know and can do, and are entitled to receive valid and reliable information about their assessment results. In Dogan's analysis, the instructional components and OTL variables are inferred indirectly by students' performance on state accountability assessments—an issue we discuss further in the next section.

Dogan's (2019) reanalysis of the 2017 NAEP mathematics Grades 4 and 8 TUDA results compels us to examine the meaning and interpretations of student performance on NAEP and the respective state accountability assessments because of the possibility of misalignment between the two types of assessments. The meanings and interpretations of students' test results have implications for content and the consequential validity of the assessment results, as well as concerns about equity and fairness for all students. We now turn to a more detailed examination of some methodological concerns with the study.

TWO METHODOLOGICAL CONSIDERATIONS

The approach used in Dogan (2019) to reanalyze the TUDA results involves reweighting the NAEP subscales to mirror the content reflected in the assessments used by the TUDA. The purpose of these analyses is to examine whether the NAEP results would differ if one were to modify scores to better reflect the content that students have been experiencing in their classrooms.

This approach involves reweighting each NAEP subscale to form a composite mathematics score that is based on a distribution of content more closely resembling the curriculum to which students in each district participating in the TUDA are theoretically being exposed. This is essentially a postadministration, statistical modification of the NAEP test blueprint. As such in this analysis, the NAEP blueprint is essentially being customized to model local (i.e., state) curricula in each district participating in the TUDA. The content emphases on the assessments in use in each district are used as a proxy for the local curricula. These aspects of the methodology raise a number of issues that will be addressed in this section.

Customizing NAEP to match the content reflected in three different assessments resulted in a variety of subscale reweightings. Some subscales required increased weights (e.g., Numbers in Grade 4, Algebra in Grade 8), whereas others required decreased weights (e.g., Data in Grades 4 and 8). The results of these analyses were consistently positive with respect to mathematics mean scores, and the changes in score means (comparing original means to the customized means), although quite small in some cases, could be viewed as being of substantively important magnitude given that even small changes in NAEP mean scores are often cited as evidence in policy debates.

Overweighting Versus Underweighting

The first issue of concern involves the implications of performing a statistical modification of the NAEP test blueprint. When NAEP is adjusted to model the content on other assessments, the result is that some subscales require underweighting, but others require overweighting. These two types of reweighting will be examined separately.

Underweighting is required when NAEP contains a greater proportion of items in a domain than does the state assessment being modeled. For example, the Grade 4 domain of Data had more extensive coverage in NAEP (10%) than it did on the examined state assessments (0%–1%). Underweighting does not pose a validity threat because it essentially creates a situation in which more items than necessary are employed to measure learning on a given domain. The underweighting simply allows for the excess coverage to be statistically removed.

The same is not true for overweighting, which occurs whenever NAEP does not emphasize a given domain as much as the assessment being modeled. An example of this occurs on the Grade 4 domain of Numbers, where NAEP dedicated a lower proportion of items (40%) than did the state assessments (54%–73%). This is potentially problematic because it represents an increase in the importance of a given domain with no corresponding increase in content coverage.

Why is this a concern? The development of a test blueprint usually follows a sequence in which the content to be measured is determined first. The relative importance of each domain is then established and reflected in the assessment by virtue of item (or score point) counts. Thus, if Domain A is judged to be twice as important as Domain B, the proportion of the assessment dedicated to measuring Domain A will generally be twice as much as the proportion dedicated to Domain B. Content experts then work within this blueprint to determine the specific aspects of each domain that should be measured given the number of items (or score points) allocated to it.

Statistically overweighting a domain is not likely to produce the same result as creating a test blueprint with a greater emphasis in that domain. The reason for this relates to how content experts make judgments about the items that are necessary to cover a given domain. For example, if the decision is made to double the emphasis on a given domain, content experts would not be likely to recommend using two items to represent each element of the domain that was previously measured by only one item. They would, instead, be more likely to use the additional items allowed by the increased emphasis to broaden the coverage of the domain, perhaps by including nuances that could not otherwise be measured because of constraints on the number of items or time available.

This implication of overweighting does not necessarily undermine the efficacy of using the reweighting methodology, but it should be considered a limitation that poses a potential threat to the validity of the regenerated scores. The seriousness of the threat is related to the degree of overweighting and the content being measured. Increasing the weighting slightly is not as great a threat as making larger increases. Judging the threat posed by a large increase also depends on the specific content of interest and its level of complexity. For example, content experts might feel an increase in items dedicated to measuring the addition of two whole numbers may not be problematic, but a similar increase in items measuring algebra would be troublesome because of the greater complexity of the domain.

State Assessments Are Assumed to Be Proxies for the Opportunity to Learn

The accuracy of achievement test score inferences and conclusions (i.e., validity) depends largely on the sensitivity of scores to instructional experiences that are focused on policy statements—for example, standards, curriculum frameworks—that express the expectations of an educational system (D'agostino, Welsh, & Corson, 2007; Stancavage et al., 2009). Burstein (1989) noted the following:

> With respect to instructional experiences, minimally, the ability to distinguish among the different educational settings . . . in which assessments are administered is necessary for an appropriate interpretation of student performance data. Information about actual topic coverage and instructional methods are of even greater value. (p. 4)

Simply stated, students' OTL measures are essential for interpreting students' test results because they provide valuable information about what content is taught, how the content is taught, and how students learn. Furthermore, OTL measures help explain why students' test scores may vary across classrooms, across schools, within and across districts (urban vs. suburban vs. rural districts), and across states.

A plethora of empirical research and policy reports has been published to inform the education field that OTL variables are significant factors in explaining students' test results (Airasian & Madaus, 1983; Brody & Good, 1986; Darling-Hammond, 1993; Leinhart, 1983; Oakes & Guiton, 1995; Schmidt, 1992; Wang, 1998; Winfield, 1991, 1993). Researchers also warn about the possibility that achievement gaps between majority and minority students could increase if OTL variables are disregarded in research (Arreaga-May & Greenwood, 1986; Madaus, West, Harmon, Lomax, & Viator, 1992). Therefore, it is very important to assess equity in students' opportunity to learn the content of the tested material so that inferences about their performance and uses of their test scores are appropriate and fair (O'Day & Smith, 1993). To illustrate the point, in 1979, the National Association for the Advancement of Colored People filed a lawsuit against the state of Florida (*Debra P. v. Turlington, 1979*), in which they argued that it was unconstitutional to deny high school diplomas to students who had not been given the opportunity to learn content that appeared on a test that was a requirement for graduation. The trial court placed a 4-year injunction on administration of the test. The injunction allowed additional time for teachers to become familiar with the test, and for students, most of whom were African American, to have an opportunity to learn the test material.

Wang (1998) noted that the OTL construct consists of two general dimensions: the amount and the quality of exposure to new knowledge. These general dimensions can be further explained by four subdimensions: content coverage, content exposure, content emphasis, and the quality of instructional delivery. Methods for measuring the subdimensions of OTL include direct observation, surveys, questionnaires, interviews, teachers' self-reports of teaching practices, analyses of classroom assessments, and ratings of teaching materials.

Collecting data about OTL variables using any subset of these different methods could provide a substantial amount of evidence about the two variables of interest in the reanalysis of the 2017 NAEP Mathematics TUDA results: content coverage (i.e., representation) and content emphasis (i.e., balance). Yet, no direct OTL data were available for the current TUDA study. Instead, the assessments employed by the respective TUDAs were used as proxies for the local curricula. This practical decision is justifiable; however, it is worth noting that the content measured by local assessments is not identical to the content covered in the functional local curricula. There is certainly going to be variation in the emphases given to specific aspects of the content, even if all major elements of the test blueprint are addressed during instruction in the classroom.

Although an important limitation to the conclusions of Dogan (2019), ultimately this variation is not judged to be a serious threat to the validity of the reweighting methodology. The assessments in use locally can be considered reasonable, if somewhat less than perfect, measures of the functional curricula. Modeling NAEP after these assessments via reweighting is judged likely to produce scores that approach the validity of the original scores generated by the local assessments. Although this concern should be considered minor, it is worth noting that this aspect of the study may introduce some additional noise to the estimation process.

CONSIDERATIONS FOR REPORTING REANALYSIS FOR STATES AND DISTRICTS

The results of the Dogan (2019) reanalysis study are likely of interest to the TUDA districts generally and particularly to those jurisdictions specifically included in the study. Of immediate interest is whether or how the findings in the study should be shared with the public, and whether analysis of this sort should be replicated in subsequent administrations. Should these results become a part of the "official" NAEP released data, or should they be considered a secondary analysis, like those conducted from the available data sets by independent researchers? It also is important to note that the relevance of the findings in this study are not limited to TUDAs. Indeed, state policymakers are likely to be just as interested in the impact of state/NAEP misalignment on NAEP's utility as a monitor on state progress. There are, however, challenges that should be addressed if NCES and the Governing Board (i.e., NAGB) consider releasing such results.

One NAEP, or More?

From the time NAEP has become widely known, there has been a single release of the data after each administration and only one set of results to inform student achievement. As an example, when the 2017 state NAEP scores were released, Kentucky fourth graders were reported to have a mean score of 240, with 40% of the students scoring at or above the Proficient level. There is no other official fourth-grade mathematics score reported for Kentucky. There is a precedent for the release of special results, however: the full population estimates. Since 2005, the NAEP program has released ancillary full population estimates for states and TUDAs based on statistical imputation methods to produce alternative scores that attempt to estimate achievement for the entire population, including students with disabilities and English learners who are excluded under current operational procedures. However, these results are presented as appendix material with the caveat that "the results of this special analysis should not be interpreted as official results" (NCES, 2018).

If the NAEP program were to disseminate results from analysis like that in Dogan (2019), careful consideration would need to be given about the timing, communication, and professional training about interpreting those scores. If they were made available to states via some secondary mechanism and not treated as official statistics, states and districts would still need help interpreting the results. Also, as states change their content standards across time, this would present ongoing issues about the development of the reports. Because the main NAEP is consistent across time, one of the primary values is the interpretability of the scores. Because any ancillary scores would depend on the current version of the state tests (that is, based on the current version of their standards versus NAEP's framework), the trends of these alternative scores could be hard to interpret.

Communications Efforts

NCES and Governing Board staff assist states in developing communications tools for each state release, which is an important and valuable service. If additional state-level information is made public by the NCES, even in appendix form, then those training efforts would need to be expanded. Although the basic concept of the reweighting analysis is not difficult to understand, it does take some specific attention to detail to understand the results. Without

some guidance, state and district leaders may reach different conclusions about the additional data. Assuming the TUDA results hold up at the state level, one state might assume that their "real" score is the reweighted score because it more closely matches what they are measuring on their state tests (and, presumably, teaching in their classrooms). Another state might conclude that it means the state should change the weights on its state tests to maximize their state NAEP results. Still another could assume that the NAEP blueprints should change, especially if that state closely matches the content emphasis of many other states.

Other Practical Considerations

The results of the initial Dogan (2019) reanalysis are interesting to examine. However, there are practical considerations for the NCES. If the studies are replicated at the national and state levels, would results from such analyses need to be made available at the same time as the main release? What about subjects other than mathematics? For all 50 states? For each TUDA district? This could amount to several hundred extra data products, as well as expensive and time-consuming special analyses of state content, unless some limits were applied. And again, because the results are dependent on state assessment blueprints, as those blueprints evolve, the studies would have to be updated to match the new blueprints. Finally, if results from the reanalysis are disseminated, it would be necessary to determine the appropriate standard errors for use in interpreting these scores.

CONCLUSION

The desire to employ the proposed reweighting strategy is understandable. Leadership in the districts that participate in the TUDA have the right to question whether NAEP is adequately measuring what their students are learning in the classroom. NCES's effort to examine reweighting as a way to monitor the validity of NAEP for the districts participating in the TUDA is a reasonable response to these questions. This situation certainly provides ample justification for investigating the reweighting procedure and considering its implementation.

The analysis provided by the Dogan (2019) study clearly supports the ongoing efforts by NCES to monitor the validity of NAEP. Evidence from the analysis has provided NCES with a reasonable approach to investigate the consequences of misalignment between NAEP subscale weights and the content emphasis of selected state assessments in terms of estimating score means. This analysis is a valuable complement to other validity studies, such as those conducted by the NVS Panel.

However, given strong interest in these results by TUDAs and other stakeholders, NCES is further faced with the question of whether and how to publicly report the results for each TUDA (and each state if the analysis is expanded). There are risks in choosing a path of public reporting. Some of the methodological issues and their implications have already been identified. However, a larger issue also needs to be considered: Is the implementation of a reweighting procedure likely to aid the NCES in accomplishing its goals?

The main risk factor lies in the violation of the principle that NAEP is not intended to represent any specific curriculum or instructional approach. Customizing NAEP using a reweighting procedure can be viewed as inconsistent with this assumption. Although it is true that the original NAEP scores would remain the primary official statistics, the program would now be producing several sets of alternative estimates based on local variations in curricula. This is a qualitatively different situation than the ancillary full population estimates discussed earlier.

Consider the scenario where a jurisdiction is provided with customized NAEP scores and informed that these scores are being provided because they offer better measurement of student instruction for that district. It seems likely that the jurisdiction would then emphasize the reweighted scores over the original NAEP scores. It also appears likely that the NCES would not be in a position from which it could dissuade the district from doing so. How could the agency that justified the provision of regenerated scores then argue these scores should not be given priority?

It also is important to consider the limitations noted in the Dogan study. First, since the reweighting method relies on data from Daro et al. (2015) and Daro et al. (in press), all limitations acknowledged in those studies apply to the current study as well. Second, the domain weights were developed based on 2017 data, so their application to 2013, 2015, and 2019 should be interpreted cautiously. Finally, the study analyzed data from nine of the 27 districts participating in the TUDA. If the NCES proceeds with implementing the reweighting strategy, it may be necessary to extend the analysis to more or all of the 27 districts.

The secondary analysis done for the NAEP TUDA scores is important and worthy of further exploration. It is important to know if trends in the TUDA scores might be a function of differences between NAEP content and the content of state assessments taken by the TUDA districts. Indeed, this is recurring theme of the NAEP validity studies. The Dogan analysis begins to provide an answer to that query. However, that information needs to be balanced by potential confusion that might be caused by an additional set of NAEP scores for some jurisdictions. For these reasons, it is the position of the NVS Panel that reweighting analyses of the sort exemplified in Dogan (2019) be limited to the role of a series of validity studies and not, in any way, be used in the reporting of any official statistics or even as a recurring set of ancillary results or appendix material. To the extent that there is a real and educationally significant mismatch between the content covered on NAEP and that in the states, the best way to ameliorate this is by modifying the NAEP frameworks, not through post hoc reweighting of the NAEP results. In the case of mathematics, the Governing Board has nearly completed an update of the framework for implementation in 2025 that will hopefully address the issue of alignment with newer state content frameworks comprehensively.

REFERENCES

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy. *Journal of Educational Measurement, 20*(2), 103–118.

Arreaga-Mayer, C., & Greenwood, C. R. (1986). Environmental variables affecting the school achievement of culturally and linguistically different learners: An instructional perspective. *NABE: The Journal for the National Association for Bilingual Education, 10*(2), 113–135.

Behuniak, P. (2015). *Maintaining the validity of the National Assessment of Educational Progress in a Common Core based environment.* San Mateo, CA: American Institutes for Research. Retrieved from https://www.air.org/sites/default/files/Validity-NAEP-Common-Core-Environment-March-2015.pdf

Brody, J., & Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 328–375). New York, NY: Macmillan.

Burstein, L. (1989, March). *Conceptual considerations in instructionally sensitive assessment.* Paper (Technical Report 333, Center for Research on Evaluation, Standards, and Student Testing, University of California–Los Angeles) presented at the annual meeting of the American Educational Research Association, San Francisco, California.

Council of Chief State School Officers. (1996). *Key state education policies on K–12 education: Content standards, graduation, teacher licenses, time and attendance.* Washington, DC: Author.

D'agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment, 12*(1), 1–22.

Darling-Hammond, L. (1993). Creating standards of practice and delivery for learning-center schools. *Stanford Law and Policy Review, 4,* 37–51.

Daro, P., Hughes, G. B., & Stancavage, F. (2015). *Study of the alignment of the 2015 NAEP mathematics items at grades 4 and 8 to the Common Core State Standards for Mathematics.* San Mateo, CA: American Institutes for Research. Retrieved from https://www.air.org/sites/default/files/downloads/report/Study-of-Alignment-NAEP-Mathematics-Items-common-core-Nov-2015.pdf

Daro, P., Hughes, G. B., Stancavage, F., Shepard, L., Kitmitto, S., Webb, D., & Tucker-Bradway, N. (in press). *A comparison of the 2017 NAEP mathematics assessment with current-generation state assessments in mathematics: Expert judgment study.* San Mateo: CA: American Institutes for Research.

Debra P. v. Turlington, 644 F.2d. 397 (U.S. Ct. App. 5th Cir. 1979).

Dogan, E. (2019). *Analysis of recent NAEP TUDA mathematics results based on alignment to state assessment content.* Unpublished manuscript.

Hughes, G. B., Daro, P., Holtzman, D., & Middleton, K. (2013). *A study of the alignment between the NAEP mathematics framework and the Common Core State Standards for Mathematics (CCSS-M)*. Palo Alto, CA: American Institutes for Research. Retrieved from https://www.air.org/sites/default/files/downloads/report/NVS_combined__study_1_NAEP_alignment_with_CCSS_0.pdf

Leinhart, G. (1983). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18(2),* 85–96.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332–1361.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4–12*. Boston, MA: Boston College Center for the Study of Testing, Evaluation, and Educational Policy. Retrieved from https://files.eric.ed.gov/fulltext/ED370772.pdf

National Assessment Governing Board. (2012). *Eligibility criteria and procedures for selecting districts for participation in the National Assessment of Educational Progress: Trial Urban District policy statement*. Washington, DC: Author.

National Assessment Governing Board. (2017). *Mathematics framework for the 2017 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education. Retrieved from https://www.nagb.gov/assets/documents/publications/frameworks/mathematics/2017-math-framework.pdf

National Center for Education Statistics. (2018). *Full population estimates* [Website]. Retrieved from https://nces.ed.gov/nationsreportcard/about/fpe.aspx

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

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Author. Retrieved from http://www.corestandards.org/read-the-standards/

O'Day, J. A., & Smith, M. S. (1993). Systemic reform and educational opportunity. In S. H. Fuhrman (Ed.), *Designing coherent education policy* (pp. 250–312). San Francisco, CA: Jossey-Bass.

Oakes, J., & Guiton, G. (1995). Opportunity and conceptions of educational equality. *Educational Evaluation and Policy Analysis, 17*(3), 323–336.

Reese, C. M., Miller, K. F., Mazzeo, J., & Dossey, J. A. (1997). *National Assessment of Educational Progress 1996 mathematics report card for the nation and the states.* Washington, DC: National Center for Education Statistics.

Schmidt, W. H. (1992). The distribution of instructional time to mathematical content: One aspect of opportunity to learn. In L. Burstein (Ed.). *The IEA study of mathematics III: Student growth and classroom processes* (pp. 129–145). New York, NY: Pergamon Press.

Stancavage, F., & Bohrnstedt, G. (2013). *Examining the content and context of the Common Core State Standards: A first look at implications for the National Assessment of Educational Progress.* San Mateo, CA: American Institutes for Research. Retrieved from https://www.air.org/sites/default/files/downloads/report/NAEP_Validity_Studies_combined_report_updated_9-19-13_0.pdf

Stancavage, F., Shepard, L., McLaughlin, D., Holtzman, D., Blankenship, C., & Zhang, Y. (2009). *Sensitivity of NAEP to the effects of reform-based teaching and learning in middle school mathematics.* Palo Alto, CA: American Institutes for Research. Retrieved from https://www.air.org/sites/default/files/downloads/report/NVS_NAEP_Sensitivity_to_Instruction_8-26-09_0.pdf

U.S. Department of Education. (2003). *NAEP validity studies: An agenda for NAEP validity research* (NCES 2003–07). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2003/200307.pdf

Valencia, S. W., Wixson, K. K., Kitmitto, S., & Doorey, N. (in press). *A comparison of NAEP reading and NAEP writing assessments with current-generation state assessments in English language arts: Expert judgment study.* San Mateo, CA: American Institutes for Research.

Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational Evaluation and Policy Analysis, 20,* 137–156.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Washington, DC: Council of Chief State School Officers. Retrieved from http://facstaff.wceruw.org/normw/WEBBMonograph6criteria.pdf

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Washington, DC: Council of Chief State School Officers. Retrieved from https://files.eric.ed.gov/fulltext/ED440852.pdf

Winfield, L. F. (1991). Resilience, schooling and development among African American youth [Special Issue]. *Education and Urban Society, 24*(1), 5–14.

Winfield, L. F. (1993). Investigating test content and curriculum overlap to assess opportunity to learn. *Journal of Negro Education, 62*(3), 288–310.

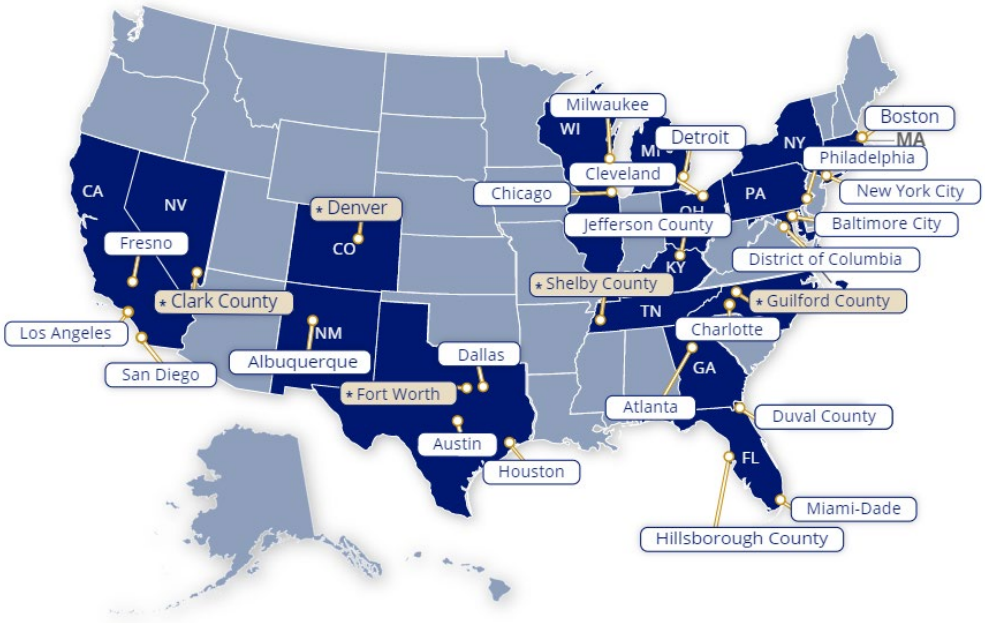
APPENDIX: ANALYSIS OF RECENT NAEP TUDA MATHEMATICS RESULTS BASED ON ALIGNMENT TO STATE ASSESSMENT CONTENT[4]

Enis Dogan
National Center for Education Statistics

Background

The National Assessment of Educational Progress (NAEP) provides an essential measure of student achievement in the United States. In addition to national and the state-level assessments, since 2002, NAEP also reports student achievement for selected urban districts in mathematics, reading, science, and writing. These are known as Trial Urban District Assessments (TUDAs). In 2017, 27 districts participated in mathematics and reading assessments at Grades 4 and 8 (Figure 1). In mathematics assessments, 20 and 23 of these districts scored significantly lower than the national public mean at Grades 4 and 8, respectively. Between 2003 and 2017, of 112 comparisons between adjacent years across participating TUDAs, there were 12 significant decreases at Grade 4 (Table 1); 11 of these were observed in 2015 or 2017. Similarly, of the five significant decreases during the same period between adjacent years at Grade 8, four were observed in 2015 or 2017 (Table 1).

Figure 1. Twenty-Seven Urban Districts That Participated in NAEP 2017 Mathematics and Reading Assessments at Grades 4 and 8



Note. Label boxes with a beige background indicate that the district began participating in the TUDA assessments in 2017.

[4] NOTE: The analysis presented in this Appendix is authored by Dr. Enis Dogan of the National Center of Education Statistics and any opinions expressed are those of the author. Although the analysis is not a product of the NAEP Validity Panel, it is included as reference and with approval from the author.

Table 1. Changes in TUDA Means Between Adjacent Administrations: 2003 to 2017 Grade 4 and Grade 8 Mathematics Assessments

Grade 4	2003 to 2005	2005 to 2007	2007 to 2009	2009 to 2011	2011 to 2013	2013 to 2015	2015 to 2017	Total
Significant increase	8	4	2	4	4	3	4	29
No change	2	6	9	14	17	10	13	71
Significant decrease	0	1	0	0	0	7	4	12
Total	10	11	11	18	21	20	21	112
Grade 8								
Significant increase	4	6	2	6	3	1	0	22
No change	6	5	9	12	17	16	20	85
Significant decrease	0	0	0	0	1	3	1	5
Total	10	11	11	18	21	20	21	112

Note. Number of significant decreases are printed in red.

These relatively negative trends in recent years coincide with many states' implementation of new assessments aligned to recently adopted college and career ready standards in mathematics, raising a legitimate question: Could these trends be a function of the differences between the contents of NAEP and state assessments? This study aims to answer this question. More specifically, the research questions are as follows:

1. How would the 2017 mathematics Grades 4 and 8 TUDA mean scores change if the NAEP subscales were weighted according to the content distribution of selected state assessments?
2. How would the mathematics Grades 4 and 8 TUDA mean scores change in 2013, 2015, and 2019 if the NAEP subscales were weighted according to the content emphasis of selected state assessments, assuming that the content emphasis of those assessments and NAEP were similar in these years compared with 2017?

Analyses rely on data from Daro et al. (in press), who compared items from the 2017 NAEP and selected state assessments in terms of content distribution among other features.

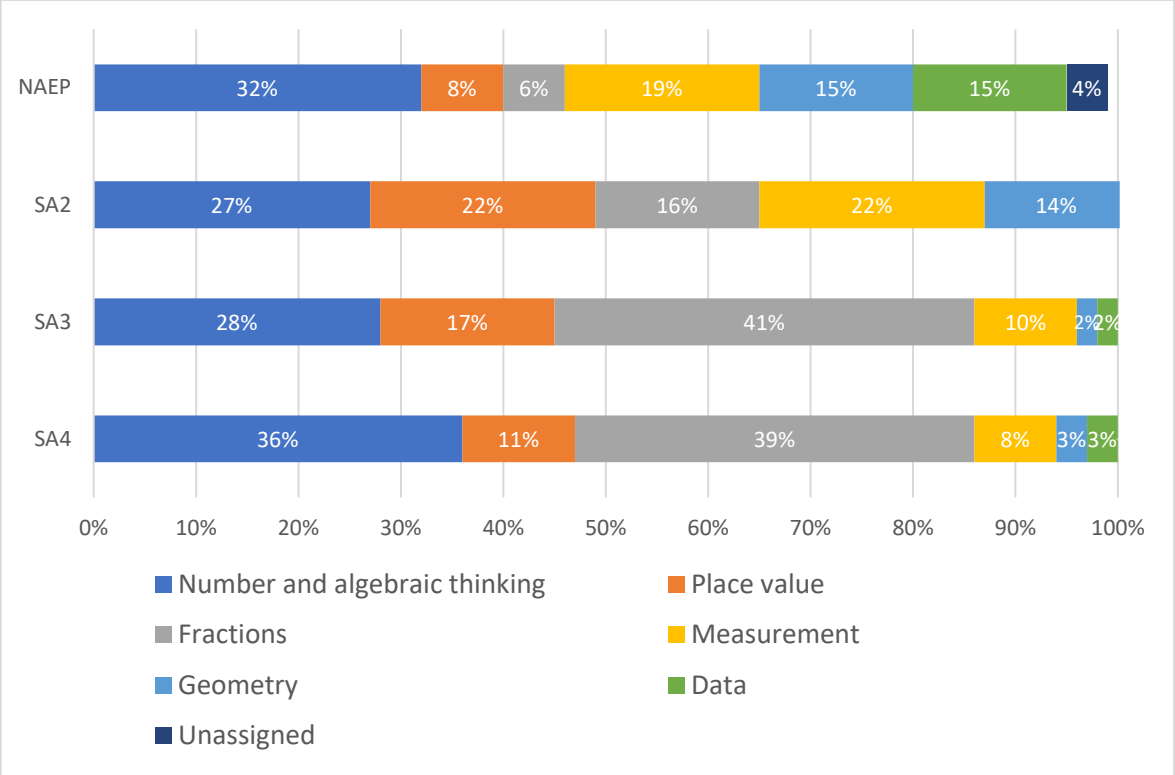
Reweighting NAEP Mathematics Subscales According to the Content Distribution of Selected State Assessments

The NAEP mathematics scale scores are computed as a weighted average of five subscales that make up the mathematics assessments: (a) Number Properties and Operations (Numbers); (b) Measurement; (c) Geometry; (d) Data Analysis, Statistics, and Probability (Data); and (e) Algebra. The relative weight of each subscale is specified in the mathematics framework. In this study, we investigate if/how the 2017 mathematics Grades 4 and 8 TUDA mean scores would change if the subscales were weighted according to the content distribution of selected state assessments instead. Data on the content distribution of NAEP and selected state assessments (Figures 2 and 4) come from Daro et al. (in press). In addition to the information on the overall content distribution, classification of each NAEP item in terms of the five domains in Daro et al. also was acquired from the researchers. As a result, classification of each individual NAEP item in terms of both the Daro et al. scheme and the NAEP framework was available for analyses discussed below.

Grade 4 Results

At Grade 4, Daro et al. (in press) classified items from the 2017 NAEP and four state assessments into six content "domains": (a) Number and Algebraic Thinking, (b) Place Value, (c) Fractions, (d) Measurement, (e) Geometry, and (f) Data. These domains do not perfectly correspond to the five subscales featured in NAEP. Nine of the 2017 TUDA districts participated in one of the three state assessments (SA) in Daro et al.: SA2 through SA4 (names withheld for confidentiality).

Figure 2. Content Distribution of 2017 Grade 4 NAEP and Selected State Mathematics Assessments According to the Daro et al. (in press) classification scheme



Note. Classifications were done at the item level but then weighted according to the contribution of each item to the assessment total score (i.e., the score points assigned to each item). The percentages in the figure are based on the proportion of the total score. Optional assessment components were not included in the analyses (Daro et al., in press).

Weights were computed for the five NAEP subscales in a way that they reflect the content emphasis of each of the three state assessments in Figure 2, one at a time, using the following the steps:

1. Assign a weight to each NAEP item *i* in each domain (w_i) by dividing the percentage of points in the state assessment for the domain the item belongs to (according to the Daro et al. classification scheme) by the same percentage in NAEP[5] (Figure 2).
2. Find the raw weight for each subscale *s* by summing the item weights across all items in that subscale ($W_s = \sum w_i$).

[5] For example, in computing weights relative to SA2, all NAEP items classified as Number and Algebraic Thinking by Daro et al. (in press) received a weight of 0.84 (.27/.32).

3. Find the sum of all subscale weights (TotW = $\sum W_s$).
4. Compute a final weight for each NAEP subscale by dividing W_s by TotW.

The resulting sets of NAEP subscale weights, computed relative to each of the three state assessments, are shown in Table 2. As seen in the table, the relative weight of the Numbers subscale increased in computed weights because state assessments put more emphasis on fractions (Figure 2), which feed into the Numbers subscale in NAEP. On the other hand, the relative weight of the Data subscale decreased in computed weights because state assessments put less emphasis on items that feed into NAEP's Data subscale.

Table 2. Subscale Weights Relative to State Assessments and According to the NAEP Framework: Grade 4 Mathematics

	Numbers	Measurement	Geometry	Data	Algebra
Weight in NAEP framework	*40%*	*20%*	*15%*	*10%*	*15%*
Weight relative to SA2	54%	18%	15%	0%	14%
Weight relative to SA3	73%	9%	2%	1%	14%
Weight relative to SA4	71%	8%	3%	0%	18%

To address the first research question, reweighted composite mean scores (2017) were computed for nine TUDAs that take either SA2, SA3, or SA4 as their state assessment. The expectation was that composite mean scores in each TUDA would improve when subscale weights were computed in a way that they mirror the content emphasis of the state assessment each TUDA takes.

The subscale weights in computing these reweighted means for a given TUDA came from the relative weights (Table 2) computed in relation to that TUDA's state assessment. For example, the weights in computing the reweighted mean for TUDA1 came from the weights relative to SA2, whereas the weights for TUDA2 came from those computed relative to SA3, and so on. Positive changes indicate that the TUDA composite mean went up when subscale weights were computed in a way that they reflected the content emphasis of the state assessment the district takes. All nine TUDAs showed positive changes in composite mean scores, ranging from 1.1 (TUDA1) to 4.6 (TUDA7) scale score points (Table 3).[6]

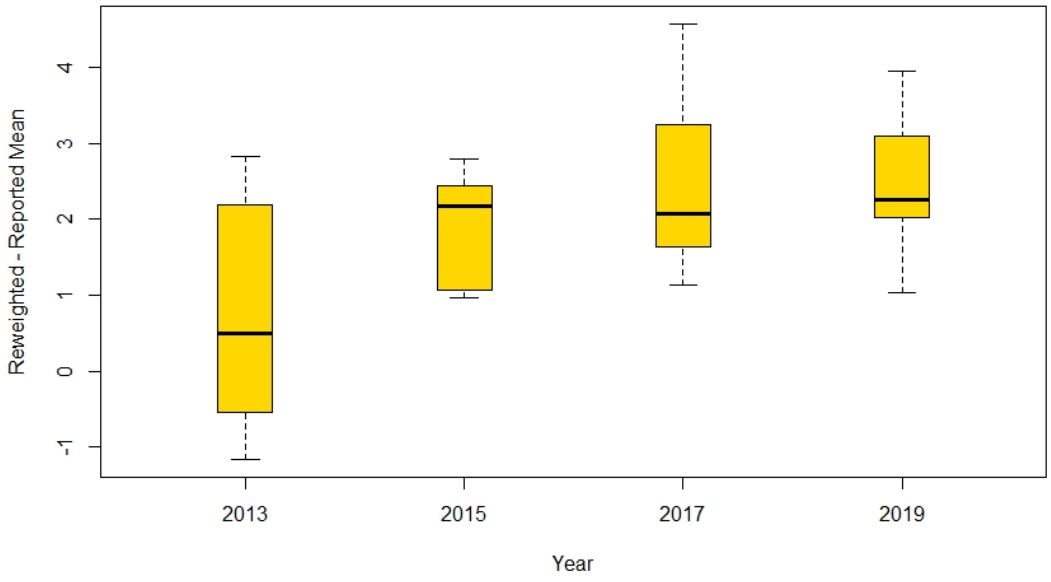
[6] These differences were not tested for statistical significance.

Table 3. Differences Between Reweighted and Reported TUDA Means: 2017 Grade 4 Mathematics

	State assessment	Reweighted – reported
TUDA1	SA2	1.1
TUDA2	SA3	1.4
TUDA3	SA3	3.0
TUDA4	SA3	2.1
TUDA5	SA3	2.1
TUDA6	SA4	1.6
TUDA7	SA4	4.6
TUDA8	SA4	3.3
TUDA9	SA4	3.5

To address the second research question, the difference between reweighted and reported means also were computed for 2013, 2015, and 2019 administrations for the same nine districts. The median difference across these districts was 0.49 in 2013, 2.18 in 2015, 2.08 in 2017, and 2.3 in 2019 (Figure 3).

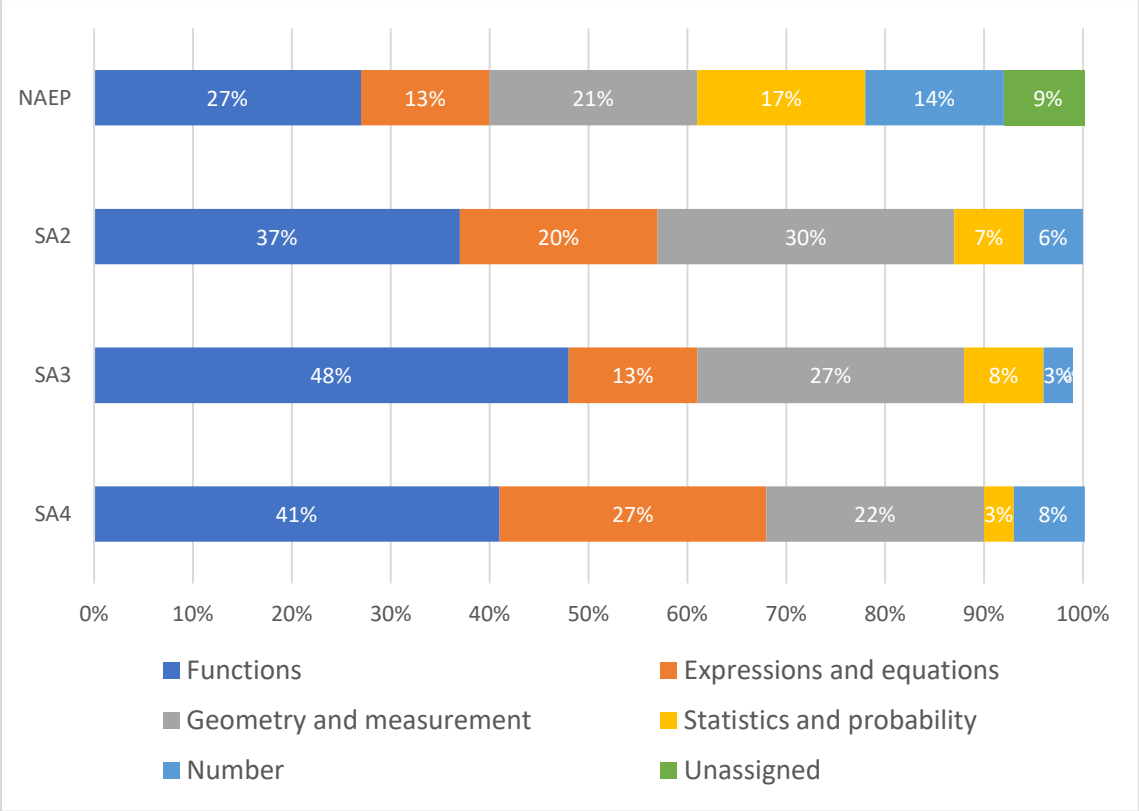
Figure 3. Differences Between Reweighted and Reported Means Across Nine TUDAs by Year: Grade 4 NAEP Mathematics Assessment



Grade 8 Results

At Grade 8, Daro et al. (in press) classified items into five domains: (a) Functions, (b) Expressions and Equations, (c) Geometry and Measurement, (d) Statistics and Probability, and (e) Numbers. As seen in Figure 4, all three state assessments put more emphasis on functions and less emphasis on data compared with NAEP at Grade 8. Weights for the Grade 8 NAEP subscales were computed following the same steps described for Grade 4. The resulting three sets of subscale weights, computed relative to the state assessments, are displayed in Table 4 along with the weights according to the NAEP framework.

Figure 4. Content Distribution of 2017 Grade 8 NAEP and Selected State Mathematics Assessments According to the Daro et al. (in press) classification scheme



Note. Classifications were done at the item level but then weighted according to the contribution of each item to the assessment total score (i.e., the score points assigned to each item). The percentages in the figure are based on the proportion of the total score. Optional assessment components were not included in the analyses (Daro et al., in press).

As seen in Table 4, the relative weight of the Algebra subscale increased in computed weights because state assessments put more emphasis on functions that feed into the Algebra subscale in NAEP. On the other hand, the relative weight of the Data subscale decreased in computed weights because state assessments put less emphasis on items that feed into NAEP's Data subscale.

Table 4. Subscale Weights Relative to State Assessments and According to the NAEP Framework: Grade 8 Mathematics

	Numbers	Measurement	Geometry	Data	Algebra
Weight in NAEP framework	20%	15%	20%	15%	30%
Weight relative to SA2	16%	19%	19%	7%	39%
Weight relative to SA3	14%	18%	19%	7%	42%
Weight relative to SA4	21%	16%	16%	2%	45%

To address the first research question, reweighted composite mean scores (2017) were computed for nine TUDAs that take SA2, SA3, or SA4 as their state assessment. The subscale weights in computing these reweighted means for a given TUDA came from the relative weights (Table 4) computed in relation to that TUDA's state assessment. All nine

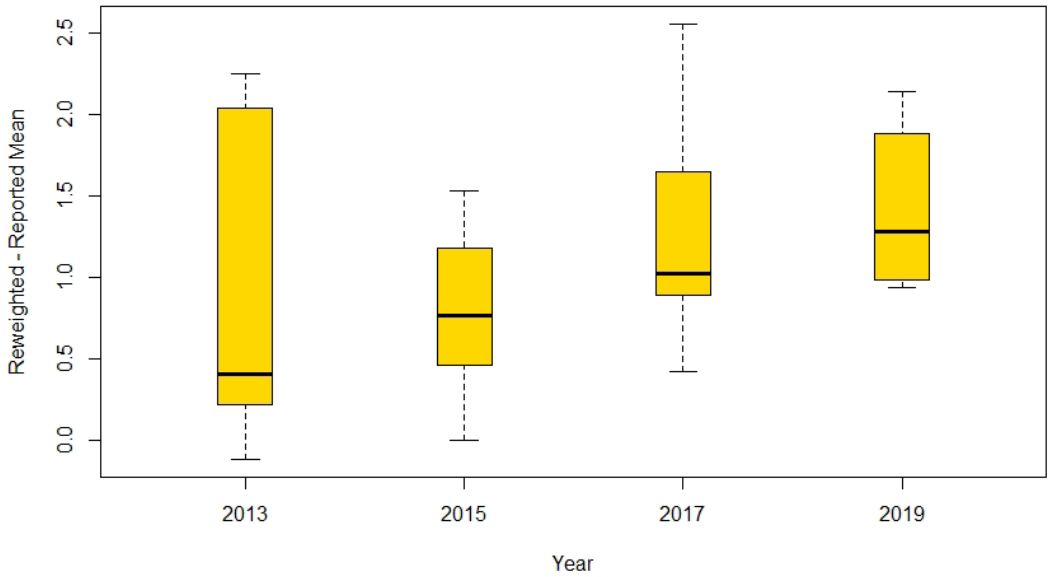
TUDAs showed positive changes, indicating that the TUDA means went up when subscale weights were computed in a way that reflected the content emphasis of the state assessment the district takes (Table 5). These changes ranged from 0.4 (TUDA3) to 2.6 (TUDA8) scale score points.[7]

Table 5. Differences Between Reweighted and Reported TUDA Means: 2017 Grade 8 Mathematics

	State assessment	Reweighted – reported
TUDA1	SA2	0.9
TUDA2	SA3	0.8
TUDA3	SA3	0.4
TUDA4	SA3	1.0
TUDA5	SA3	0.9
TUDA6	SA4	1.6
TUDA7	SA4	1.7
TUDA8	SA4	2.6
TUDA9	SA4	1.4

To address the second research question, the difference between reweighted and reported means was also computed for the 2013, 2015, and 2019 administrations for the same nine districts. The median difference across these districts was 0.41 in 2013, 0.76 in 2015, 1.02 in 2017, and 1.28 in 2019 (Figure 5).

Figure 5. Differences Between Reweighted and Reported Means Across Nine TUDAs by Year: Grade 8 NAEP Mathematics Assessment



In sum, at both grades, the results showed positive changes in the TUDA means, for all cases in 2015, 2017, and 2019 and in majority of cases in 2013, when subscale weights were

[7] These differences were not tested for statistical significance.

adjusted in a way that mirror the content emphasis of the state assessment each TUDA takes. The median difference between reweighted and reported means tended to increase between 2013 and 2015 and then leveled off at Grade 4, and it tended to increase in each subsequent year since 2013 at Grade 8.

Discussion

NAEP frameworks are not meant to be aligned to any particular state's curriculum or learning standards. On the other hand, NAEP has always sought a balance between maintaining the essence of the constructs being measured to be able to report on trends and reflecting changes in the educational objectives and curricula across the country in its assessment frameworks. Keeping this balance has become arguably more challenging recently as many states began implementing new college and career ready standards in mathematics and English language arts/literacy, along with assessments aligned to these new standards. Shifts in the standards and assessments have obvious implications for NAEP. Daro et al. (in press) show that there are important differences between the content emphasis of NAEP and three state assessments used in nine of the 27 urban districts participating in the NAEP assessments. In general, the state assessments examined in Daro et al. put greater emphasis on Numbers at Grade 4 and Algebra at Grade 8 and less emphasis on Data at both grades. Building on these data, the current study showed that when the weights of subscales in computing NAEP composite means were adjusted to reflect the content emphasis of the aforementioned state assessments, the 2017 NAEP mean for all nine districts that use these assessments went up at both grades. This also was true when the same weights were applied for the 2015 and 2019 assessments. When applied to the 2013 assessments, however, this was not the case. Three districts showed a negative change at Grade 4, and one showed a negative change at Grade 8 when the subscale weights were adjusted as described earlier. This change might be due to differences in content emphasis in state assessments in 2013 compared with 2017. It also is possible that the differences in content emphasis between state assessments and NAEP mattered less in earlier years when the states had transitioned to new assessments aligned to new standards only recently.

This study has several limitations. Because the study relies on data from Daro et al. (2015) and Daro et al. (in press), all limitations acknowledged in those studies apply to the current study as well. In addition, an important limitation to the reweighting method is that the weights were applied to scores obtained from existing NAEP item pools without removing or adding actual content to these pools. In this regard, the reweighting method provides only a proxy to results we would expect to obtain if the NAEP item pools were reshaped to reflect the content emphasis of state assessments. Furthermore, the fact that data from only nine of the 27 districts participating in the NAEP TUDA program were examined in the study limits the generalizability of the findings to all districts. In addition, findings regarding 2013, 2015, and 2019 data should be interpreted more cautiously because the subscale weights applied to the scores from these years were derived from the content analysis of the 2017 NAEP and state assessments.

References

Daro, P., Hughes, G. B., & Stancavage, F. (2015). *Study of the alignment of the 2015 NAEP mathematics items at grades 4 and 8 to the Common Core State Standards for Mathematics.* San Mateo, CA: American Institutes for Research. Retrieved from https://www.air.org/sites/default/files/downloads/report/Study-of-Alignment-NAEP-Mathematics-Items-common-core-Nov-2015.pdf

Daro, P., Hughes, G. B., Stancavage, F., Shepard, L., Webb, D., Kitmitto, S., & Tucker-Bradway, N. (in press). *A comparison of the 2017 NAEP mathematics assessment with current-generation state assessments in mathematics: Expert judgment study.* San Mateo: CA: American Institutes for Research.