Teachers' Views of Their Practices Related to Common Core State Standards-Aligned Assessments

ETS RR–19-40

Heather M. Buzick
Anna Rhoad-Drogalis
Cara C. Laitusis
Teresa C. King

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Teachers' Views of Their Practices Related to Common Core State Standards-Aligned Assessments

Heather M. Buzick,[1] Anna Rhoad-Drogalis,[2] Cara C. Laitusis,[1] & Teresa C. King[1]

1 Educational Testing Service, Princeton, NJ
2 Ohio State University, Columbus, OH

A fundamental claim for Common Core State Standards (CCSS)-aligned assessments is that they will lead to better teaching practices. The purpose of this study is to seek evidence in support of this claim by surveying teachers about their instructional practices, test preparation strategies, and test score use both before and after the introduction of CCSS-aligned assessments. Baseline and trend data were collected via five Web-based surveys, administered over 2 years to elementary and middle school English language arts and mathematics teachers in one state, New Jersey. Responses to the first three surveys ($n_1 = 402$ teachers, $n_2 = 469$ teachers, and $n_3 = 175$ teachers from 4% to 6% of New Jersey schools) are summarized and described; results from the remaining surveys are omitted due to low response. Challenges to collecting empirical evidence in support of the validity argument and theory of action for a new assessment are discussed.

Keywords Common Core State Standards; assessment; validity argument; teachers' practices

doi:10.1002/ets2.12277

A new set of K – 12 education standards in English language arts (ELA), literacy, and mathematics, the Common Core State Standards (CCSS; http://www.corestandards.org), was released in the United States in 2010 (see Porter, McMaken, Hwang, & Yang, 2011). Subsequently, the development of assessment systems aligned to the CCSS commenced. A fundamental claim for, or desired consequence of, CCSS-aligned assessments is that they would lead to better teaching practices by measuring standards that "encourage best practices in teaching and learning" and by "providing meaningful feedback to educators" (Conley, 2014, pp. 7 – 8). This claim can be evaluated via Kane's (2006) argument-based approach to validity and Bennett's (2010) application of theory of action to educational assessment, which drive the collection of evidence on the use and interpretation of assessment scores as well as the consequences of assessment use itself.

Bennett, Kane, and Bridgeman (2011) identified several areas of needed research in the context of building a validity argument and developing a theory of action for CCSS-aligned assessments. Of the seven recommendations in the report, we aimed to address two in the current study through the collection of survey data from teachers both prior to administration of CCSS-aligned assessments and then again during the first and second operational administrations. The two recommendations were (a) "Collect data from key stakeholders documenting how assessment results are used, noting both intended and unintended consequences of score use" and (b) "To help in making the case for changes in teaching and learning practice postulated by the theory of action, begin collecting data now so that existing practices can be documented" (p. 4).

The purpose of the current study was to measure teachers' views of their instructional practices, test preparation strategies, and test score use both before and after administration of CCSS-aligned assessments to estimate changes associated with the introduction of the assessments. In 2014, we invited Grades 3 – 8 mathematics and ELA teachers across the state of New Jersey to complete a survey following the administration of the state-developed summative assessment. In 2015, we targeted the same population of teachers to invite to complete a survey with the same questions as the 2014 survey, after the CCSS-aligned assessments were administered. We again invited teachers to participate in the survey in 2016, the second year that the CCSS-aligned assessments were administered. In the course of collecting empirical evidence in support of the validity argument and theory of action for a new assessment, we met substantial research design and measurement challenges. Herein we describe the study, summarize the results and the shortcomings of the data we collected, and reflect on both the measurement challenges and the possibility of innovative ways to evaluate the consequences of new educational assessments.

Corresponding author: H. M. Buzick, E-mail: hbuzick@ets.org

Background

By 2013, all states had adopted either the CCSS or state-specific college- and career-ready standards, with more than 40 states choosing to adopt the CCSS (Achieve, 2013). The federal government funded several consortia to design and develop assessments aligned to the CCSS. Two consortia, the Partnership for the Readiness of College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (Smarter Balanced), developed assessments aligned with the CCSS in mathematics and ELA and literacy to administer to all students except those with the most significant cognitive disabilities. Those assessments were first administered operationally during the 2014–2015 school year. Previously, states developed their own summative assessments, which were aligned to their own state standards. In the years between June 2010, when the CCSS were finalized, and 2014–2015, when the first PARCC and Smarter Balanced assessments were administered, some states, including New Jersey, were transitioning their state assessment to partially align with the CCSS (New Jersey Department of Education, n.d.).

Both the PARCC and Smarter Balanced assessments were designed following criteria for *high-quality assessments* (Council of Chief State School Officers [CCSSO], 2014). This criteria included the aim of "providing timely data that inform instruction" (CCSSO, 2014, p. 14) and assessing higher order thinking aligned with college- and career-ready standards (CCSSO, 2014, Sections B.4 and C.4). An independent evaluation by Herman and Linn (2013) of the assessments from PARCC and Smarter Balanced found that the development plans and assessment claims would likely support assessments of *deeper learning* (their definition was grounded in Norman Webb's depth of knowledge classification scheme; Webb, Alt, Ely, Cormier, & Vesperman, 2005).

Despite the intended benefits of the CCSS-aligned assessments, after pilot testing in the 2013–2014 school year, some educators and the public raised specific concerns about the tests. For example, Kaufman et al. (2015) surveyed individuals from the American Teacher and American School Leader Panels in February 2015. Teachers' main concern about the tests that were about to be administered to their students was that they were going to be too difficult. This is in contrast to a report by McClellan, Joe, and Bassett (2015), commissioned by the National Network of State Teachers of the Year. They reported perspectives that a multistate sample of State Teachers of the Year had regarding the CCSS-aligned assessment administered in their state and the previous state summative assessment. Overall, the teachers reported positive opinions about content and construct validity as well as depth of knowledge. One of the state assessments was the New Jersey Assessment of Skills and Knowledge (NJ ASK), New Jersey's state summative assessment prior to PARCC. Teachers in the sample also responded positively to NJ ASK.

Prior to the CCSS, several survey studies had been conducted on teachers' attitudes and responses to assessments. Work by Hoover and Abrams (2013) and Vogler (2002) investigated how teachers' use of data from summative assessments used to make high-stakes decisions impacts their classroom instruction. A national study of teachers conducted by the National Board on Educational Testing and Public Policy examined how teachers' views on state-mandated testing programs varied depending on whether they taught in states with assessments used to make high- or low/moderate-stakes decisions (Abrams, Pedulla, & Madaus, 2003). Firestone, Schorr, and Monfils (2004) conducted a 3-year study focusing on teachers' attitudes toward summative assessments and on understanding how teachers modified their instructional practices and test preparation in response to New Jersey's statewide assessment. Returning to the CCSS, Hamilton et al. (2016) and Kaufman et al. (2016) reported findings from a February 2015 survey of the American Teacher Panel and the American School Leader Panel. Their nationally representative teacher sample reported on their familiarity with the CCSS in mathematics and ELA and their professional development needs in specific areas. Furthermore, Edgerton and Desimone (2018) surveyed teachers in three states on the extent that CCSS-aligned content was covered in their classes. Although results from surveys on perceptions and behaviors related to the CCSS have been published, we are not aware of any studies that have collected direct evidence of changes in trends in teachers' instructional practices as they relate to CCSS-aligned assessments.

Method

Our current study was driven by three categories of research questions:

1 Did teachers' instructional practices related to teaching assessed material in mathematics and ELA change from immediately before to after administration of the new CCSS-aligned assessments? [Instructional practices]

Table 1 Survey Administration Schedule

Survey topics	Rounds of data collection	Number of teacher respondents/schools
Instructional and test preparation practices (*instructional practices*)	1. Summer 2014 (Baseline, NJ ASK)	402/122[a]
	2. Summer 2015 (Trend 1, PARCC)	175/75[a]
	3. Summer 2016 (Trend 2, PARCC)	81/38
Test score use/attitudes related to validity (*score use*)	1. Fall 2014 (Baseline, NJ ASK)	469/137[a]
	2. January and June 2016[b] (Trend 1, PARCC)	68/38

Note: NJ ASK = New Jersey Assessment of Skills and Knowledge; PARCC = Partnership for the Readiness of College and Careers.
[a]Results are reported from this survey. [b]Delays in PARCC score reports delayed the administration of this survey from Fall 2015 to January and June 2016.

2 What is the frequency and type of test preparation practices that teachers are engaged in? Did test preparation practices change immediately before and after administration of the CCSS-aligned assessments? [Test preparation practices]
3 How do teachers use score reports, and do they think that the test scores measure what they are intended to measure? Did these behaviors and perceptions change from immediately before to after administration of the CCSS-aligned assessments? [Test score use/attitudes related to validity]

We collected data over multiple years (see Table 1), prior to the administration of CCSS-aligned assessments and after, to answer these research questions. We intended to administer one survey on all three topics; however, there was a large time gap between the test administration and when teachers were able to view individual student score reports. Consequently, we scheduled separate survey administrations with instructional and test preparation practices combined into one survey and test score use/attitudes related to validity combined into another survey. Herein, we refer to the instructional and test preparation practices survey as the *instructional practices survey* and the test score use/attitudes related to validity as the *score use survey*.

Survey Development and Administration

The surveys were developed for this study through collaboration with personnel in the NJDOE and researchers with expertise in validity theory, teaching, and survey development. Some items were modified from the teacher questionnaire used by Firestone et al. (2004). The survey questions were piloted with a focus group of four teachers, two staff members in the assessment division of the NJDOE, one expert in assessment development and policy related to the CCSS, and two researchers with expertise in teaching. We revised the surveys based on the collective feedback prior to the Round 1 administration.

The instructional practices survey included items on instructional practices (36 items), test preparation practices (22 items), teacher sentiment (9 items), and background (9 items). The score use survey included the same background questions (9 items) as well as items on attitudes about the validity of overall math scores (8 items), math subscores (15 items), overall ELA scores (8 items), ELA subscores (11 items), test score use (36 items), and performance-level descriptors (1 item). Most of the items were Likert-type items in which teachers were asked to specify their level of agreement or frequency of behavior. Respondents were asked to explain or expand on select answers in several open-ended questions. Each survey took about 10 min to complete, based on the results from focus group participants. The surveys are available online, and the items appear in the results section.

For each administration, we sent e-mail invitations with unique online survey links to all New Jersey public elementary and middle school principals. Principals were asked to forward their unique survey links to teachers in their schools who taught mathematics and ELA in Grades 3 – 8. Principals received four contacts: an introductory e-mail and three reminder e-mails. For each survey administration, to encourage participation, we used a lottery incentive; upon completion of the survey, participants could choose to be entered into a lottery in which three people were randomly selected to win $200. We administered each round of the instructional practices survey immediately after the state testing window. We administered each round of the score use survey immediately after districts shared student test scores with teachers.

Sample

Our target population was all public school teachers in New Jersey who taught mathematics and/or ELA/reading to students in Grades 3–8 and had at least half of their students participate in the regular state assessment. Across the five data collections, respondents came from 282 schools, approximately 15% of eligible schools. Table 1 shows the number of teachers who responded to each survey as well as the number of schools represented.

We do not have a direct measure of the teacher response rate, because we could not observe which principals forwarded the e-mail to their teachers, nor did we have information on the total number of eligible teachers. Based on the unique survey links sent to principals, for our two baseline surveys, 6% of eligible New Jersey schools were represented by one or more teachers in the teacher practice survey, and 7% of eligible schools were represented by teachers' responses to the test score survey. This return is lower than an expected 10%–15% response rate from directly contacting teachers and not providing a tangible incentive. After each e-mail blast, approximately 90 principal e-mail addresses were returned as undeliverable, and we received another 10–20 automatic out-of-office replies. Excluding them raises each response rate by about a percentage point. Respondents to our first trend survey on instructional practices, administered after the PARCC assessments were introduced, represented just 4% of eligible New Jersey schools. Respondents to the second trend year for instructional practices and the first trend year for score use came from approximately 2% of eligible schools. Due to the very low response rates for our final two surveys, we report on the first three surveys: baseline instructional and test preparation practices, baseline score use, and instructional and test preparation practices the year after PARCC was first administered (see Table 1). We were not able to collect a sufficient number of responses on the score use survey after PARCC was first administered nor on the instructional and test preparation practices survey after the second year that PARCC was administered. We cannot report longitudinal trends in score use; therefore we report the pre-PARCC score use data as a snapshot.

Among respondents to the baseline score use survey, 31.35% taught ELA, 31.6% taught math, and the remaining 35.4% taught both math and ELA. Teaching experience was similar to the baseline instructional practice survey sample with respondents averaging 13.2 years ($SD = 8.2$) of teaching full-time in the state and 7.7 years ($SD = 6.6$) of teaching the same grade level. In participants' schools, on average, 28.2% ($SD = 29.2\%$) of students qualified for free or reduced-price lunch, which was lower than the statewide average of 32.7% based on a one-sample z-test ($\mu = 32.7, z = -3.3, p < .001$). There was little overlap between the baseline instructional practices survey and the baseline score use survey in the schools in which teacher respondents taught. Specifically, 5% of schools in the baseline instructional practices survey were represented in the baseline score use survey.

Analysis and Results

We employed descriptive statistics and graphical displays of response data to describe response frequencies immediately before and after the introduction of the PARCC assessment in New Jersey. Our numerical and graphical summaries of the data are at the item level.

Among the 122 schools represented in the baseline instructional practices survey respondent sample, 25% were represented by the teachers who responded to the Trend 1 instructional practices survey. We adjusted the post-PARCC survey responses to account for differences among survey respondents between the two surveys of instructional and test preparation practices. We used the R package ebal (Hainmueller, 2014) to use entropy weighting (Hainmueller, 2012) to weight the second year of data from the instructional practices survey. We used the following set of school-level covariates from both the baseline and Trend 1 year: percentage of students receiving free or reduced-price lunch, total number of students, percentage of students proficient in mathematics and ELA, student to teacher ratio, and suspension rate. School-level covariates from both baseline and trend sample years were used in the weighting. We chose to match on the first moments of the school-level covariates from the baseline sample of schools, because including the second moments only trivially moved the Trend 1 covariate means closer to the baseline covariates.

Because we weighted the 1 year trend data back to the original baseline instructional practices survey sample, here we describe the characteristics of teachers and schools from the baseline sample. Among the baseline instructional practices survey respondents, 28.2% taught ELA only, 25.5% taught math only, and the remainder (46.3%) taught both subjects. Respondents averaged 13.0 years ($SD = 8.6$) as full-time teachers in the state and 7.3 years ($SD = 6.3$) teaching the same grade. The mean percentage of students in respondents' schools who qualified for free or reduced-price lunch ($M = 50.5$,

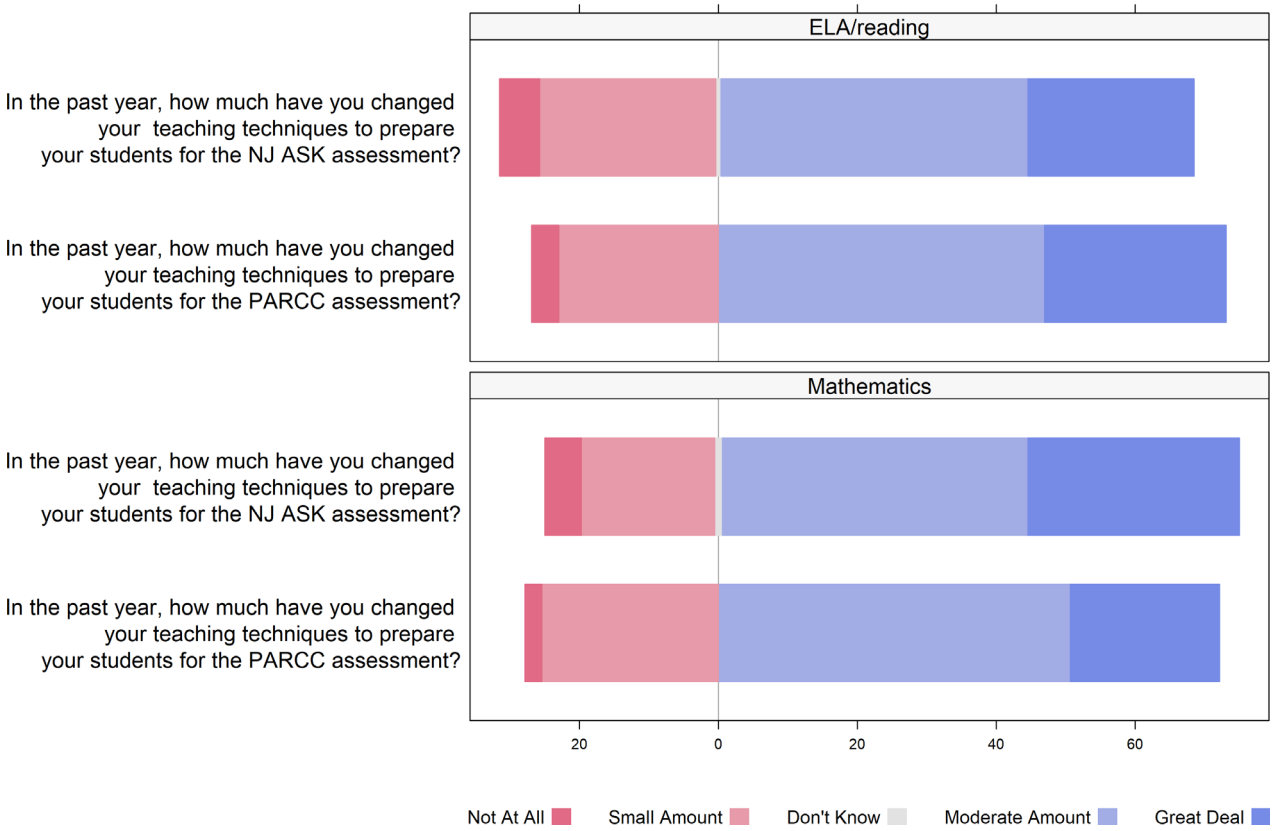

Figure 1 Teacher-reported changes in teaching practices related to preparing students for the state assessment (mathematics: $n = 266$; English language arts: $n = 276$).

SD = 32.7) was higher than the 2013–2014 school average statewide (reported as 32.7% on the state department of education website) based on a one-sample *z*-test ($\mu = 32.7$, $z = 11.0$, $p < .0001$).

We manually coded responses to two open-ended questions posed to teachers after the PARCC assessments were administered to summarize their perceptions on the negative and positive consequences of the PARCC assessments. All analyses and graphing were conducted in R (R Core Team, 2013).

Teachers' Views on Their Instructional Practices

We directly asked teachers about how much they changed their teaching techniques to prepare students for the NJ ASK to establish a baseline with which to compare responses from teachers after initial and subsequent administrations of PARCC. Most teachers reported changing teaching techniques to prepare students for the state assessment, both before and after administration of the PARCC assessments (see Figure 1; cf. Kane, Owens, Marinell, Thal, & Staiger, 2016, Table D6). Recall that New Jersey had partially integrated the CCSS into NJ ASK after the standards were adopted in 2010, so we expected that some teachers would have already made some changes.

In the remaining instructional practice items, we found little evidence of central tendency bias on the frequency of behavior items. That is, respondents tended to use the full set of response options on items. We also did not have evidence of acquiescence bias on the level of agreement items. That is, no respondents agreed with every item. In Figure 2, we display teachers' responses about their ELA instructional practices related to the prior state assessment (2014) and the PARCC assessment (2015). The responses are ordered by the most frequently occurring practice before the PARCC assessments were administered appearing at the top and the smallest percentage of daily responses at the bottom.

In general, we found that the teacher respondents engaged in the practices listed in Figure 2 more frequently after the PARCC assessments, that is, in 2015, relative to immediately before, in 2014. This finding is illustrated by larger percentages for the two darkest bars in 2015 versus 2014. While these trends are consistent with our expectations for most items,

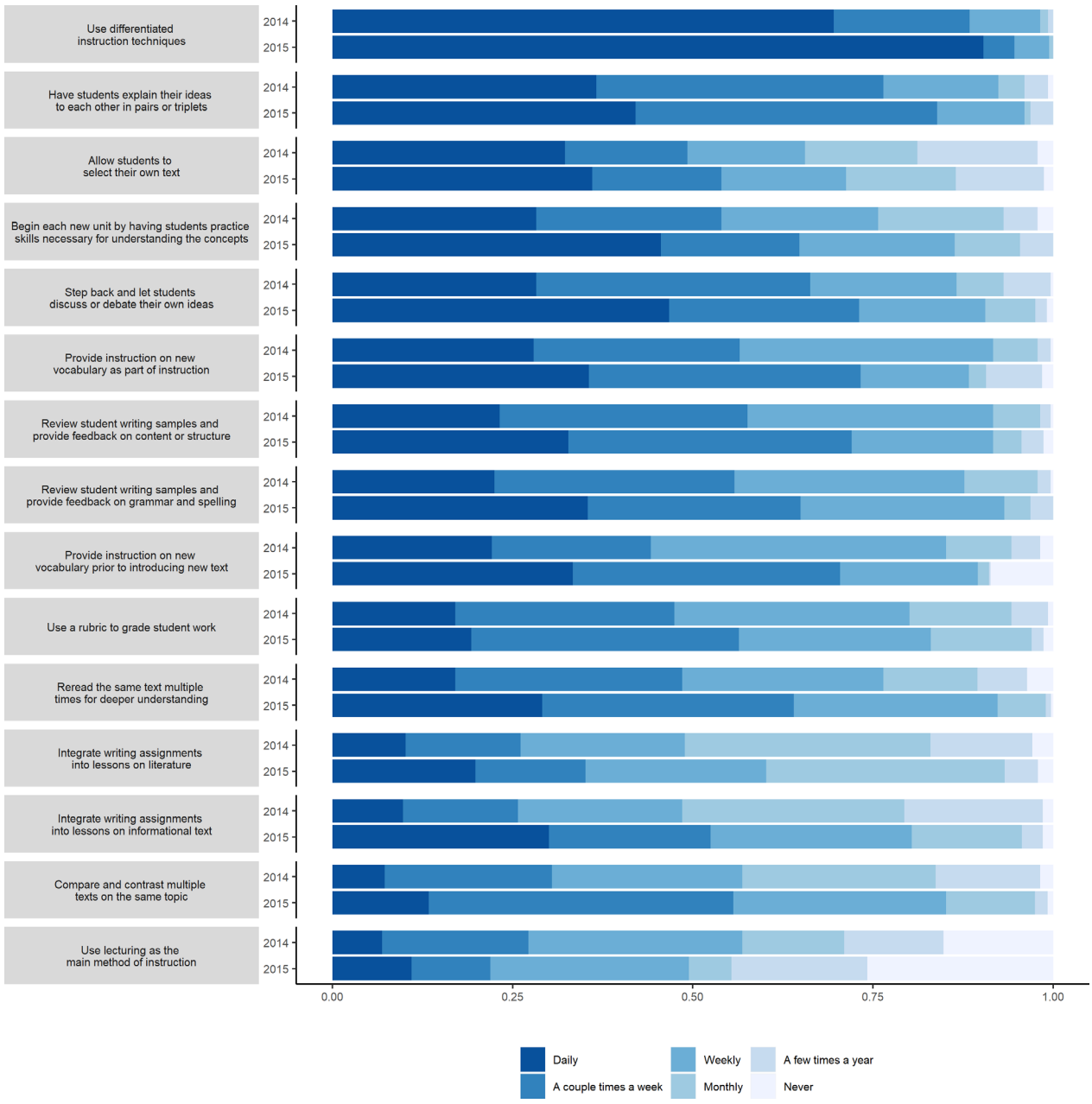

Figure 2 English language arts instructional practices survey responses, $n = 276$. The year 2014 was prior to the administration of Partnership for the Readiness of College and Careers assessments, and the year 2015 was after.

the changes are counter to our expectations for two practices. Specifically, using lecturing as the main method of instruction had more daily responses from teachers after the PARCC assessments were administered relative to before and the frequency of providing feedback on grammar and spelling increased. But since these are not linked to deep learning, we expected their frequency to decrease.

In comparison, we did not estimate increased frequencies in reported use of most of the mathematics instructional practices after the PARCC assessments were administered (Figure 3). Some of the decreases in frequencies were expected. Specifically, we expected teachers to engage less frequently in having students complete routine mathematics worksheets, emphasizing that following procedures, using lecturing, and teaching algorithms for solving problems are important, and we estimated trends in this direction. In terms of increasing trends, teachers more frequently had students explain their ideas to each other in pairs or triplets, used manipulatives, reviewed portfolios or journals to identify

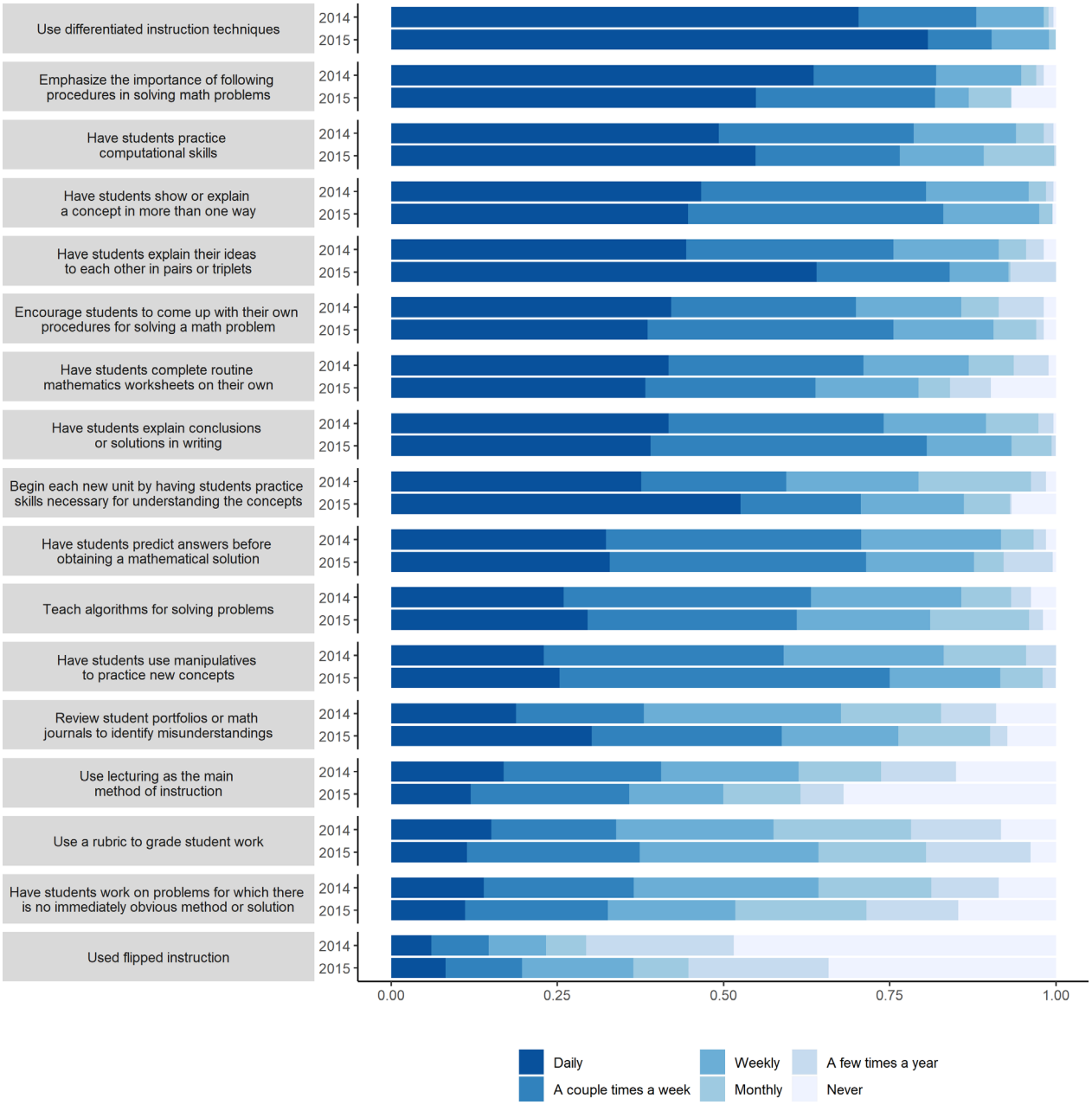

Figure 3 Mathematics instructional practices survey responses, $n = 266$. The year 2014 was prior to the administration of Partnership for the Readiness of College and Careers assessments, and the year 2015 was after.

misunderstandings, and practiced new skills with students before each unit. These increases were aligned with our expectations.

Figure 4 displays responses on the frequency of teachers' test preparation activities (cf. Kane et al., 2016, Figure 10 and Table D9). Teachers reported infrequent use of test items and test preparation materials at other times during the school year, both before and after PARCC was administered. Of particular note, fewer teachers reported engaging in most of the test preparation activities daily after the PARCC assessment relative to the year prior.

Teachers also responded to questions about their perspectives on and sentiments toward instructional practices related to assessment and the CCSS (see Figure 5). There were minimal differences on most of the items before and after PARCC was first administered, although there is some evidence of a trend toward more negative sentiments toward testing in the year after PARCC.

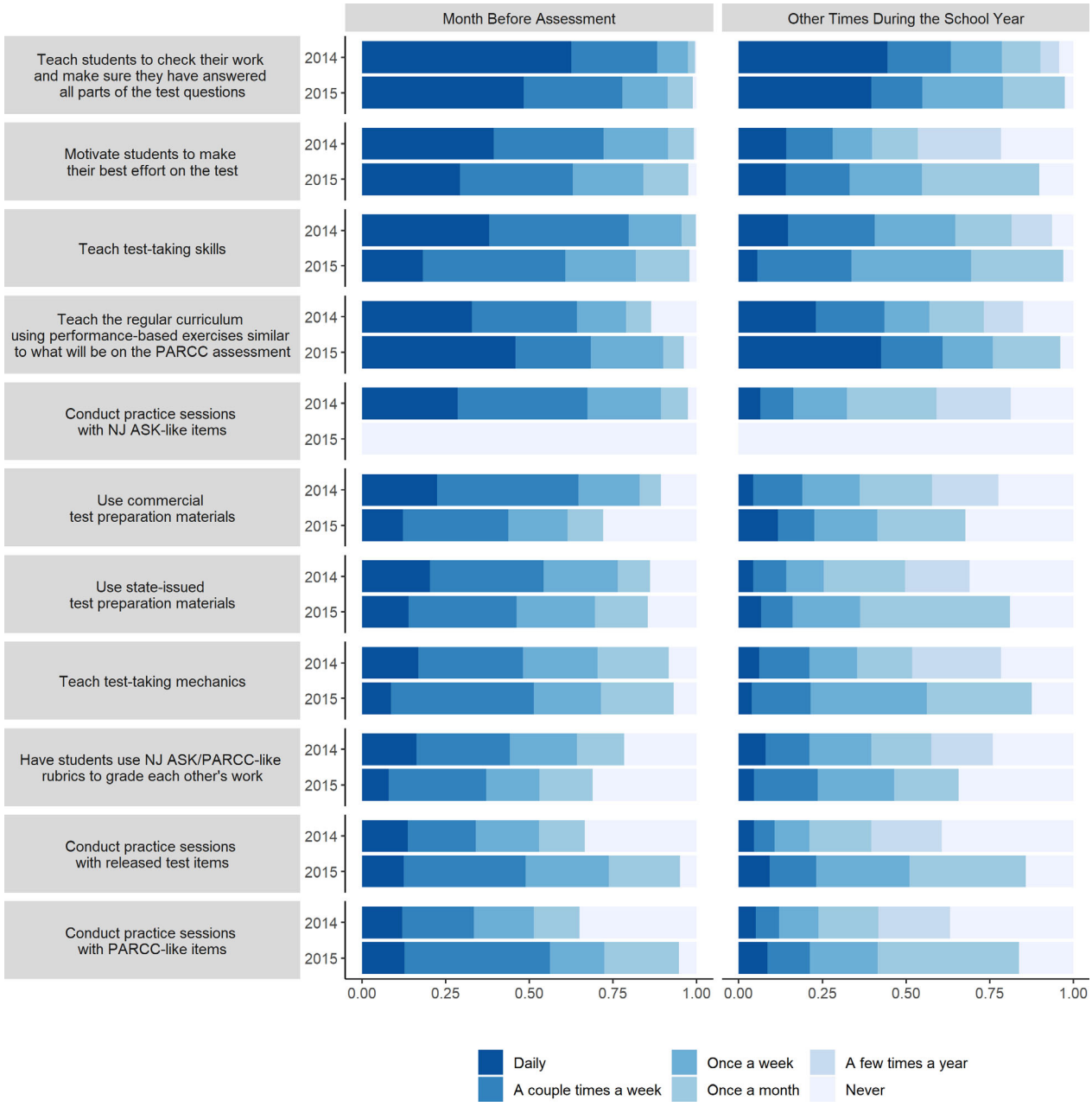

Figure 4 Teachers' frequency of test preparation activities, $n = 374$.

Teachers' Use of Test Scores

We first asked teachers whether they saw their students' test scores. Most respondents (95% of 469 teachers) reported seeing student test scores from the prior year's state assessment. Eighty-one percent of them saw the scores of their current students, and 73% saw the scores of their students from the prior academic year; 60% of them saw test scores for both their current and prior students. Viewing subscores was less common. Forty-two percent of 444 teacher respondents saw student subscores from the prior year's state assessment.

The following is focused on survey responses pertaining to teachers' use of state assessment test scores in the year prior to administration of the PARCC assessments. Because we could not obtain a sufficient sample within 1 year from when the PARCC assessments were first administered, we do not report on those responses. Figure 6 shows teachers' plans to do more or less of each activity in the current school year based on student test scores on the 2014 NJ ASK taken the previous

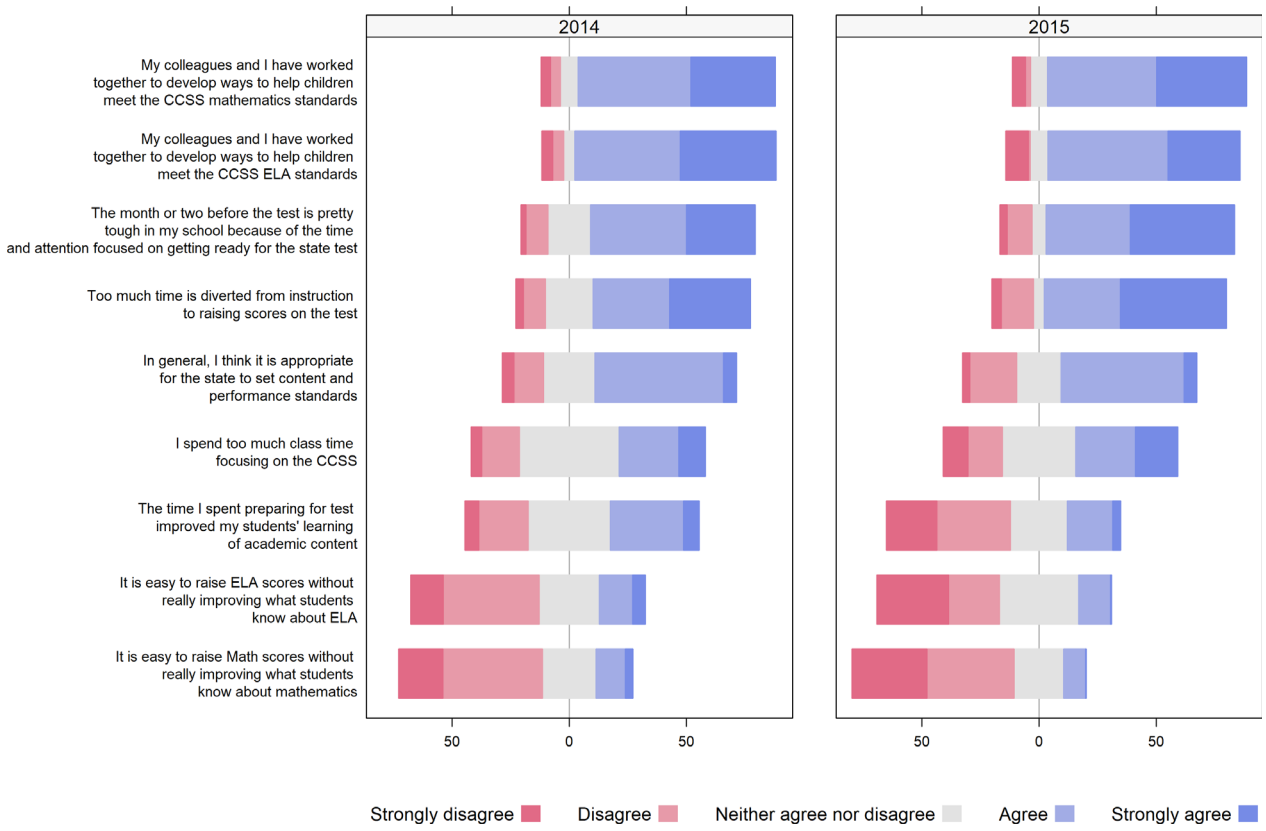

Figure 5 Teacher perspectives and sentiments regarding instructional practices and assessment ($n = 374$ for general questions; $n = 266$ for mathematics questions; $n = 276$ for English language arts questions).

spring. The majority were already reporting that they would be doing more based on test scores even before PARCC was administered.

In the year prior to administration of the PARCC assessments, many respondents felt that the test scores represented students' mathematics knowledge, but fewer felt similarly about ELA (see Figure 7). In contrast, few teachers indicated that they agreed or strongly agreed that subject-specific scores represented their effectiveness as teachers (27% of math teachers and 17% of ELA teachers; see Figure 8). However, around half of teachers agreed or strongly agreed that student scores helped them set goals to improve instructional practices (52% of math teachers and 48% of ELA teachers). Mathematics teachers who responded to the survey were mixed on whether or not state assessment mathematics test scores were representative of their instruction, test preparation, or effectiveness. In comparison, most ELA teachers did not agree that test scores reflected their instruction, test preparation, or effectiveness.

On the score use survey, we asked teachers about their knowledge and use of performance-level descriptors. Almost one fifth of our respondents in the fall of 2014 had yet to fully learn about performance-level descriptors. In comparison, almost 50% claimed to understand them well. Almost two thirds of the respondents said that they used performance-level descriptors to interpret students' proficiency levels on the prior year's state assessment.

On all of our surveys, we asked teachers about their knowledge of the CCSS. Although our baseline data were collected a year prior to implementation of the PARCC assessments, which were aligned to the standards, the CCSS were put in place for several years before the new assessments were administered. More than half of our respondents indicated that they understand the standards well enough to implement them fully in class, with an estimated increase of about 10 percentage points from before to after administration of the CCSS-aligned assessments. But as shown in Figure 9, about one third of respondents felt that they did not know the CCSS well enough to fully implement them in class, even several years after the CCSS were implemented (cf. Hamilton et al., 2016; Kaufman et al., 2016, Figure 12).

We also gathered open-ended responses to two questions asking teachers to describe any positive and negative consequences of the PARCC assessments. Sixty-nine percent of respondents gave responses for both positive and negative

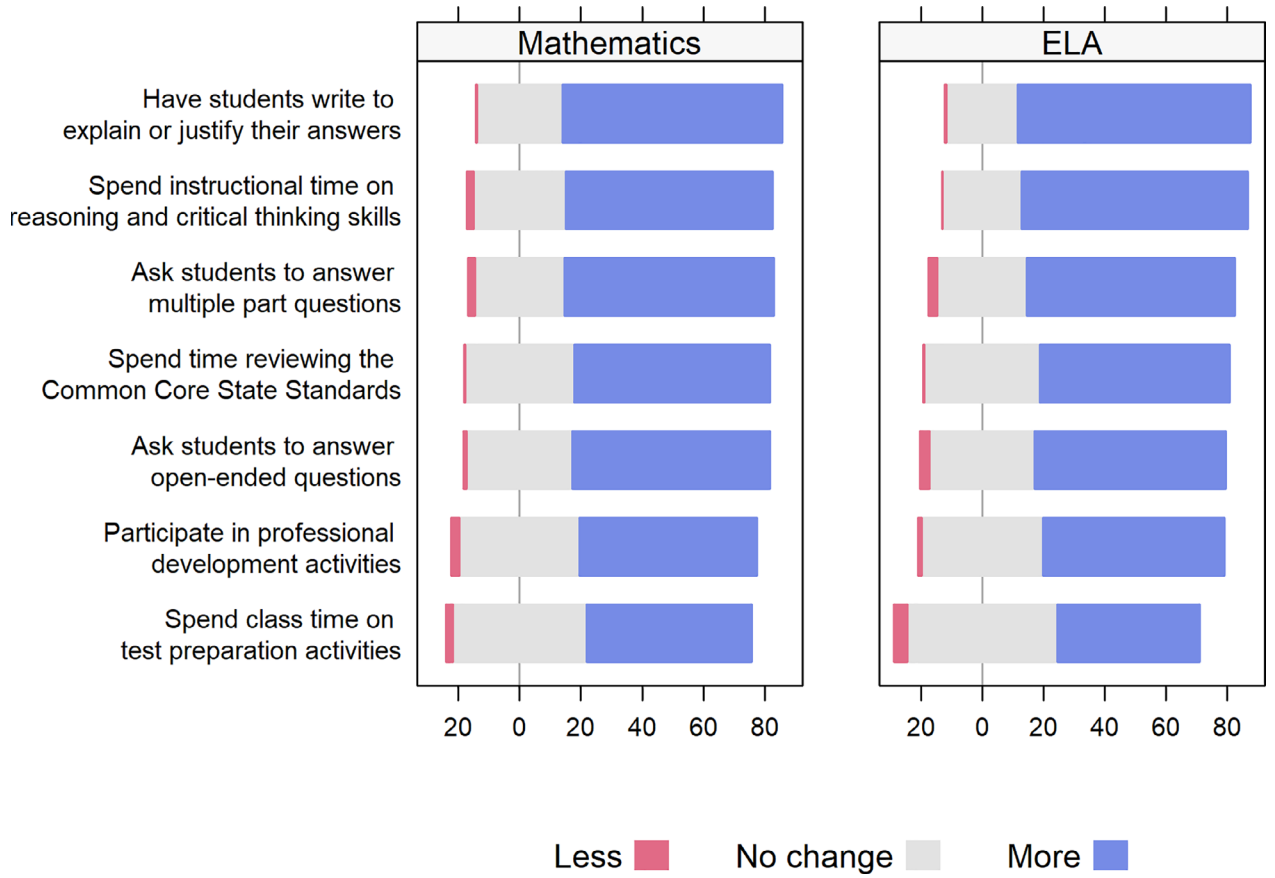

Figure 6 Teachers' plans for activities based on student test scores from the 2014 New Jersey Assessment of Skills and Knowledge assessment (mathematics: $n = 281$; English language arts: $n = 286$).

consequences, 22% responded to the negative consequences question only, 7% did not respond to either question, and the remaining 2% offered a positive consequence only. Categories of teachers' positive consequences of the PARCC assessments and the frequency of teachers whose responses we coded in the category are summarized in Table 2. Categories of negative consequences and the frequency of teachers' responses coded in each category are shown in Table 3.

Research Design and Measurement Challenges

This study is intended to provide policy makers with timely information on how the new CCSS-aligned assessments have an impact on teacher practices and test score use. Teachers were first surveyed a year prior to administration of the new assessments, immediately after the previous state assessment administration, about their views and perceptions about the instructional practices in which they had previously engaged during the school year. But states had already adopted the new standards, and some had been introducing them into the curriculum and existing assessments. We intended to collect trend data over two administrations of the new CCSS-aligned assessments to smooth out the variation we would expect due to a significant change within the educational system. Our three significant challenges to gathering data on potential changes to teachers' practices and use of test scores that are associated with a new assessment were type of measurement, timing, and sampling.

Type of Measurement

We chose to survey a large sample of teachers, asking them questions about their instructional practices, test preparation strategies, and test score use. This decision was motivated by time and cost constraints. Large-scale surveys are more efficient to administer than direct observation. Furthermore, we were interested in collecting information on teachers'

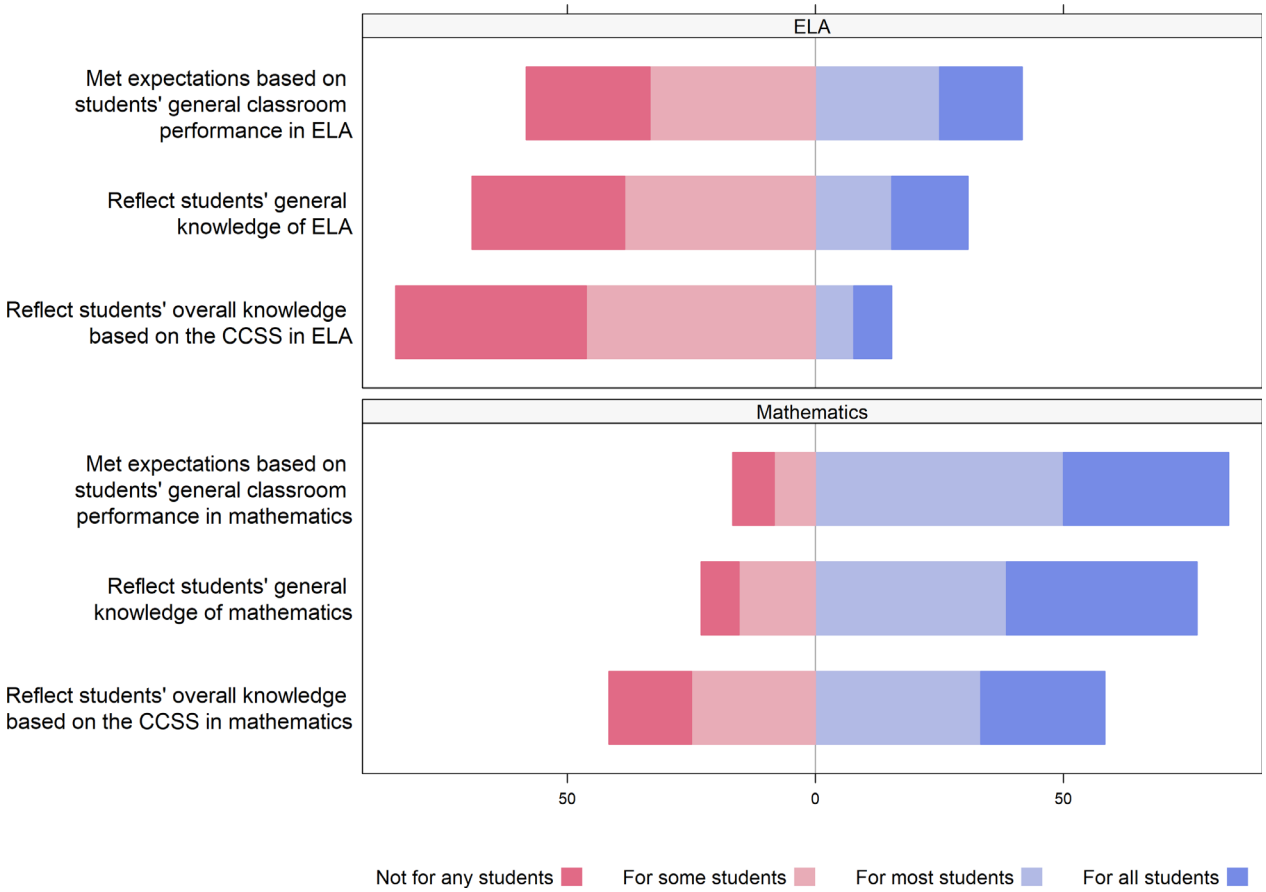

Figure 7 Teachers' sentiments regarding test scores on the 2014 New Jersey Assessment of Skills and Knowledge assessment the year prior to administration of the Partnership for the Readiness of College and Careers assessments (mathematics: $n = 215$; English language arts: $n = 226$).

practices related to assessment that we could compare to Firestone et al. (2004) who used a large-scale survey of teachers in the same state. But there are potential limitations to self-report retrospective questionnaires on teaching practices. For example, Mayer (1999) suggested that teacher surveys might be sufficient for describing the quantity of instructional practices, but not the quality (cf., Desimone, 2009). Furthermore, while the complete versions of the surveys used for this study were piloted with focus groups of teachers and administrators, and reviewed and revised by survey methodology and educational measurement experts, no independent reliability and validity evidence was available prior to administration.

Timing

Summative assessments are delivered late in the spring, near the summer. Our surveys were administered after the assessment window and ran through the summer. We heard from several principals who said they did not want to burden their teachers with a survey during the summer. We also received numerous out-of-office responses from the principals. This timing also introduced variability: Some teachers may have responded to the survey toward the end of the school year when stress levels are higher, while others responded when school was no longer in session.

Timing was also a challenge because administering the survey within a year of the new assessment being introduced meant that teachers already had knowledge of the new standards and likely began incorporating them in their instruction and test preparation while the prior state assessment was still in place. The length of time between when the summative assessments were administered and when the teachers see the score reports was also a challenge. During our time period of study, the state assessment was administered in the spring, and the score reports became available in the fall. In the first year that PARCC was introduced, score reports came out even later, with many teachers reporting that they still had

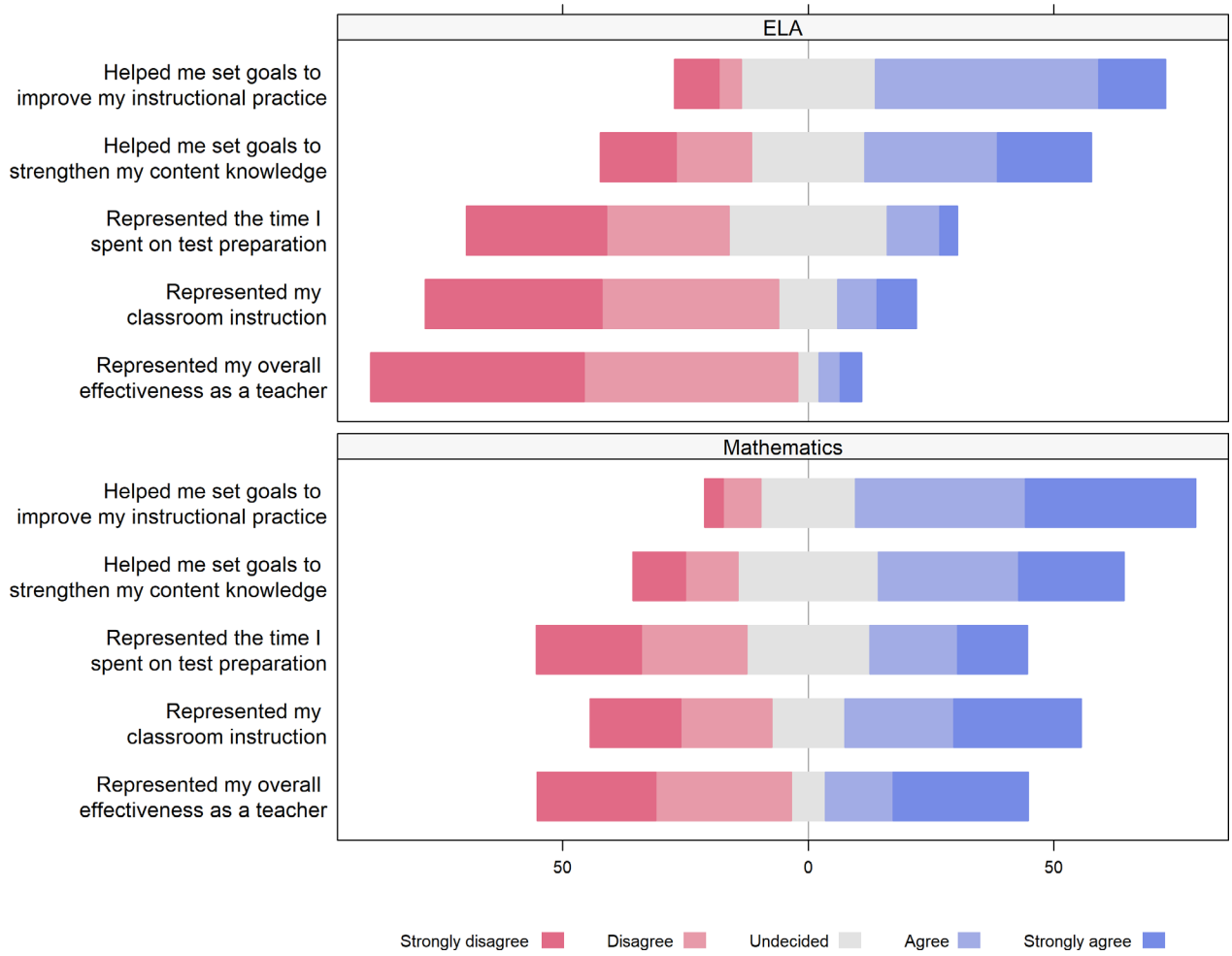

Figure 8 Teacher sentiments about test scores the year prior to administation of the Partnership for the Readiness of College and Careers assessments (mathematics: $n = 215$; English language arts: $n = 226$).

not seen scores by the following January. Timing challenges impeded our ability to collect a large sample of data about practices related to assessment and score use in real time.

Sampling

Only a small percentage of New Jersey schools were represented in our survey, so we cannot make claims about the data representing the state. For our sampling approach, we chose to target a large state sample. There are both benefits and limitations to this approach relative to collected data from a smaller targeted area, such as a district. The response rate is expected to be higher when a smaller area is targeted, but results may be less generalizable to other contexts. Larger samples can support more complex statistical modeling and can accommodate the inclusion of covariates of interest, such as grade level and school socioeconomic status. But it can be labor intensive and time consuming to get contact information for a large sample of teachers (see Matlock et al., 2016, who went to district websites to get teacher e-mail addresses). Given the need to collect information over time to estimate trends, a larger sample could be better used to accommodate teachers moving out of district but within state and to deal with missing data from teachers not responding each year. But we found a much lower response rate for the follow-up surveys intended to collect trend data. It is likely that our strategy to contact teachers indirectly through their principals contributed substantially to this low response rate. Splitting up the surveys to account for the timing of teachers' access to student test scores meant that we sent up to five surveys to principals over 2 years.

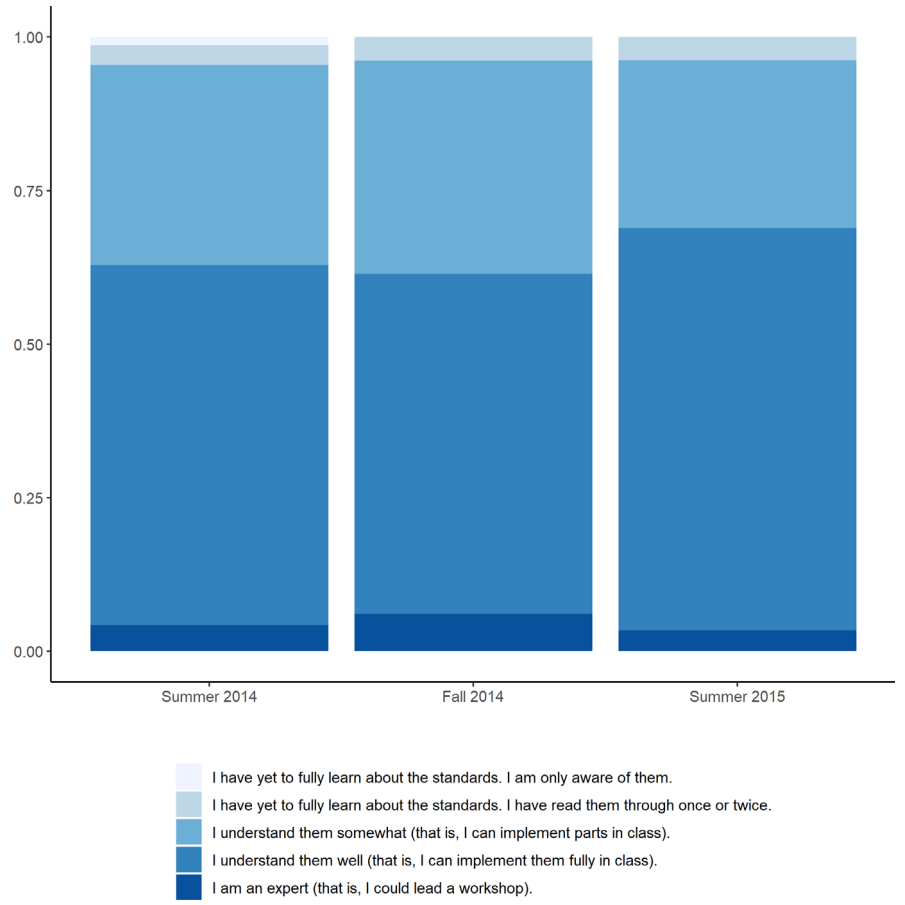

Figure 9 Teachers' knowledge of the Common Core State Standards ($n = 374$).

Table 2 Categories of Teachers' Open-Ended Responses About the Positive Consequences of the Partnership for the Readiness of College and Careers Assessments

Positive consequence of PARCC assessments	Frequency
Technology-related benefits	46
CCSS-related changes	44
Test readiness	33
Test taking strategies	19
Teacher and school changes	17
Total unique positive consequences	159

Note: CCSS = Common Core State Standards; PARCC = Partnership for the Readiness of College and Careers.

Innovation in Evaluating the Consequences of New Educational Assessments

Bennett (2010) stated that "in modern validity theory and its application, far more attention has been directed at the instrument technical-quality concerns more central to assessment programs than to [educational assessment] program effects" (p. 71). He went on to point out that

> evaluating these effects in a way that allows strong causal attributions would be extremely difficult, if not impossible. What should be achievable is to gather qualitative and quantitative data, examine whether those data are at least consistent with the postulated effects, and attempt to discount plausible competing hypotheses for seemingly positive results. Through an ongoing process of gathering evidence, support might either be amassed or the case for the theory of action weakened, suggesting the need for changes to the theory, to the assessment design, or to program implementation. (p. 87)

Table 3 Categories of Teachers' Open-Ended Responses About the Negative Consequences of the Partnership for the Readiness of College and Careers Assessments

Negative consequence	Frequency
Student dispositions	86
Too much time	62
PARCC difficulty	27
Teacher-level effects	24
Technical issues	19
Student preparedness	19
Special education, EL, and below grade level student issues	19
Test format	13
School-level effects	13
Parent-level effects	11
Opt out effects	6
General negative	1
Total unique negative consequences	300

Note: EL = English learners; PARCC = Partnership for the Readiness of College and Careers.

Lane and Stone (2002) and Haertel (2013), among others, also discussed the concept of evaluating educational assessment program effects in terms of consequences. The current study can contribute to the accumulation of evidence about the consequences of changing K – 12 large-scale assessments.

A lesson learned from the current study is that planning for an evaluation of the consequences of a new assessment should commence in the earliest stages of assessment development. If that were the case, we would have had more control over the research design to overcome some of the challenges described herein. That would have also given us an opportunity to develop a plan to triangulate the teachers' perspectives on the impact of the CCSS-aligned assessments with principals' or other school and district administrators' views of the changes. Given the level of negative public opinion and discourse on the CCSS-aligned assessments (e.g., Jochim & McGuinn, 2016), in hindsight, we would have collected data over smaller time increments, expanded the study to more than one state, and directly invited teachers to participate without going through their principals. This change would have improved our chances of higher response rates for the collection of trend data, increased our sample sizes each year, and allowed us to evaluate and possibly correct for nonresponse bias.

Discussion

In this study, we focused on a single state, collecting teachers' responses to survey questions related to instructional practice, test preparation strategies, and test score use, both before and after the introduction of one of the CCSS-aligned assessments, the PARCC assessment. The results provide a snapshot of responses from an unrepresentative group of several hundred teachers from one state. Due to the low response rates, the data and analysis results cannot be used to generalize claims about teachers in New Jersey or directly about the consequences of delivering assessments aligned with the CCSS. But the direction of the trend data was often consistent with what we expected based on the claims of the assessment program for teaching practices. In terms of the claim that the CCSS-aligned assessment would provide meaningful feedback to educators, we were not able to observe evidence in support of or refuting this claim due to the delays in the availability of the score reports and low response rates.

Aside from providing some evidence, albeit weak, of the effects of the CCSS-aligned assessments on a set of teachers' instructional practices and a description of pre-PARCC score use, this study also serves as a larger-scale pilot for using these items in other studies. We encourage the further accumulation of evidence on the consequences of large-scale state accountability assessments.

References

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice*, *42*, 18 – 29. https://doi.org/10.1207/s15430421tip4201_4

Achieve. (2013). *Closing the expectations gap*. Retrieved from http://www.achieve.org/files/2013ClosingtheExpectationsGapReport.pdf

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, *8*(2–3), 70–91.

Bennett, R. E., Kane, M., & Bridgeman, B. (2011, March). *Theory of action and validity argument in the context of through-course summative assessment*. Princeton, NJ: Center for K–12 Assessment & Performance Management, Educational Testing Service.

Conley, D. T. (2014). *The Common Core State Standards: Insights into their development and purpose*. Washington, DC: CCSSO. Retrieved from http://www.ccsso.org/Documents/2014/CCSS_Insight_Into_Development_2014.pdf

Council of Chief State School Officers. (2014). *Criteria for procuring and evaluating high-quality assessments*. Retrieved from http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*, 181–199. https://doi.org/10.3102/0013189X08331140

Edgerton, A. K., & Desimone, L. M. (2018). Teacher implementation of college- and career-readiness standards: Links among policy, instruction, challenges, and resources. *AERA Open*, *4*(4), 1–22. https://doi.org/10.1177/2332858418806863

Firestone, W. A., Schorr, R. Y., & Monfils, L. (2004). *The ambiguity of teaching to the test*. Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9781410609946

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives*, *11*(1–2), 1–18. https://doi.org/10.1080/15366367.2013.783752

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*, 25–46. https://doi.org/10.1093/pan/mpr025

Hainmueller, J. (2014). *ebal: Entropy reweighting to create balanced samples* (R package Version No. 0.1-6). Retrieved from https://CRAN.R-project.org/package=ebal

Hamilton, L. S., Kaufman, J. H., Stecher, B. M., Naftel, S., Robbins, M., Thompson, L. E., … Opfer, V. D. (2016). *What supports do teachers need to help students meet Common Core State Standards for mathematics? Findings from the American Teacher and American School Leader Panels*. Santa Monica, CA: RAND Corporation. https://doi.org/10.7249/RR1404

Herman, J. L., & Linn, R. L. (2013). *On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia* (CRESST Report No. 823). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Hoover, N. R., & Abrams, L. M. (2013). Teachers' instructional use of summative student assessment data. *Applied Measurement in Education*, *26*, 219–231. https://doi.org/10.1080/08957347.2013.793187

Jochim, A., & McGuinn, P. (2016). The politics of the Common Core assessments: Why states are quitting the PARCC and Smarter Balanced testing consortia. *Education Next*, *16*(4), 44–53.

Kane, M. T. (2006). Validation. Educational *Measurement*, *4*(2), 17–64.

Kane, T. J., Owens, A. M., Marinell, W. H., Thal, D. R., & Staiger, D. O. (2016). *Teaching higher: Educators' perspectives on Common Core implementation*. Center for Education Policy Research. Retrieved from http://cepr.harvard.edu/files/cepr/files/teaching-higher-report.pdf

Kaufman, J. H., Hamilton, L. S., Stecher, B. M., Naftel, S., Robbins, M., Garber, C., … Opfer, V. D. (2015). *What are teachers' and school leaders' major concerns about new K–12 state tests? Findings from the American Teacher and American School Leader Panels*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_reports/RR1200/RR1294/RAND_RR1294.pdf

Kaufman, J. H., Hamilton, L. S., Stecher, B. M., Naftel, S., Robbins, M., Thompson, L. E., … Opfer, V. D. (2016). *What supports do teachers need to help students meet Common Core State Standards for English language arts and literacy? Findings from the American Teacher and American School Leader Panels*. Santa Monica, CA: RAND Corporation. https://doi.org/10.7249/RR1374

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice*, *21*(1), 23–30. https://doi.org/10.1111/j.1745-3992.2002.tb00082.x

Matlock, K. L., Goering, C. Z., Endacott, J., Collet, V. S., Denny, G. S., Jennings-Davis, J., & Wright, G. P. (2016). Teachers' views of the Common Core State Standards and its implementation. *Educational Review*, *68*, 291–305. https://doi.org/10.1080/00131911.2015.1070333

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, *21*, 29–45. https://doi.org/10.3102/01623737021001029

McClellan, C., Joe, J., & Bassett, K. (2015). *The right trajectory: State teachers of the year compare former and new state assessments*. Retrieved from http://www.nnstoy.org/wp-content/uploads/2015/11/Right-Trajectory-FINAL.pdf

New Jersey Department of Education. (n.d.). *Historical context: Overview of New Jersey's statewide testing program*. Retrieved from http://www.state.nj.us/education/assessment/history.shtml

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core standards: The new US intended curriculum. *Educational Researcher*, *40*, 103–116. https://doi.org/10.3102/0013189X11405038

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Vogler, K. E. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. *Education*, *123*, 39.

Webb, N. L., Alt, M., Ely, R., Cormier, M., & Vesperman, B. (2005). *The WEB alignment tool: Development, refinement, and dissemination*. Washington, DC: Council of Chief State School Officers.

Suggested citation:

Buzick, H. M., Rhoad-Drogalis, A., Laitusis, C. C., & King, T. C. (2019). *Teachers' views of their practices related to Common Core State Standards-aligned assessments* (Research Report No. RR-19-40). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12277

Action Editor: James Carlson

Reviewers: Richard Tannenbaum and Caroline Wylie