

California Department of Education

Measures for a College and Career Indicator: Research Brief on Multiple Measures

August 5, 2014

Educational Policy Improvement Center (EPIC)

Introduction

In September of 2012, Governor Jerry Brown signed into law Senate Bill (SB) 1458, which calls for California's school accountability system to shift from a near-exclusive reliance on state test scores to a broader range of measures demonstrating student achievement. At the high school level, starting in the 2015–2016 school year, the Academic Performance Index (API) will include an indicator composed of measures reflecting students' college and career preparedness.

To determine the measures that will comprise this new indicator, the State Superintendent of Public Instruction and the State Board of Education will consider input from regional public meetings, a statewide survey, and recommendations from the Public Schools Accountability Act (PSAA) Advisory Committee. To further support this decision-making process, the California Department of Education (CDE) has contracted with the Educational Policy Improvement Center (EPIC) to conduct analyses of six different types of potential measures of college and career preparedness, each type analyzed in a series of white papers and a summary report.

This white paper considers the benefits and limitations of multiple measures of college and career preparedness. The previous papers in this series evaluated measures against an analytical framework consisting of 10 criteria. This white paper uses theory and practice related to measures of college and career preparedness as its frame of reference and primary organizing structure. The paper also considers the historical context for how accountability measures have been used in American education. The paper explores a set of criteria to consider when designing a multiple-measures system, then examines current multiple-measures systems used across the nation, and summarizes emerging themes from the literature on multiple measures for accountability of college and career preparedness. The paper concludes with a consideration of cutting-edge concepts and uses of multiple measures to ascertain college and career preparedness.

Bases for Incorporating Multiple Measures

Perhaps the most compelling argument that can be made for including multiple measures is based on the limitations of single measures. Most often, these are standardized tests. The Sandler Foundation (2014), a California-based philanthropic organization, advocates using multiple measures because “standardized assessments in a handful of core subjects aren't enough to adequately measure school performance.” Over the past decade, the demands of No Child Left Behind (NCLB), and federal education assessment policy more generally, have led to the use of scores on English and mathematics tests exclusively, which overlooks the information that could be generated on student performance in other subject areas and has the effect of narrowing the curriculum (Manna, 2011).

Additionally, prioritizing a handful of measures has tended to overemphasize cut scores that define levels of proficiency. Cut scores can inadvertently lower the bar by

transforming an intended floor into a ceiling that sets the target on which teachers and students focus. Meanwhile, students whose test scores are well below a cut score are often overlooked in favor of those who are very close to the cut score, because moving the latter students over the line makes schools look better. Problems arise particularly when such single measures have questionable validity (Marion & Gong, 2003) and can send the wrong signal to students about their academic potential and capabilities. This is especially salient when predicting individual student college and career preparedness. Scores on single measures might send the message to students that they are not prepared when, in fact, they may still be capable of success. This issue is particularly problematic when the measures' thresholds have been arbitrarily set (Francis et al., 2005), potentially discouraging students who should be seeking postsecondary opportunities. When schools are judged on a single measure, this may result in a school's ability to prepare students for a variety of postsecondary pathways being disregarded. Research has found that students can qualify for a variety of postsecondary pathways even if they do not reach English and mathematics cut scores that the system has designated (WestEd & EPIC, 2013).

Furthermore, systems that include multiple measures are more likely to achieve accuracy, consistency, and reliability (Brookhart, 2009; Gong & Hill, 2001; Marion & Gong, 2003) and increase a state's capacity to identify unintended consequences by analyzing student achievement and school outcomes from a variety of perspectives (Marion & Gong, 2003; Mikulecky & Christie, 2014; Raudenbush, 2004). Additional advantages of multiple measures include a system that can encourage and support effective teaching of critical content (Gong & Hill, 2001) and help to attenuate each individual measure's validity limitations (ASCD, 2013). These advantages are particularly important in high-stakes accountability systems. Raudenbush (2004) warns against making high-stakes decisions based on comparisons of school-level test score averages. A multiple-measures system provides greater insight into the complex interplay of school-level variables that influence and shape student achievement and lead to college and career preparedness. Such insights provide the basis for fairer, more valid accountability determinations regarding any individual school.

How Measures Have Been Used for Accountability

If multiple measures are so good, why aren't they the norm in state accountability systems today? Much of the reason can be traced to the ability to measure in the first place and the tendency of states to devote as few resources as possible to gauging the performance of the educational system. Mikulecky and Christie (2014) traced the progression of accountability from the turn of the 20th century to the present. From 1900 to the 1980s, accountability systems served accreditation purposes and measured school quality through input factors and indicators that could be easily counted, such as percentages of faculty with degrees, existence of curriculum plans, and numbers of library books. The 1990s saw the widespread implementation of academic standards that could be tested systematically in ways that allowed comparisons of schools on a common scale. These accountability systems often relied on school report cards to give each school an A–F grade. Federal mandates dominated the first decade of the 21st

century in the form of NCLB, which attached federal-level incentives and consequences to school-level scores on standards in mathematics and reading. Consequently, states relied exclusively on standardized tests of content knowledge to evaluate learner progress and school quality. However, these measures were not linked directly to college and career preparedness (Conley & Darling-Hammond, 2013).

Shortly before NCLB, California's standards rated among the best in the nation (American Federation of Teachers, 2008; Klein et al., 2005), particularly in mathematics (Finn, Petrilli, & Vanourek, 1998; Stotsky, 2005). Its Academic Performance Index used scores on the California Standards Tests in mathematics, English, science, and social studies, as well as the California High School Exit Examination, to compare school performance and rank schools. Ironically, NCLB's focus on singular measures actually created twin accountability systems in California and elsewhere, with districts and schools receiving ratings on both state and federal systems (EdTrust, 2014).

The federal accountability model under NCLB expected significant annual growth for all students, with particular attention on the performance of subgroups (Raudenbush, 2004). The use of Adequate Yearly Progress (AYP) as the primary measure of school effectiveness increased the attention schools gave to subgroups that were performing less well, but also moved states away from any system that measured college and career preparedness through multiple measures. Instead, states became fixated on achievement "gap gazing" (Gutiérrez, 2008). While notions such as AYP are powerful and potentially quite valuable and useful, Sandler (2014) prefers that AYP be used as a floor accompanied by rigorous, peer-reviewed measures that create incentives for schools to address college and career preparedness for all students.

Accountability in the past half-dozen years has evolved as the U.S. Department of Education granted states waivers from the strictest and most inflexible requirements of NCLB. Of the 45 states that submitted waiver requests, 43 have been approved. Race to the Top legislation passed in 2009 used a competitive grant application process as a mechanism to encourage states to link test scores to teacher evaluations (Mikulecky & Christie, 2014). A number of states have used the waiver process to explore multiple measures and to conceptualize college and career preparedness more comprehensively. This evolution of policy has led to greater openness toward multiple measures in state accountability systems.

Criteria for Designing a Multiple-Measures System

The natural tendency when considering multiple measures is to focus on the measures themselves and to consider them individually and then in concert. However, multiple measures constitute what amounts to a system rather than just a combination of individual elements. Assembling multiple measures, then, requires attention both to the technical aspects of the measures and to the process used to select them. This section offers an overview of the criteria involved in designing a multiple-measures system, in other words, the process used to create and build support for a multiple-measures

system. This discussion is followed by examples of multiple-measures systems currently in use in a number of states.

When constructing a multiple-measures accountability system, states should begin by carefully specifying how they will use multiple measures. Will the accountability system seek to measure multiple constructs, employ multiple measures to assess a single construct, or gather data from multiple administrations of the same measure of a single construct (Brookhart, 2009; Ehren & Swanborn, 2012; Gong & Hill, 2001)? All three approaches have appeared in the literature as definitions or examples of multiple measures. The majority of this paper focuses on single measures of multiple constructs (Linn, 2005), which requires a state to define clearly all the constructs it seeks to measure, the scope of its multiple-measures system, and associated trade-offs of the system's breadth and affordability (Sandler, 2014).

A state considering multiple measures must ensure the design meets its goals and requirements. Mikulecky and Christie (2014) convened a panel of 12 school accountability experts who identified this process as an opportunity for a state to define and publicize the "North star" of its education policy. Doing so sends a message that can focus educator efforts and improve public understanding of state education goals. Perie, Park, and Klau (2007) urged system designers to consider both goals and theories of action to increase the likelihood that a multiple-measures system provides the intended data and achieves desired outcomes.

Mikulecky and Christie established conditions that are necessary but not sufficient to create a multiple-measures system. Principally, they endorsed alignment from kindergarten to college as a necessary prerequisite to a state's capacity to collect data at all levels and conduct longitudinal analyses at all levels. Next, states must provide all students with opportunities to learn, including programs such as Advanced Placement, International Baccalaureate, and dual enrollment. Finally, states must choose measures wisely from among a wide range of possibilities that include, for example, course-taking and test-taking patterns, course-passing and test-passing results, industry certifications, postsecondary enrollment, and percentages of high school graduates needing remediation prior to earning college credit.

A review of the literature produced seven criteria that states should consider when designing a multiple-measures system:

1. Stakeholder collaboration
2. Design method
3. Breadth of coverage
4. Measurement/reporting type
5. Combination method
6. Ability to compare
7. Stakes

The order does not represent a step-by-step process. In developing each criterion, its descriptors, and their corresponding levels of complexity for this white paper, EPIC

reviewed a wide range of high-quality studies on the design of accountability systems, particularly the benefits and limitations associated with multiple measures.

Table 1. Criteria for Designing Multiple-Measures State Accountability System

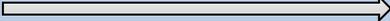
Criterion	Most Complex  Least Complex				
	Stakeholder collaboration <i>Whose voices will be included or excluded?</i>	Consensus	N/A	Hybrid	N/A
Design method <i>To what extent will the design be unique or dependent on models used previously by other states?</i>	Independent	Piggyback	Cyclical	Patchwork	Status quo
Breadth of coverage <i>What are the trade-offs between comprehensive coverage and capacity to collect and analyze data?</i>	Comprehensive	Unified	N/A	Overlapping	Isolated
Measurement/ reporting type <i>What are the trade-offs between complexity and transparency?</i>	(Quasi)- longitudinal	Successive groups	Status	Mandatory reporting	Optional reporting
Combination method <i>Which should be more highly valued: simplicity or accuracy?</i>	Matrix	Compensatory (including weighting)	N/A	Complementary	Conjunctive

Table 1 shows descriptors for each of the first five criteria, which are explained in greater detail in the sections that follow. Each criterion’s descriptors are listed on a scale from most to least complex. Greater complexity typically involves greater trade-offs. Consequently, the most complex approach to each criterion may not be appropriate to satisfy the state’s goal.

Criteria 6 and 7 are not included in Table 1 because they do not lend themselves to a tabular display. Ability to compare (Criterion 6) encompasses a variety of comparative options that states should consider when designing a multiple-measures system. Therefore, Criterion 6 is discussed in detail in an ensuing section. Criterion 7 considers the stakes of the system, which is a continuous, not a dichotomous (on/off), variable. Considering all the possible incentives and consequences is a crucial design conversation with many dimensions. It is also addressed in an ensuing section.

Stakeholder Collaboration

Mikulecky and Christie (2014) emphasize the importance of a coalition to foster public buy-in and cohesive messaging. Generating such consensus depends upon the level of stakeholder collaboration the state is willing to facilitate during the design stage. Blank (1993) states that selecting education indicators is a process that “requires interaction and consensus among different kinds of stakeholders.” Blank specifically endorsed engaging a group wider than policymakers and educators, seeking researchers to ensure inclusion of “central and critical” variables, and data managers to guarantee data aggregation and reporting that is appropriate for, and feasible within, state infrastructure. However, such consensus would require considerable resources, especially time, to convene diverse groups of stakeholders that could inform the design of a cohesive state system.¹

Hentschke, Wohlstetter, Hirman, and Zeehandelaar (2011) investigated multiple measures of school performance by California charter schools. They reported anecdotal evidence that suggest benefits of a hybrid system incorporating consensus and top-down approaches. Interviews and focus groups of charter-school stakeholders informed Hentschke et al.’s findings: some stakeholders sought additional understanding of data sources, reporting procedures, and how “data would help them make the managerial decisions necessary to foster school-level change.” Meanwhile, other participants who had greater familiarity with data systems expressed interest in generating their own indices and creating unique accountability reports to best serve their schools. The flexibility required in a hybrid system could increase the burden on a state to identify classroom and school leader capacities at local levels and then act upon them. Also, a hybrid system could generate unintended consequences of inadequate coverage and a lack of fair comparisons if schools adopt unreliable or insufficient measures.

Linn (2005) reported limited advantages of a top-down approach, such as the possibility of visible results in less time. Sandler (2014) noted that a uniform set of measures may facilitate comparisons and rankings, but could stifle innovation and local decision-making by disempowering education leaders.

¹ It is worth noting parenthetically that the PSAA Advisory Committee already serves much of this function in California and that it could be used as the hub for gathering and synthesizing additional input on

Design Method

In a foundational study on creating indicator systems, Shavelson et al. (1987) explored benefits and limitations of five methods. The independent method requires development and field-testing of a unique, comprehensive system. The piggyback method borrows from existing data systems, adding iteratively to obtain more comprehensive data. The cyclical method employs as-needed, time-series data collection from smaller samples, complementing existing statewide data by exploring components of schooling that a current system does not address. The patchwork method cobbles together previously unexplained components through existing data sources such as the National Assessment of Educational Progress (NAEP). As its name implies, the status quo method relies on data currently available with no attempts to augment its scope.

Shavelson et al. considered the independent and piggyback methods as being best suited to identify emerging problems in a system, inform improvement of policy and practice, and monitor areas currently being ignored. However, these methods require high levels of technical expertise, political support, and acceptance of burdens to state and local entities. The independent option offers the greatest stability over time, but may require redistribution of authorities across and within agencies. The piggyback option also depends largely upon interagency cooperation. Though the cyclical method can inform improvements and monitor a system’s blind spots, it cannot detect emerging problems. The cyclical method can only react to problems identified by other means or suspected to be in existence. The status quo and patchwork methods present fewer feasibility concerns, particularly in technical expertise, respondent burden, and interagency cooperation. However, those methods provide the fewest benefits. Table 2 reports design costs that Shavelson et al. estimated for the National Science Foundation, but the ratios could be applied to estimate impacts on a state’s budget.

Table 2. Indicator Design Cost Estimates by Method

Method	1987 Dollars	2014 Dollars ²	Maximum Ratio to Status Quo
Status quo	\$40,000	\$84,362	1:1
Patchwork	\$150,000	\$316,356	3.75:1
Cyclical	\$500,000 – 1 million	\$1.05 – 2.11 million	25:1
Piggyback	\$2.5 – 6 million	\$5.27 – 12.65 million	150:1
Independent	\$23 – 34 million	\$48.51 – 71.71 million	850:1

Source: Shavelson et al. (1987)

² 2014 dollars are estimated using a 2.80% annual inflation average and rounded.

Breadth of Coverage

A state must choose how comprehensive or isolated a system it requires. Rather than incorporating all potential options, a state will need to make strategic choices on a manageable set of measures. In practice, this implies a handful of measures with careful consideration regarding what each contributes to an overall determination of college and career preparedness.

A key trade-off is between breadth (comprehensive) and focus (isolated). Shavelson et al. (1987) recommended prioritizing measures that are “comprehensive and parsimonious.” Sandler (2014) echoed that approach, noting that measuring multiple indicators will undoubtedly cost more. Blank (1993) strongly advocated for fewer indicators in favor of minimizing complex reporting and directing resources to the most critical areas.

However, such isolation leaves a system susceptible to blind spots that could be alleviated by additional indicators. A unified model could allow a system to detect blind spots if the state strategically chose measures to examine a wide range of indicators without redundancy.

The overlapping model, which is characterized by iterative additions over time to an existing system, could increase reliability through repeated measurements of related constructs (Gong & Hill, 2001). But an overlapping model can end up mixing “traditional compliance-based forms of accountability” with newer performance-based models (Jos & Tompkins, 2004). The result could be “the accountability paradox,” in which the system yields “good administrative judgment” but simultaneously threaten “the very qualities that support responsible judgment” (Jos & Tompkins, 2004).

Measurement/Reporting Type

Gong and Hill (2001) classify mandatory reporting into three models: (a) status, (b) successive groups, and (c) quasi-longitudinal. The status model measures a construct at a specific, single time point. The successive groups model compares pre- and post-scores of a construct, examining both within-group and between-group differences. The quasi-longitudinal model compares growth in a construct from one year with that of other years. Balancing complexity and transparency is an essential consideration in this criterion.

The status model is the most straightforward. It provides a cross-section of a school's or state's current level with respect to a given indicator. It will detect any measurable difference that is statistically significant (Gong & Hill, 2001). It targets growth only in comparison to a previous year, typically examining the group average of Year 1 to that of Year 2. However, such findings may lead to questionable causal inferences. Large group sizes can be called statistically significant largely based on the size of the sample, not necessarily the meaningful nature of the difference. Furthermore, groups examined, for example 10th grade students, may not present the necessary stability across years to make informed comparisons to another year's cohort. Finally,

comparing a group average relays little information about how to change course if sufficient progress has not been made.

The successive groups model may better align to a long-range goal such as increasing the percentage of eighth grade students taking algebra. The successive group model tracks adequate annual progress toward long-range goals. However, this method requires more political will and public patience. Furthermore, diffuse measures such as dual-enrollment participation or completion and industry certifications may depend upon factors outside a school's control.

The quasi-longitudinal model relies on more complex statistical analysis but allows comparisons of real and expected growth in the aggregate and for subgroups. Raudenbush (2004) endorses this approach, particularly using value-added estimates that statistically adjust "for school differences in the entry status of the students the school serves." Such methods can be highly reliable when averaged over spans of two or more years.

Two other possibilities include mandatory reporting and optional reporting. Nineteen states employ mandatory reporting without accompanying measurement of college and career preparedness indicators. Illinois, Kentucky, and Maryland report at least five indicators that they do not measure. Optional reporting exists in concept, though no evidence of states using such an approach surfaced in reviews of Elementary and Secondary Education Act waivers or associated documents.

Combination Method

Once a state has chosen the measures for its indicator, four basic options exist for combining measures: matrix, compensatory, conjunctive, and complementary. Choice of method creates trade-offs associated with the validity for conclusions that can be drawn from a multiple-measures system (Brookhart, 2009; Gong & Hill, 2001).

The matrix method may best serve situations that require simplicity, fairness, and cost as essential criteria (Harris, 2013). The matrix does not actually involve literal combination. Instead, matrices are used to set levels for each of the multiple measures and then examine each in relation to its individual threshold.

For example, Table 3 presents a hypothetical approach for measuring three college and career preparedness indicators and establishing achievement levels for each. For simplicity's sake, the matrix is organized into high, medium, and low levels for each measure. Combinations with blue text indicate higher performance, green text indicates mid-range performance, and red text indicates lower performance.

Table 3. A Matrix Approach to Combining Multiple Measures

High ACT, High CR, High WK	High ACT, High CR, Mid WK	High ACT, Mid CR, High WK
Mid ACT, High CR, High WK	High ACT, Mid CR, Mid WK	Mid ACT, Mid CR, High WK
Mid ACT, High CR, Mid WK	High ACT, High CR, Low WK	High ACT, Low CR, High WK
Low ACT, High CR, High WK	Mid ACT, Mid CR, Mid WK	High ACT, Mid CR, Low WK
High ACT, Low CR, Mid WK	Mid ACT, High CR, Low WK	Mid ACT, Low CR, High WK
Low ACT, High CR, Mid WK	Low ACT, Mid CR, High WK	Mid ACT, Mid CR, Low WK
Mid ACT, Low CR, Mid WK	Low ACT, Mid CR, Mid WK	High ACT, Low CR, Low WK
Low ACT, High CR, Low WK	Low ACT, Low CR, High WK	Mid ACT, Low CR, Low WK
Low ACT, Mid CR, Low WK	Low ACT, Low CR, Mid WK	Low ACT, Low CR, Low WK

ACT = ACT results; CR = percentage of students in need of college remediation; WK = WorkKeys results
Blue = higher performance; Green = mid-range performance; Red = lower performance

If a state chose ACT results as a college and career preparedness indicator, a defensible choice for a mid-range composite score might be 20, because that was the national average during exam administrations from 2011 to 2013 (ACT, 2014). A mid-range score of 21 might be defensible because several states use that threshold in their accountability systems because it corresponds to the college readiness benchmark on the ACT test. If a state chose percentage of students needing college remediation, a defensible medium level might fall between 60%³ and 68%.⁴ And if a state chose the percentage of students earning WorkKeys' National Career Readiness Certificates at or above the silver level, a defensible mid-range level might be 69%, based on data collected from more than 3.8 million WorkKeys' examinees from 2006–2011 (LeFebvre, Clark, Burkum, & Kyte, 2013).

The matrix method allows for more nuances than a single number conveys, and it targets opportunities for improvement. Also, it allows states to approach schools that are inconsistent across measures differently than it would if the state only had a single number on which to base decisions. The complexity of this approach is evident, so it would take time to socialize the profession and the public to any such model. An additional limitation is the matrix method's assumption that all measures are valued equally.

Weighting multiple measures addresses this issue by attaching multipliers to the value of each measure. Doing so generates a single number, which can then be used to rank groups. For example, if a state chose the same three indicators as the matrix example, ACT results, college remediation, and industry certifications would each receive a .33 or 33% multiplier producing an approximate total of 1 or 100%. That would mean the state placed equal value on each indicator. Also called a composite or index, weighting allows a state to reflect its priorities by increasing the value of some measures over others.

Harris (2013) noted several benefits of weighted indices, such as their ability to produce rankings from simple calculations and their intuitive role in society, most notably the

³ National statistic cited by The National Center for Public Policy and Higher Education (2010).

⁴ Greene & Foster (2003).

Dow Jones Industrial Average, Consumer Reports ratings, and the Weather Channel's heat index. However, Harris cautions that weighting is prone to random error. This is why weighted indexes vary so much from year to year.

Brookhart (2009) delineates conjunctive, compensatory, and complementary combination methods. Systems using the conjunctive method require schools to meet or exceed thresholds on all measures. Systems using the compensatory method allow higher performance on some measures to offset lower performances on others. In systems using the complementary method, meeting or exceeding the threshold on a single measure indicates sufficiency, even if the thresholds of some or all other measures are not met.

NCLB provided an example of a conjunctive approach by requiring AYP on several measures (and for all subgroups). However, NCLB's safe-harbor provision operated in complementary fashion, granting adequate progress to a school if a subgroup did not pass but the percentage of students below the threshold in that subgroup decreased by 10% or more. Weighting is an example of a compensatory model such as the grading policies in most teachers' classrooms (Brookhart, 2009).

Conley (2012) asserts the essential nature of this decision in system design, a decision that is not merely technical. If a state believes that students need to do all things equally well in order to be considered prepared for college and career, a conjunctive approach is warranted. If a state believes a student can offset weaknesses in some areas with strengths in others, within a variable range of skills, college and career preparedness should take on a compensatory approach. A design that fails to choose appropriately in this area will produce classification errors that can either identify students as being prepared for college and careers when they are not or not being prepared when they are.

Ability to Compare

As mentioned previously, the ability to compare does not fit a hierarchy. It operates more like a checklist. Therefore, system designers should ask: What comparisons are included or excluded by particular design decisions? Affordability may drive responses to this question (Sandler, 2014). When making its recommendations for the National Science Foundation, Shavelson et al. (1987) advocated for systems to be able to make as many of the following comparisons as resources allow:

- against normative standards
- across nations
- across populations domestically
- within subjects over time

Porter (1991) amplified the national-comparison component, noting that state systems should also allow for state-, district-, and school-level comparisons. The more comparisons a system enables, the more flexibility a state has to employ innovative methods and make more sophisticated decisions in response to a wider range of data.

Stakes

Several researchers have criticized high-stakes accountability (Amrein & Berliner, 2002; Raudenbush, 2004) in part because much of the current accountability infrastructure relies upon a limited number of measures being used for high-stakes purposes. Stakes, or decisions about how incentives or consequences attach themselves to outcomes, should be proportional to the confidence that system designers have in the measures themselves and the desire to influence local educational practice in addition to measuring it.

The subsequent section presents several states' approaches to multiple measures. While no broad generalizations can be made about how states use multiple measures, it is interesting that the states that do employ multiple measures share few common characteristics. This suggests that multiple measures are feasible in a wide range of settings and that it may be that specific characteristics of a state's policy system are more influential in the decision to use multiple measures than are the technical aspects of the measures themselves.

States' College and Career Preparedness Indicators

The number of possible permutations of college and career preparedness measures exceeds the number of states in the Union. This illustrates an important point made by Mikulecky and Christie (2014), that "no single formula or definition guarantees" college and career preparedness. A combination of findings from Mikulecky and Christie and the Education Commission of the States (2014) reveal an array of indicators states used when adopting multiple measures for college and career preparedness, including:

- ACT WorkKeys
- ACT/SAT participation and/or results*
- Advanced Placement (AP) participation and/or results*
- Career Technical Education certifications/competencies
- College-going rate*
- Dual enrollment participation and/or completion*
- Industry certifications earned*
- International Baccalaureate (IB) participation and/or results*
- Percentage of students enrolled in postsecondary programs*
- Percentage of students needing college remediation*
- Percentage of students taking algebra in Grade 8*
- Percentage of students taking higher-level courses

Asterisks reflect the nine indicators Mikulecky and Christine identified as essential for consideration. But three of those indicators have been incorporated in accountability systems of four or fewer states. This suggests that the popularity of a particular college and career preparedness indicator does not necessarily correlate with opinions Mikulecky and Christie gathered from a panel of 12 experts.

Indicators chosen for inclusion in an accountability system may depend upon the state's preference for measurement/reporting type as discussed in the section on design criteria. Some states use indicators to measure college and career preparedness by calculating that indicator into school ratings. Other states use indicators to report college and career preparedness on school report cards, but do not calculate that indicator in school ratings. Other states measure *and* report an indicator. The decision to measure, report, or both could affect state education policy and student outcomes. For instance, no state currently measures the percentage of students needing college remediation or the percentage of students taking algebra in Grade 8.

Remediation is an important measure because delaying enrollment in credit-bearing courses generally results in a decrease in college success overall. Bailey, Jeong, and Cho (2010) reported students in need of remediation suffer a variety of costs: financial (e.g., additional fees and debt), psychological (e.g., the shock of ostensibly returning to high school work), and opportunities (e.g., lost time and earnings), as do taxpayers. Scott-Clayton and Rodriguez (2012) studied data from 100,000 students at a large, urban community-college system, finding remedial math classes diverted one-quarter and remedial reading diverted 70% of students who "would have earned a B or better in the relevant college course." Accurate identification of students in need of remediation could have dramatic impacts on a P-20 system and students' academic and financial postsecondary experiences.

Additionally, taking algebra in Grade 8 predicts the number of high school mathematics courses taken (Spielhagen, 2006), which in turn has a direct effect on admission to competitive postsecondary schools and professional programs (Schiller and Müller, 2003). Also, math course-taking behavior strongly correlates to students' earnings 10 years after high school even when controlling for student demographic, family, and school variables, as well as highest educational degree attained, college major, and occupation (Rose & Betts, 2004). Some of Rose and Betts' estimates suggest algebra credits have the largest effect. However, these factors also correlate strongly with income and the availability of programs that provide students these opportunities.

Despite empirical findings that consider college remediation rates and eighth-grade algebra participation as strong indicators of potential postsecondary preparedness, EPIC identified two states measuring college remediation rates (Hawaii and Missouri) and no state measuring eighth-grade algebra participation for accountability purposes.

Finally, the use of indicators that measure both college and career preparedness (e.g., dual enrollment, postsecondary enrollment, cumulative high school GPA) should be balanced against those that focus exclusively on college preparedness (e.g., ACT/SAT, AP and IB, courses taken to meet college entrance requirements, percentages of students needing college remediation) or exclusively on career preparedness (e.g., industry certifications, WorkKeys, CTE courses taken).

The remainder of this section details the approaches seven states have adopted to incorporate a college and career preparedness indicator in some fashion in their

accountability systems. Variation is considerable from state to state. Each state represents a different way of thinking about a multiple-measures system.

Georgia

Georgia’s College and Career Ready Performance Index uses achievement (60%), progress (25%), and achievement gap (15%) measures to calculate an overall score out of 100 points for high schools. The achievement measure includes 18 separate indicators across three categories: content mastery, post-high school readiness, and graduation rates. The post-high school readiness category accounts for 30% of the achievement measure, or 18% of the overall score, and includes eight indicators. Table 4 illustrates these indicators. Each indicator is assigned a benchmark percentage against which school performance is measured.

Table 4. Georgia's Post-High School Readiness Indicators

Indicator	Description
1. Course-taking behavior	Percentage of graduates completing one of the following: <ul style="list-style-type: none"> • Career Technical/Agricultural Education (CTAE) pathway • advanced academic pathway • fine arts pathway • world language pathway
2. Career preparedness	Percentage of CTAE pathway completers earning one of the following: <ul style="list-style-type: none"> • national industry-recognized credential • IB Career-Related Certificate • passing score on a state-recognized, end-of-pathway assessment (beginning in 2014–2015)
3. College-course preparedness	Percentage of graduates: <ul style="list-style-type: none"> • entering two- or four-year in-state colleges not requiring remediation or learning-support courses • scoring program-ready on ACT’s Compass • scoring at least 22 out of 36 on the composite ACT • scoring at least 1550 out of 2400 on the combined SAT • scoring 3 or higher on two or more AP exams, or • scoring 4 or higher on two or more IB exams
4. Dual or concurrent enrollment	Percentage of graduates earning high school credit(s) for <ul style="list-style-type: none"> • accelerated enrollment via ACCEL • Dual HOPE Grant • Move On When Ready • Early College • Gateway to College • Advanced Placement courses, or • International Baccalaureate courses
5. Postsecondary writing preparedness	Percentage of students scoring at <i>Meets</i> or <i>Exceeds</i> on the Georgia High School Writing Test
6. Postsecondary reading preparedness	Percentage of students achieving a Lexile measure greater than or equal to 1275 on the American Literature End-of-Course-Test (EOCT)
7. Postsecondary overall academic preparedness	Percentage of EOCT assessments scoring at the <i>Exceeds</i> level
8. Attendance	Student Attendance Rate (%)

For instance, the benchmark percentage for the course-taking indicator is 100, meaning that 100% of a school's students are expected to complete one of the applicable course pathways. If 50% of graduates complete an applicable pathway the school is awarded 5 points for this indicator. The points earned over all eight indicators are summed and divided by 80 (each indicator is worth 10 points). The resulting percentage is multiplied by .30 to generate the weighted post-high school readiness score. Thus, Georgia follows a model that has compensatory, conjunctive, and complementary elements.

Florida

Unlike Georgia, Florida's model is primarily compensatory. It uses a 1,600-point scale to calculate A–F grades for high schools. Similar to the planned revisions to California's API, no more than 50% of Florida's high school grades are determined by state standardized test scores that measure performance, learning gains for all students, and learning gains for the bottom 25% of students for mathematics and reading. Graduation rate accounts for 18.75% of the high school grade. An acceleration component worth 18.75% measures student participation and performance in AP, IB, or other courses where students can earn college credit. Lastly, the college readiness component, worth 12.5% of the overall school grade, measures the performance on the ACT, SAT, Florida College Entry-Level Placement Test, or the Postsecondary Education Readiness Test. The flexibility allowed in the college readiness component introduces a complementary element to Florida's model.

Texas

Texas's core framework has elements of a matrix because it equally weights four criteria: student achievement, student progress, closing performance gaps, and postsecondary readiness, each of which is graded *high* or *low*. However, Texas does produce a number to be used for ranking, similar to compensatory approaches seen in Florida and California. Texas's Performance Index Framework grades schools on a three-point scale: met standard, met alternative standard for alternative education and charters, or improvement required. Like Texas's overall framework, its postsecondary readiness measure includes four equally weighted indicators: (a) percentage of students meeting the postsecondary readiness level on one or two State of Texas Assessments of Academic Readiness (STAAR) assessments, (b) graduation rates, (c) the percentage of students who graduated under the Recommended High School Program or Distinguished Achievement Program, and (d) the percentage of graduates meeting college-ready criteria on the reading/English language arts and mathematics Texas Assessment of Knowledge exit-level test, the SAT, or ACT. Texas also injects a complementary component by allowing schools to earn distinction designations for participation and advanced performance on STAAR assessments, SAT/ACT, or AP/IB exams.

Kentucky

Figure 1 presents Kentucky's college- and career-readiness assessment model. Schools in Kentucky receive an overall score out of 100 points based on three

components: Next Generation Learners, Next Generation Instructional Programs and Support, and Next Generation Professionals. Worth 70% of a school's overall score, the Next Generation Learners component measures: (a) student achievement, (b) achievement gaps, (c) individual student growth, (d) graduate rate, and (e) college and career readiness. Kentucky allows students to earn college-ready, career-ready, or both college- and career-ready status. Figure 1 shows the combinations of possibilities students must meet with benchmark scores on the ACT test, ACT's Compass (college placement test), Kentucky Online Testing Program (KYOTE), or an industry certificate. Students can earn both college- and career-ready status by meeting a benchmark on the ACT test, ACT's Compass, or KYOTE *and* meeting the benchmark on the Kentucky Occupational Skills Standards Assessment (KOSSA) or earning an industry certification. Overall, Kentucky's model is compensatory, but its college- and career-ready bonus takes a conjunctive approach.

College Ready	Career Ready		Bonus: College and Career	
	Academic	Technical	College	Career
Must meet the benchmark on one of the following:	Must meet the benchmark for one of the following:	Must meet the benchmark or earn one of the following:	Must meet the benchmark or earn one of the following: ⁵	Must meet the benchmark or earn one of the following:
ACT	ASVAB	KOSSA	ACT	KOSSA
ACT's Compass	ACT's WorkKeys ⁶	Industry Certificate	ACT's Compass	Industry Certificate
KYOTE			KYOTE	

Figure 1. Kentucky College Ready, Career Ready, and College *and* Career Ready

New Mexico

New Mexico calculates A–F school grades with 30% based on student achievement, 30% based on achievement growth for the highest- and lowest-performing students, 17% on graduation rate, 15% on college and career readiness (CCR), and 8% on opportunity to learn. Student performance on one of the 10 following indicators accounts for one third (5%) of the CCR score:

1. PSAT or National Merit Scholarship Qualifying Test
2. SAT
3. College Board's ACCUPLACER assessment
4. ACT's PLAN assessment
5. ACT
6. ACT's Compass assessment
7. one Advanced Placement exam

⁵ Meeting the college-ready academic requirement means students have to satisfy the career-ready academic requirement to earn the bonus distinction of college and career ready.

⁶ Students have to meet the benchmark for the three NCRC WorkKeys assessments.

8. one International Baccalaureate exam
9. concurrent or dual enrollment
10. Career Technical Education (CTE) course pathway completion

Student participation on the same indicator accounts for two thirds (10%) of the CCR score.

Students may make multiple attempts, with multiple indicators, in multiple years. The most successful indicator is retained, making New Mexico's CCR model complementary.

Oklahoma

Oklahoma uses student achievement and student growth to calculate A–F grades based on a 100-point scale for high schools. Achievement and growth are each worth 50% of the total school grade, but growth is calculated overall and for the bottom quartile of students, with each worth 25%. High schools may add as many as nine college and career preparedness indicator-related bonus points to their final grades. Schools with cohort graduation rates above 90% acquire five bonus points. Schools can acquire a bonus point for meeting or exceeding state-determined thresholds on each of the following indicators:

- participation or performance in advanced coursework (e.g., AP, IB, concurrent/dual enrollment, Advanced International Certificate of Education, or CTE courses that lead to an industry certificate)
- participation or performance on SAT or ACT
- graduating low-achieving, 8th grade students from high school on time, or
- improving year-to-year growth on three of the other bonus-point categories.

Oklahoma's model is primarily compensatory, modified by a complementary bonus structure. Pennsylvania operates similarly to Oklahoma, but adopted different indicators.

Missouri

Missouri uses academic achievement (40%), graduation rates (21%), college and career readiness (21%), subgroup achievement (10%), and attendance (7%) to create Annual Performance Report (APR) scores for high schools on a 140-point scale.⁷ Missouri's college and career preparedness calculations include three indicators: (a) percentage of graduates who scored at or above the state standard for participation and performance on the ACT test, SAT, ACT's Compass, or ASVAB; (b) percentage of graduates who earned a qualifying score on an AP, IB, or Technical-Skills Attainment assessment; and (c) percentage of graduates who attend postsecondary education/training.⁸

⁷ Percentages do not equal 100% due to rounding.

⁸ Postsecondary training/education includes: entering the military within six months of graduating high school or completing a department-approved, career-education program that leads to occupational placement within six months of training.

For each college and career preparedness indicator above, students earn points for their schools similar to the way California students earn API points for their schools. For instance, there are four levels of performance for indicator (a) above that each graduating student falls into, regardless of whether that student takes the ACT test, SAT, ACT's Compass, or ASVAB. Table 5 illustrates Missouri's calculation using SAT score levels as an example. The calculation weights the numbers of graduates at each level by a corresponding multiplier.

Missouri's college and career preparedness approach contains conjunctive, compensatory, and complementary elements. Schools are required to meet or exceed the thresholds for all three college and career preparedness measures, making this aspect of the indicator conjunctive. However, high performance on one measure (e.g., the SAT) can compensate for poor performance on another (e.g., the ACT). Finally, students are free to choose between the different measures, a feature more in line with complementary systems.

Table 5. Missouri's Calculation of the SAT as a College and Career Preparedness Indicator

SAT Math and Reading Scores Combined ⁹	Multiplier for Number of Students at Level
Graduates scoring 1190–1600	x 1.25
Graduates scoring 990–1180	x 1.00
Graduates scoring 870–980	x 0.75
Graduates scoring below 870	x 0.25
Graduates not participating	x 0

Note. The calculation adds the products of each level and divides by the total number of graduates to generate the percentage of graduates scoring at or above the state standard on the SAT. This percentage is used to calculate status and progress ratings that are added together to produce the total points for the first college and career readiness indicator.

Single- or Zero-Indicator States

Five states measure single indicators of college and career preparedness: Arkansas, Delaware, New Hampshire, Nevada, and New York. An additional 11 states measure none. Sandler (2014) describes growing interest among state policymakers in multiple measures, calling now a time to seize momentum. Multiple-measure college and career preparedness indicators could be used in accountability systems to diagnose gaps in schools, target supports and interventions, collect broader data to inform policy decisions, provide transparent public reporting of more student and school outcomes, and set improvement goals. ASCD (2013) declared multiple measures as an imperative: "Any comprehensive determination of student proficiency, educator effectiveness, or school quality must be based on more than just standardized test scores and should use a variety." Accountability systems that reduce student achievement to a single number inherently limit a school's ability to examine variations, and it is these variations

⁹ Calculated for mathematics and reading portions of the SAT, which total 1600 points. The SAT writing section is excluded.

that provide conditions for creativity, discovery, and innovation, all educational outcomes that are more challenging to capture using single, standardized measures.

Emerging Themes in the Use of Multiple Measures

Three themes run through the theoretical and practical literature on multiple measure: (a) the challenges inherent in adopting a multiple-measures system, (b) the effects that selecting one or many measures has on the system as a whole, and (c) the unintended consequences that arise from not including critical elements of college and career preparedness.

Challenges Inherent in Adopting a Multiple-Measures System

Gong and Hill (2001) identified many challenges that may keep some states from adopting multiple measures:

1. To what extent will resources (i.e., money, available expertise, and time) allow for design, development, administration, and scoring of a multiple-measures system?
2. How will adopting a multiple-measures system affect instructional time in schools?
3. To what extent has a state's expertise to this point prepared it for the complex, technical tasks of deciding and monitoring one or more methods of combining multiple measures?
4. What validity concerns should a state anticipate before making inferences and decisions from a multiple-measures system?
5. What are the trade-offs between immediate and incremental implementation of a multiple-measures system?
6. What state-specific trade-offs exist between a simpler system that allows for a narrow focus with a target of immediate gains in a specific area and a more complex system that offers a more comprehensive, and therefore likely a more long-term, view?

Effects of Choices Across the System as a Whole

More measures and research support understanding of college preparedness than career preparedness. A previous EPIC white paper identified career preparedness as receiving less attention in the literature, partly because it does not lend itself to a straightforward listing of knowledge and skills that are prerequisite for success in all careers, as well as closer proximity of college students to researchers, especially those at universities (Conley, 2014c). Consequently, states should be aware that career preparedness is more difficult to capture in a multiple-measures system. States should consider population needs, both current and future, when choosing across classifications of measures, which include measures of college *and* career preparedness, career preparedness exclusively, and college preparedness exclusively. States must determine if equal or different measurement are warranted for each classification.

Potential measures of both college and career preparedness include dual-enrollment participation or completion and percentage of students enrolled in postsecondary programs. Participation and completion are different measures with different implications, an issue discussed further in a subsequent section.

New Mexico's system may appear to provide the most breadth with 10 indicators and a complementary approach. However, eight of its 10 indicators measure college preparedness and one jointly measures college and career preparedness, meaning roughly 85% of its indicators push toward college as an outcome. By contrast, Georgia has eight indicators, four of which apply directly to college preparedness, three of which can correspond either to college or career preparedness, and one for career preparedness exclusively. Finally, Kentucky offers three pathways, essentially equalizing its emphasis on college, career, and college *and* career preparedness. However, it should be noted that Kentucky sets a single requirement for students to demonstrate college preparedness but two requirements for students to demonstrate either career preparedness or college and career preparedness (see Figure 1).

The course-taking behavior, postsecondary reading, and attendance components of Georgia's college and career preparedness indicator provide indirect examples of additional approaches to measuring college *and* career preparedness with a shared indicator. Percentages of students taking higher-level courses could potentially measure both college and career preparedness if the definition of "higher-level courses" includes advanced CTE or other career-related coursework. However, states more commonly restrict such definitions of "higher-level" to traditional academic courses.

Though some arguments exist for career preparedness assessments (e.g., ACT WorkKeys or ASVAB) having the ability to diagnose students' college and career preparedness, no independent evidence exists to support claims that career preparedness assessments also predict college preparedness. Measures exclusively assessing career preparedness offer slightly more options, including WorkKeys, ASVAB, CTE certifications or competencies, and industry certifications earned. Some states have developed their own assessments, such as the Kentucky Occupational Skills Standards Assessment. A previous EPIC white paper thoroughly details career preparedness measurements (Conley, 2014c).

Measures of college preparedness consist of measures of performance and measures of participation. Performance measures exclusively for college preparedness include ACT/SAT results, AP results, IB results, and percentage of students needing college remediation. Participation measures include college-going rate, percentage enrolled in upper-level courses, percentage of students taking Algebra in Grade 8, course-taking behavior, and participation in ACT, SAT, AP, or IB. Performance measures prioritize excellence and may provide insight into instructional quality. By contrast, participation measures prioritize equity and may provide insight into inputs and processes.

Currently, 15 states incorporate AP/IB participation and performance data into their public high school accountability systems or plan to do so. However, the role of participation rates and exam scores varies across state systems. Florida uses both

AP/IB participation and performance to determine accountability scores. Texas uses AP/IB participation and performance to award bonus points for school grades or to identify exemplar schools. Nevada exemplifies the most common approach, measuring only AP/IB performance.

Measures of Opportunity to Learn represent an additional classification worth examining. For example, measuring a school's advanced course offerings (in AP or IB) would allow a system to examine its student coverage and fair comparisons of schools in terms of college preparedness. Regarding career preparedness, measurement of CTE course availability would provide comparable insights to measuring AP or IB course availability. A previous EPIC white paper on course-taking behavior asserted that beyond issues of basic access, high schools do not present equal opportunities to students, particularly lacking parity in rigorous coursework and materials/resources necessary for effective instruction (Conley, 2014a).

College and Career Preparedness Indicator Components Missing from Multiple-Measures Systems

Though no single formula guarantees a state can measure all elements that contribute to college and career preparedness, some crucial elements seem noticeably absent from multiple-measures systems currently in use by states. These include measures of expository writing, speaking and listening, academic mindset, goal orientation and aspirations, learning techniques such as study skills, metacognitive skills such as employing a range of learning strategies, proficiency in languages other than English, and creativeness and expressiveness in the arts and in core academic classes.

Though a high school graduate's later use of writing may depend upon the type of college or career pursued, written communication is an indispensable skill for many postsecondary options. Despite the student value in writing ability, no state but Georgia directly includes writing among its college and career preparedness indicators. In fact, many states that use the SAT in their college and career preparedness indicators opt to examine only mathematics and reading scores, ignoring the writing section. Similarly, EPIC found little performance assessment among multiple-measures systems.

Furthermore, metacognitive skills have yet to be included in any statewide accountability system (though the seven California Office to Reform Education districts are piloting some metacognitive measures). Additionally, credits earned in languages other than English may be required for graduation in some states (one credit is required to graduate from a California high school, but in most states only college readiness pathways require language; some states do not require any). However, states that include graduation rates among their college and career preparedness indicator calculations may not have the data-collection capacity or may not examine their data thoroughly enough to identify when course-taking of languages other than English determines growth or decline in graduation rates. This would be an ideal opportunity for the cyclical approach (Shavelson et al., 1987).

The absence of measures of a much wider range of skills and abilities associated with

college and career success is striking. In a previous white paper on innovative measures, EPIC examined 38 clusters of college majors identified by the College Board. Of those, 32, or 84%, depended on performance demonstrations or writing as the primary means of assessment. EPIC also examined admissions policies at higher education systems and found that the flagship universities in 36 of 50 states require a minimum of two years of courses in languages other than English for admission. More selective public universities (e.g., University of Michigan, University of North Carolina at Chapel Hill, University of Wisconsin–Madison, University of Texas–Austin, and several campuses of the State University of New York) have even higher standards to demonstrate the level of high school academic rigor in languages other than English necessary for admission (Conley, 2014b). It seems that measures of a much wider range of skills and strategies could provide considerable value as indicators of college and career preparedness.

New Conceptions of Accountability

Conley and Darling-Hammond (2013) endorse uses of multiple measures that “contribute to a comprehensive picture of the quality of learning in classrooms, schools, schools systems, and states.” They recommend supplementing standardized test cut scores with additional data to reduce the likelihood of misclassification, yielding student “profiles of information for evaluating and conveying insights” rather than a single benchmark score. This information can be aggregated upward to reach accountability decisions about individual schools and the system as a whole.

It is important to note that adopting a multiple-measures system does not automatically enhance an accountability system’s quality or the quality of inferences drawn from it. However, Chester (2005) showed that combining performance indicators, a performance index, a growth calculation, and AYP status enabled Ohio to classify schools as Excellent, Effective, engaged in Continuous Improvement, on Academic Watch, or in Academic Emergency, resulting in an expanded definition of effectiveness, better identification of school levels, and a higher degree of confidence in the accountability system’s legitimacy.

Ultimately, a state’s decision to adopt multiple measures of college and career preparedness depends upon how it wants to manage an array of factors including stakeholder collaboration, design method, breadth of coverage, measurement and reporting method, data integration, ability to make comparisons, and level of stakes. Such determinations are not easily made because they rely upon complex and often competing values. Adding to the challenge is the fact that the use of multiple measures for accountability purposes has been on hold for well over a decade and is only now again beginning to be viewed as a viable option for state policy. The potential for a highly valid accountability system that influences practice in positive ways and that has strong practitioner buy-in and ownership is one of the key factors that makes it worthwhile to consider the technical challenges of a multiple-measures approach to a college and career preparedness indicator.

References

- ACT. (2014). National ranks for test score and composite score. Retrieved from <http://www.actstudent.org/scores/norms1.html>
- American Federation of Teachers. (2008). *Sizing up state standards, 2008*. Retrieved from <https://www.aft.org>
- Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education Policy Analysis Archives*, 10(18), 1–74.
- Association for Supervision and Curriculum Development. (2013). *ASCD policy points: Multiple measures of accountability*. Retrieved from <http://www.ascd.org/ASCD/pdf/siteASCD/publications/policypoints/Multiple-Measures-of-Accountability.pdf>
- Bailey, T., Jeong, D. W., & Cho, S. W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, 29(2), 255–270.
- Blank, R. K. (1993). Developing a system of education indicators: Selecting, implementing, and reporting indicators. *Educational Evaluation and Policy Analysis*, 15(1), 65–80.
- Brookhart, S. M. (2009). The many meanings of "multiple measures." *Educational Leadership*, 67(3), 6–12.
- Chester, M. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice*, 24(4), 40–52.
- Conley, D. T. (2012). Does common mean “the same”? Implementing new state standards and assessments equitably and fairly. *The State Education Standard*, 13(1), 23–28.
- Conley, D. T. (2014a). *Measures for a college and career indicator: Research brief on course-taking behavior*. Eugene, OR: Author.
- Conley, D. T. (2014b). *Measures for a college and career indicator: Research brief on innovative measures*. Eugene, OR: Author.
- Conley, D. T. (2014c). *Measures for a college and career indicator: Research brief on career assessment*. Eugene, OR: Author.
- Conley, D. T., & Darling-Hammond, L. (2013). *Creating systems of assessment for deeper learning*. Stanford, CA: Stanford Center for Opportunity Policy in Education.
- EdTrust. (2014). Accountability. Retrieved from <http://www.edtrust.org/west/our-work/agenda-at-a-glance/accountability>

- Education Commission of the States. (2014). *50 state analysis: School accountability 'report cards'*. Retrieved from <http://ecs.force.com/mbdata/MBquest3RT?Rep=ar10>
- Ehren, M. C., & Swanborn, M. S. (2012). Strategic data use of schools in accountability systems. *School Effectiveness and School Improvement*, 23, 257–280.
- Finn, C. E., Jr., Petrilli, M. J., & Vanourek, G. (1998). The state of state standards. *Fordham Report*, 2(5).
- Francis, D. J., Fletcher, J. M., Stuebing, K. K., Lyon, G. R., Shaywitz, B. A., & Shaywitz, S. E. (2005). Psychometric approaches to the identification of LD IQ and achievement scores are not sufficient. *Journal of Learning Disabilities*, 38(2), 98–108.
- Gong, B., & Hill, R. (2001). *Some considerations of multiple measures in assessment and school accountability* [PowerPoint slides]. Retrieved from http://www.nciea.org/publications/MultiMeasures_Gong01.pdf
- Greene, J., & Foster, G. (2003). *Public high school graduation and college readiness rates in the United States*. Retrieved from http://www.manhattan-institute.org/html/ewp_03.htm
- Gutiérrez, R. (2008). A 'gap-gazing' fetish in mathematics education? Problematizing research on the achievement gap. *Journal for Research in Mathematics Education*, 357–364.
- Harris, D. N. (2013). *How might we use multiple measures for teacher accountability?* Retrieved from http://www.carnegieknowledgegenetwork.org/briefs/multiple_measures/
- Hentschke, G., Wohlstetter, P., Hirman, J., & Zeehandelaar, D. (2011). Using state-wide multiple measures for school leadership and management: costs incurred vs. benefits gained. *School Leadership and Management*, 31(1), 21–34.
- Jos, P. H., & Tompkins, M. E. (2004). The accountability paradox in an age of reinvention. *Administration and Society*, 36, 255–81.
- Klein, D., Braams, B. J., Parker, T., Quirk, W., Schmid, W., & Wilson, W. S. (2005). *The state of state math standards, 2005*. Retrieved from <http://www.math.jhu.edu/~wsw/ED/mathstandards05FINAL.pdf>
- LeFebvre, M., Clark, H., Burkum, K., & Kyte, T. (2013). *The condition of work readiness in the United States*. Retrieved from <http://www.act.org/research/policymakers/pdf/ConditionWorkReadiness.pdf>
- Linn, R. L. (2005). *Issues in the design of accountability systems*. Retrieved from <http://www.cse.ucla.edu/products/reports/r650.pdf>
- Manna, P. (2011). *Collision course: Federal education policy meets state and local realities*. Washington, DC: CQ Press.
- Marion, S., & Gong, B. (2003). *Evaluating the validity of state accountability systems* [PowerPoint slides]. Retrieved from

http://www.nciea.org/publications/RILS2003_BGSM03.pdf

- Mikulecky, M., & Christie, K. (2014). *Rating states, grading schools: What parents and experts say states should consider to make school accountability systems meaningful*. Retrieved from <http://www.ecs.org/docs/rating-states,grading-schools.pdf>
- National Center for Public Policy and Higher Education. (2010). *Beyond the rhetoric (Improving college readiness through coherent state policy): The gap between enrolling in college and being ready for college*. Retrieved from http://www.highereducation.org/reports/college_readiness/gap.shtml
- Perie, M., Park, J., & Klau, K. (2007). *Key elements for educational accountability models*. Retrieved from http://www.ccsso.org/Documents/2007/Key_elements_for_educational_2007.pdf
- Porter, A. C. (1991). Creating a system of school process indicators. *Educational Evaluation and Policy Analysis*, 13(1), 13–29.
- Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* Retrieved from http://www.ets.org/Media/Education_Topics/pdf/angoff9.pdf
- Rose, H., & Betts, J. R. (2004). The effect of high school courses on earnings. *Review of Economics and Statistics*, 86(2), 497–513.
- Sandler Foundation. (2014). *Questions and answers about multiple measures*. Retrieved from http://www.sandlerfoundation.org/wp-content/uploads/Sandler-Multiple-Measures-Report_Questions-and-Answers.pdf
- Schiller, K. S., & Muller, C. (2003). Raising the bar and equity? Effects of state high school graduation requirements and accountability policies on students' mathematics course taking. *Educational Evaluation and Policy Analysis*, 25, 299–318.
- Scott-Clayton, J., & Rodriguez, O. (2012). *Development, discouragement, or diversion? New evidence on the effects of college remediation* (No. 18328). Cambridge, MA: National Bureau of Economic Research.
- Shavelson, R. J., McDonnell, L., Oakes, J., Carey, N., & Picus, L. (1987). *Indicator systems for monitoring mathematics and science education*. Retrieved from <http://www.rand.org/content/dam/rand/pubs/reports/2007/R3570.pdf>
- Spielhagen, F. R. (2006). Closing the achievement gap in math: Considering eighth grade algebra for all students. *American Secondary Education*, 34(3), 29–42.
- Stotsky, S. (2005). The state of state English standards, 2005. Retrieved from the Thomas B. Fordham Institute website: http://edex.s3-us-west-2.amazonaws.com/publication/pdfs/FullReport%5B01-03-05%5D_7.pdf
- WestEd & EPIC. (2013). *National Assessment of Educational Progress grade 12 preparedness research project job training programs curriculum study: Final report*. San Francisco, CA, and Eugene, OR: Authors.