Docket ID ED-2016-OESE-0047

COMPARABILITY OPTIONS FOR STATES APPLYING FOR THE INNOVATIVE ASSESSMENT AND ACCOUNTABILITY DEMONSTRATION AUTHORITY: COMMENTS SUBMITTED TO THE UNITED STATES DEPARTMENT OF EDUCATION REGARDING PROPOSED ESSA REGULATIONS[1]

National Center for the Improvement of Educational Assessment

Susan Lyons and Scott Marion

September 7, 2016

Executive Summary and Key Policy Recommendations

John King, Secretary of Education, proposed new regulations under title I, part B of the Elementary and Secondary Education Act of 1965 (ESEA) to implement changes made to the ESEA by the Every Student Succeeds Act (ESSA) enacted on December 10, 2015, including the ability of the Secretary to provide demonstration authority to a State educational agency (SEA) to pilot an innovative assessment and use it for accountability and reporting purposes under title I, part A of the ESEA before scaling such an assessment statewide. This document is focused on the comparability requirements spelled out in §200.77 of the draft regulations in large part because this is one of the trickier issues for states to wrestle with and it was such a prominent feature of the proposed regulations.

As spelled out in the full document that follows, the recommendations contained herein are based on insightful contributions from some of the most prominent measurement, accountability, and innovation experts in the United States. The document provides a robust conceptualization of comparability and discusses how such a conceptualization should be applied to states proposing an innovative assessment and accountability system. We then provide a framework for designing options to evaluate comparability that considers the types of measures (items) and student sample used. As called for in §200.77, we offer more than a dozen potential approaches for evaluating comparability beyond the three proposed by ED in §200.77. We do not mean for this to be an exhaustive list, rather it should be considered a set of illustrative exemplars to highlight key aspects of the proposed framework.

[1] To be cited as: Lyons, S. & Marion, S. F. (2016). Comparability options for state applying for the Innovative Assessment and Accountability Demonstration Authority: Comments submitted to the United States Department of Education regarding proposed ESSA regulations. Retrieved from www.nciea.org.

Key Policy Recommendations

We offer several recommendations, highlighted in bold text, in the full document that follows. We summarize them briefly here, but urge the reader to review the context and associated explanation for the recommendations found in the full document.

1. The U.S. Department of Education (ED) should not focus too narrowly on establishing strict comparability between the old and new assessment systems, because by doing so, ED will end up constraining innovation.
2. States' evidence of comparability of assessment results should focus at the level of the proficiency (achievement level) classifications across the two assessment systems.
3. States must propose, as part of innovative pilot applications, how they intend to document and evaluate comparability within and among the pilot districts when the innovative assessment system affords some degree of local flexibility. ED should NOT require specific methods for evaluating comparability because such evaluations will be context dependent, but such information should be included in pilot applications and reviewed by peers.
4. States should submit evidence that the innovative assessment system is aligned to the state standards and has performance expectations that are consistent with the state assessment and can therefore be employed to support the same uses in the statewide accountability system. In other words, in addition to evidence of consistency in proficiency classifications, states should be expected to provide evidence of alignment to the content standards so that the state can document the extent to which all students are being provided an opportunity to learn the required content standards at the expected level of cognitive complexity.
5. The most compelling evidence of alignment for the two assessment systems will be based on the alignment of each system to the *content standards* rather than alignment of one assessment system to the other.
6. Pilot to non-pilot comparability analyses must begin with establishing a common set of achievement level descriptors that is shared across the two assessment systems. If a state wishes to use different achievement level descriptors for the innovative assessment system, a rationale for that decision and a discussion of how those differences impact the planned comparability analyses should be provided.
7. States must propose a specific approach or approaches for evaluating comparability tied to the context of the state and the proposed innovative learning system that includes a comprehensive approach to comparability evaluations including within-pilot comparability analyses as well as pilot to non-pilot comparability studies.
8. Any comparability proposal should be evaluated according to the inferences that the design or designs can defend. If ED maintains the three options for comparability proposed in the draft regulations, we strongly recommend AGAINST having a higher bar for options that differ from the three proposed by ED.
9. When feasible or where evidence may be lacking strength, states should consider multiple approaches to comparability evaluations to provide a more complete picture of the degree of comparability in the achievement levels across the two assessment systems.
10. We strongly recommend AGAINST setting a standard criterion, or "comparability bar," for determining how comparable is comparable enough because the intended uses and the contextual factors surrounding the evaluation of comparability are critical. We offer

suggestions in the full document for considering reasonable expectations for the amount of variability that can be expected across the two assessment programs such as contextualizing the differences in results across the two systems in terms of the variability observed in the state system either within a given year or from year-to-year.

11. As the innovation reaches critical mass and spreads across the state, comparability between the two assessment systems becomes less important than the comparability of results among districts *within* the innovative system of assessments.
12. If the evidence for comparability across the two systems of assessment is strong, comparability need not be re-evaluated every year. Once it has been established, the state should provide evidence that the processes and procedures in place are sufficient for replicating the program across years and then perhaps auditing the comparability after two or three years to confirm these results.

Specific Regulatory Recommendations

In order to be as constructive as possible, we provide specific potential changes to the proposed §200.77 regulations that are coherent with the general recommendations offered above. The proposed §200.77 regulations are copied below and we use ~~strikethrough~~ and underlined text to indicate recommended deletions and additions, respectively.

(4) Provide for comparability to the State academic assessments under section 1111(b)(2) of the Act, including by generating results that are valid, reliable, and comparable for all students and for each subgroup of students under section 1111(b)(2)(B)(xi) of the Act, as compared to the results for such students on the State assessments and to other districts participating in the pilot. Consistent with the SEA's or consortium's evaluation plan under §200.78(e), the SEA must plan to annually determine comparability during each year of its demonstration authority until the state can demonstrate that the evidence for comparability across the two assessment systems is strong and the processes and procedures in place are sufficient for replicating the results. ~~period in one of the following ways:~~ States must provide evidence regarding the comparability of the inferences associated with the achievement level determinations tied to the specific state context and to the nature of the proposed innovation that are:

- Comparable between participating pilot districts and non-pilot districts; and
- Comparable among participating pilot districts (when the innovative system allows participating districts opportunities to select and administer different assessments as part of the overall assessment system)

[**NOTE TO ED: We recommend deleting the three options below and incorporating these options as well as the recommendations in our document into a non-regulatory guidance**.]

(i) ~~Administering full assessments from both the innovative and statewide assessment system to all students enrolled in schools participating in the demonstration authority, such that at least once in any grade span (e.g., 3-5, 6-8, or 9-12) and subject for which there is an innovative assessment, a statewide assessment in the same subject would also be administered to all such students. As part of this demonstration, the innovative assessment and statewide assessment need not be administered to an individual student in the same school year.~~

~~(ii) Administering full assessments from both the innovative and statewide assessment system to a demographically representative sample of students and subgroups of students under section 1111(c)(2) of the Act, from among those students enrolled in schools participating in the demonstration authority, such that at least once in any grade span (e.g., 3-5, 6-8, or 9-12) and subject for which there is an innovative assessment, a statewide assessment in the same subject would also be administered in the same school year to all students included in the sample.~~

~~(iii) Including, as a significant portion of the innovative and statewide assessment systems in each required grade and subject in which both assessments are administered, common items that, at a minimum, have been previously pilot tested or field tested for use in either the statewide or innovative assessment system.~~

~~(iv) An alternative method for demonstrating comparability that an SEA can demonstrate will provide for an equally rigorous and statistically valid comparison between student performance on the innovative assessment and the existing statewide assessment, including for each subgroup of students under section 1111(b)(2)(B)(xi) of the Act~~.

Introduction

The Every Student Succeeds Act (ESSA) provides states with a significant opportunity to develop an innovative assessment system that supports the state's vision for student-centered, personalized learning or other systems designed to promote deeper and more engaged learning. While there are a number of provisions in ESSA that states can leverage to build these systems, the Innovative Assessment and Accountability Demonstration Authority (hereafter known as the "innovative pilot" or the "Demonstration Authority") authorized under Section 1204 provides states with an unprecedented opportunity to develop next generation approaches to assessment that transcend the standardized tests commonly used to evaluate student and school performance.

This document is focused on the comparability requirements spelled out in §200.77 of the draft regulations in large part because this is one of the trickier issues for states to wrestle with and it is such a prominent feature of the proposed regulations. The recommendations contained in this document were drafted by Susan Lyons, Ph.D. and Scott Marion, Ph.D., Associate and Executive Director, respectively at the National Center for the Improvement of Educational Assessment (Center for Assessment) but initial and subsequent drafts of the recommendations were vetted and endorsed by the following leading measurement specialists with expertise in comparability:

- Randy Bennett, Ph.D., Norman O. Frederiksen Chair in Assessment Innovation in the Research & Development Division at Educational Testing Service
- Henry Braun, Ph.D., Boisi Professor of Education and Public Policy and Director, Center for the Study of Testing, Evaluation, and Education Policy
- Robert Brennan, Ph.D., E. F. Lindquist Chair of Measurement and Testing in the College of Education at The University of Iowa and Founding Director of the Center for Advanced Studies in Measurement and Assessment (CASMA)
- Derek Briggs, Ph.D., Professor and Chair of the Research and Evaluation Methodology Program at the University of Colorado, Boulder
- Linda Cook, Ed.D. Former director of the Center for Validity Research, Educational Testing Service
- Joan Herman, Ed.D. Director Emerita of the Center for Research on Student Standards and Testing (CRESST) at UCLA
- Stuart Kahl, Ph.D., Founder and former president of Measured Progress
- Richard Luecht, Ph.D., Professor of Educational Research, Measurement and Evaluation at University of North Carolina at Greensboro
- Lauress Wise, Ph.D., Principal Scientist and former president at Human Resources Research Organization (HumRRO)

Additionally, the following professionals from the Center for Assessment, all experts in comparability, contributed to and endorse these recommendations:

- Juan D'Brot, Ph.D., Senior Associate
- Nathan Dadey, Ph.D., Post-doctoral fellow

- Chris Domaleski, Ph.D., Associate Director
- Erika Hall, Ph.D., Senior Associate
- Joseph Martineau, Ph.D., Senior Associate
- Thanos Patelis, Ph.D., Senior Associate

Finally, we received important and useful feedback from the following leaders in innovative assessment and accountability systems:

- Linda Darling-Hammond, Ed.D., President of the Learning Policy Institute and Charles E. Ducommun Professor of Education Emeritus at Stanford University
- Paul Leather, Deputy Commissioner, New Hampshire Department of Education
- Lillian Pace, Senior Director of National Policy at KnowledgeWorks
- Maria Worthen, Vice President for Federal and State Policy at iNACOL

Focusing on Comparability

The draft federal regulations would require that states "provide for comparability to the State academic assessments under section 1111(b)(2)." The comparability requirement is only necessary when a state is proposing to use the innovative assessment system with a subset of school districts. In spite of the challenges of implementing and evaluating the comparability of two assessment systems operating within the state at once, we strongly support starting the innovative pilot with a subset of districts because truly innovative assessment systems are likely to require considerable support and commitment for successful implementation and to build the body of validity evidence and program processes are strong enough to responsibly scale statewide.

The issue of comparability across the two systems is of primary concern for two reasons. First, because states must incorporate assessment results from the pilot districts into the state accountability system alongside the results generated from the non-pilot districts, the assessment systems must produce results that are comparable enough to support their simultaneous use in the single statewide accountability system. Secondly, requiring that the assessment systems produce comparable results ensures that states will not view the innovative assessment and accountability demonstration authority as a way to relax the rigorous expectations for student achievement established under the current state assessment systems. The innovative assessment systems designed under the demonstration authority must be aligned to the intended content standards and produce annual summative determinations that are consistent across the two assessment programs. This does not require scale score comparability, but does require the ability to meaningfully compare the achievement level classifications for use in the accountability system.

To address these two major concerns, states will be asked to provide evidence of comparability of assessment results, which we **recommend should focus at the level of the proficiency classifications** across the two assessment systems. Evidence of comparability would support the notion that in general, schools that are participating in the innovative assessment system could be expected to have similar distributions of students into performance classifications had the school instead participated in the statewide standardized assessment system. This is not to say that we

would expect all districts that participate in the innovative pilot to exhibit similar levels of achievement as the non-pilot districts. Pilot districts will be most certainly a non-random sample and the innovative learning model associated with the assessment system should influence achievement, the performance of students in each group of schools may well differ. However, it should remain the case that the performance standards in both pilot and non-pilot settings support the same interpretations.

Though the two primary concerns mentioned above, comparability for school accountability and comparability of expectations for student achievement, are defensible, a narrow focus on pilot to non-pilot comparability misses the bigger picture in two important ways: 1) by potentially inhibiting innovation, and 2) by failing to address additional, and potentially more important, comparability questions. **First, if the U.S. Department of Education (ED) focuses too narrowly on establishing strict comparability between the old and new assessment systems, it is likely that the assessment systems designed under this new option for flexibility—which is intended to drive innovation—will not be innovative**. There are a variety of reasons why there may be legitimate differences in the results produced by the two or more assessment systems. States likely would take advantage of the innovative assessment and accountability demonstration authority for one of four reasons: 1) to measure the state-defined learning targets more efficiently (e.g., reduced testing time), 2) to measure the learning targets more flexibly (e.g., when students are ready to demonstrate "mastery"), 3) to measure the learning targets more completely and/or deeply, or 4) to measure targets from the standards that are not measured in the general statewide assessment (e.g., listening, speaking, extended research, scientific investigations). Therefore, requiring the results produced across the old and new systems to tell the exact same story about student achievement has the very real potential to prevent meaningful innovation. **To quote one of the leading experts on score comparability, Dr. Robert Brennan, when asked about comparability between the innovative and standardized assessment systems, "perfect agreement would be an indication of failure."**

The emphasis on pilot to non-pilot comparability misses an important set of potential threats to equity due to local flexibility under the demonstration authority. Because local assessment information can now be used to inform accountability determinations, the comparability of assessment system scores within and across pilot districts will be an important comparability challenge faced under the Demonstration Authority. Allowing for local flexibility in the assessment results used for accountability determinations is new territory for states. This type of innovation will call for new, close relationships between LEAs and SEAs in order to arrive at common understandings about the content, content alignment, assessment quality, quality control, ownership, and data sharing. Ensuring that the innovative assessment system is producing results that are comparable within and among innovative districts will require new ways to conceptualize the gathering of comparability evidence as discussed in detail in Lyons, Evans, & Marion, 2016 and Lyons, Marion, Pace & Williams, 2016. Comparability within and among pilot districts is necessary but not sufficient for pilot to non-pilot comparability. To provide evidence of comparability across the innovative and current assessment systems, states first will need to demonstrate how they are going to evaluate comparability within and among pilot districts. **Therefore, we recommend that as part of innovative pilot applications, states propose how they intend to document and evaluate comparability within and among the pilot districts. We do NOT recommend that ED require specific methods for evaluating**

these levels of comparability because such evaluations will be context dependent, but information on approaches to evaluating comparability among pilot districts should be included in pilot applications and reviewed by peers.

Defining Comparability

Comparability is a judgment based on an accumulation of evidence to support claims about the meaning of test scores and whether scores from two or more tests or assessment conditions can be used to support the same interpretations and uses. In this way, assessments are not dichotomously determined to be comparable or not, but like validity, comparability is a judgment about the strength of the theory and evidence to support the comparability of score interpretations for a given time and use. This means that evidence used to support claims of comparability will differ depending on the nature (or grain-size) of the reported scores. For example, supporting claims of raw score (number correct) interchangeability—the strongest form of comparability—would likely require the administration of a single assessment form with measurement properties that are the same across all respondents (i.e., measurement invariance). Most state assessment systems with multiple assessment forms fail to meet this level of score interchangeability. Instead, the design of most state assessment systems aims to be "comparable enough" to support scale score interchangeability. This level of comparability typically requires that the multiple test forms are designed to the same blueprint, administered under almost identical conditions, and scored using the same rules and procedures. Still, many states are currently struggling to meet this level of comparability due to challenges with multiple modes of administration—paper, computer, and devices (see DePascale, Dadey & Lyons, 2016). In this way, comparability is an evidence-based argument, and the strength of evidence needed will necessarily depend on the type and use of the score being supported. As shown in Figure 1, comparability lies on a continuum and rests on two major critical dimensions: the comparability of content and the comparability of scores, and that each of these may exists at different degrees of granularity.

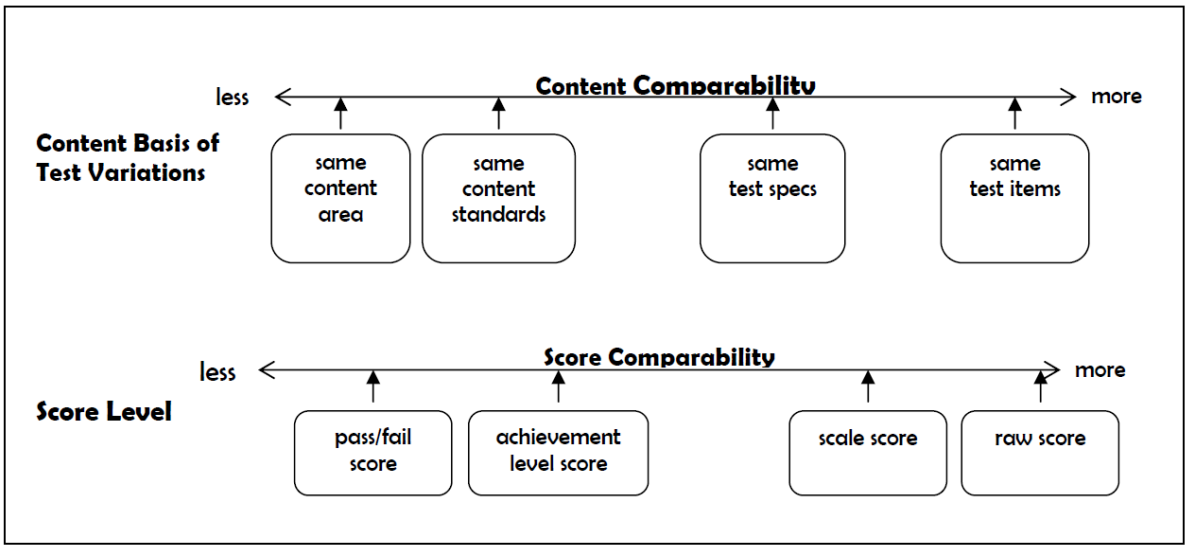

Figure 1. Comparability Continuum (Winter, 2010, p. 5)

Reiterating our earlier recommendation, comparability must be required at the level of the annual determinations. This means that evidence is provided to support the notion that the distribution of student achievement classifications in one district would be similar, all things equal, if that schools' students had participated in another district's assessment system (either pilot or non-pilot).

Evidence to Support Claims of Comparability across the Innovative and Standardized Assessment Systems

As noted above, the proposed regulations (§200.77) are focused primarily on the comparability between the pilot and non-pilot districts focused only on score comparability and not on content similarity at all. The methods for gathering evidence to support a comparability claim are not a series of analyses after the fact, but rather begin with the design of the innovative assessment and accountability pilot itself. In traditional standardized assessment programs, comparability is generally established by planning for it in the assessment system design (e.g., embedding items), evaluating the degree of comparability achieved (e.g., analyses of differential item functioning), and then, if necessary, adjusting the measurement scales to account for differences (e.g., equating). Providing evidence of comparability for the innovative assessment system will require discussion related to each of these steps, even if the methods related to each step are different. Three key questions below can guide the process of designing a pilot to produce comparability results:

1. How does the design of the innovative assessment system yield evidence to support comparability claims? Innovation and comparability appear at odds, which is why comparability must be explicitly designed for in the innovative assessment model.
2. How will the state evaluate the degree of comparability achieved across differing assessment systems (pilot/non-pilot)? What criteria will the state use to judge the results as comparable enough to support their intended purposes?

This paper does not offer additional guidance to support states in responding to question 1 above. Instead, the purpose of the current paper is to propose methods by which states could gather evidence of comparability across the innovative and standardized assessment systems. **As called for in the Notice of Proposed Rulemaking (NPRM), this document offers a broader conceptualization of comparability than what is found in the NPRM and proposes a framework along with exemplar options for evaluating comparability beyond the three options listed in the NPRM**. Additionally, the final section of this paper includes a discussion of criteria for establishing the degree of comparability necessary for supporting the intended uses.

Two Major Categories of Evidence

To evaluate the comparability of the achievement levels across an innovative assessment pilot and the statewide standardized assessment system, states should provide evidence related to each of two categories:

1. the alignment of each assessment system to the content standards, and
2. the consistency of achievement classifications across the two systems.

We recommend that states submit evidence that the innovative assessment system is aligned to the state standards and have performance expectations that are consistent with the state assessment and can therefore be employed to support the same uses in the statewide accountability system.

Evidence of Alignment

The innovative assessment system will be drawing from the same content standards as the traditional assessment system, but the way in which the standards are selected, prioritized or measured may lead to different or improved inferences about what students know and can do. Current statewide standardized assessment systems assess a non-random sample of the grade-level content standards. An innovative assessment system may have a different content sampling procedure, predicated on different curricular priorities, which could result in a different—and perhaps more valid—picture of what students know and can do. Additionally, the innovative assessment system may prioritize and measure the standards that are covered by the innovative system of assessments differently than the state standardized assessment system. This also means that the innovative assessment system may measure the state standards that are prioritized (if that is part of the design) more deeply and more thoroughly and therefore better embody the intent of the content standards than the standardized assessment system. **Therefore, we strongly recommend that evidence of alignment for the two assessment systems should come from alignment to the *content standards* rather than alignment to one another.**

There are a number of widely-used methodologies to evaluate the alignment of an assessment or assessment system to the content and depth of knowledge of the state standards (e.g., Web Alignment Tool, NCIEA's methodology for evaluating test content relative to CCSSO's Criteria for High Quality Assessments). **We recommend that the innovative assessment system, like the statewide assessment system, should be expected to provide evidence of alignment to the content standards so that the state can document the extent to which all students have learned the required content standards at the expected level of cognitive complexity.**

Evidence of Consistency in Classifications across Assessment Systems

In addition to evidence of content alignment, states participating in the demonstration authority should also be expected to provide evidence that the rigor of the performance expectations for the innovative assessment system are similar or more rigorous than those of the statewide standardized assessment system. This evidence supports the claim that not only do the assessment systems measure the same set of content standards (albeit with potentially different prioritizations), but the annual determinations reflect the same levels of achievement on those content standards as the state assessment. It is important to note that the options presented for gathering this evidence will not generally allow for equating or linking the scores scales of the two assessment systems. In other words, the goal of these analyses is to evaluate the relative rigor of the performance standards of both assessment systems, not to put the assessment results on the same score scale. To

this end, there are a number of design options available to states, each of which can be used with a variety of analytic techniques. **We recommend that each of these methods should first begin with establishing a common set of achievement level descriptors that is shared across the two assessment systems. If a state wishes to use different achievement level descriptors for the innovative assessment system, a rationale for that decision and a discussion of how those differences impact the planned comparability analyses should be provided.**

ED outlined three possible approaches for evaluating comparability between pilot and non-pilot school districts. The draft regulations invite commenters to offer additional approaches for evaluating pilot to non-pilot comparability. To summarize, the options offered in the draft regulations include:

1. Administering both assessment systems (or just the standardized assessment) to all students enrolled in pilot schools at least once per grade span,
2. Administering both assessment systems to a representative sample of students enrolled in pilot schools at least once per grade span, and
3. Embedding a set of anchor items that are the same within each grade and subject area across the pilot and non-pilot assessment systems.

The purpose of this section of the paper is to propose additional, alternative options that should be viable for evaluating comparability across pilot and non-pilot assessment systems.

In Table 1 we provide a framework to assist ED and interested states in thinking through the factors that might be considered in designing options for evaluating the consistency of the achievement classifications across assessment programs. **These factors include the sample of students included, the measures administered, and the time of administration (i.e., concurrent or non-concurrent)**. The combinations of these factors result in different design options for evaluating comparability. The design options provided in Table 1 do not represent an exhaustive list of the possibilities, but rather, they are included to demonstrate the reality that there are multiple viable ways to generate evidence of comparability of the annual determinations produced from different assessment programs. **We recommend that states be required to propose a specific approach or approaches for evaluating comparability that are tied to the specific context of the state and the proposed innovative learning system. Further, we recommend having the state-proposed approaches evaluated as part of the initial peer review process where there should NOT be a higher bar for options that differ from the three proposed by ED. Rather, any comparability proposal should be evaluated according to the inferences that the design or designs can defend.** While some designs will produce evidence of comparability that is more compelling than others, **we recommend, where feasible or where evidence may be lacking strength, that states consider multiple approaches to comparability evaluations to provide a more complete picture of the degree of comparability in the achievement levels across the two assessment systems.**

Table 1. *Design Options for Evaluating the Comparability in Rigor of Performance Standards across Innovative and Standardized Assessment Systems* (Note: The numbers in the table are tied to the multiple options listed in Appendix A).

	All Students	Some Students	No Students in Common
Both Measures	Concurrent (in past): 4. "Pre-equating"	Concurrent: 1. a) *Both assessment systems to all students in the <u>same</u> select grade levels* 2. Both assessment systems to a sample of students in select grade levels 8. Both assessment systems to a sample of students in every grade level Not Concurrent: 1. b) *Statewide assessment once per grade span in lieu of innovative assessment (i.e., state and innovative assessment in different grades)* 9. Conditioning on past performance 10. Leveraging the Student Longitudinal Data System (SLDS) for mobile students	Concurrent: 5. Random assignment of assessment system to classrooms
Some Measures	Concurrent: *3. Embedded common items across both systems* 6. Common innovative tasks 11. Common writing task 12. Short form of the state assessment		
Third Measure in Common	Concurrent: 13. Common independent assessment 14. Relationship to desired external outcome variables		Concurrent: 7. Propensity score matching
Other			Concurrent: 15. Judgmental ratings relative to Achievement Level Descriptors 16. Standard setting design

In Appendix A, we provide brief descriptions and key use considerations for each of the 16 comparability designs shown in Table 1. However, we highlight several design options below to better illustrate the qualitative differences in methods offered by framework proposed here. Further, we offer commentary regarding the feasibility and viability of each of the options presented.

1. **Both assessment systems to all students in select grade levels**[2]
   Some students, both measures, concurrent or not concurrent
   1a) Administering full assessments from both the innovative and statewide assessment system to all students enrolled in schools participating in the demonstration authority, such that at least once in any grade span (e.g., 3-5, 6-8, or 9-12) and subject for which there is an innovative assessment, a statewide assessment in the same subject would also be administered to all such students. While any method that assesses the same students on both measures is the gold standard design, there may be significant challenges associated with double-testing students. To increase feasibility, states can consider alternating the grade levels in which each content area is assessed.

   1b) All students within the pilot districts would participate in the statewide standardized assessment in lieu of the innovative assessment system once per grade span. This would allow for a direct comparison of achievement across years for the same students taking each of the assessment systems once the pilot is in its second year.

   Commentary: We and the expert panelists recognized that option 1 would offer evidence useful for evaluating comparability, but it should be noted that the two conditions—testing the same students with both assessment systems or testing students in different grades with one of the assessment systems—can support different inferences. Testing the same students with both assessment systems in the same year can support direct comparability inferences, while assessing students in different grades with the different assessment systems would provide less compelling evidence of comparability. However, the alternate grades approach is likely more feasible because it does not require double-testing.

2. **Both assessment systems to a sample of students in select grade levels**[3]
   Some students, both measures, concurrent
   Administering full assessments from both the innovative and statewide assessment system to a demographically representative sample of students and subgroups of students under section 1111(c)(2) of the Act, from among those students enrolled in schools participating in the demonstration authority, such that at least once in any grade span (e.g., 3-5, 6-8, or 9-12) and subject for which there is an innovative assessment, a statewide assessment in the same subject would also be administered in the same school year to all students included in the sample.

   Commentary: Option 2 offers considerable potential for generating strong comparability evidence, but suffers from feasibility problems because of the practical challenges of assessing only a portion of the students in the pilot districts. Further, creating an adequate sample that appropriately represents the subgroup proportions in the state can be

very challenging. In order to increase feasibility, states may consider sampling at the school-level, or alternating the grade levels in which each subject test is administered (e.g., ELA in 3rd grade and math in 4th grade).

3. **Embedded common items across both systems**[4]
   All students, some common measures, concurrent
   Including, as a significant portion of the innovative and statewide assessment systems in each required grade and subject in which both assessments are administered, common items that, at a minimum, have been previously pilot tested or field tested for use in either the statewide or innovative assessment system. This option may be limited in its feasibility if the innovative assessments are substantially different from the standardized assessment system.

   Commentary: Our expert panel was critical of any option that relied on the use of embedded items to evaluate comparability unless the two sets of assessments were designed to measure the same content in the same or very similar ways. In that case, it would be hard to see how the innovative pilot could be very innovative. In most cases, the innovative assessments will be different enough from the state assessment so that any embedded items would be so novel to the students from the "different" system that the results across conditions cannot be validly compared. For example, if the innovative system relies on extended performance tasks, but students participating in the state assessment had not experienced such tasks, not only will it be obvious that the innovative tasks are from an assessment that does not count for them, but they will likely be very disadvantaged in demonstrating their knowledge and skills if they had not experienced such formats previously.

4. **"Pre-equating"**
   Some students, both measures, concurrently administered ***in the past***
   This option would be available to those states where the innovative and traditional assessment systems existed simultaneously within the state prior to approval for a demonstration authority. For example, a state that is moving to an interim assessment option that already has a long history of use within the state. Evaluating the degree of comparability across the systems for prior years would be suitable for sustaining a comparability argument for the first one to three years of the innovative pilot. A state that takes advantage of this option would need to provide evidence that the current implementation and scoring processes of the innovative assessment system has not changed over the years, and is therefore likely to continue to produce comparable results.

   Commentary: The advantage of this option is that it does not require any sort of double-testing, so it is a very feasible option. Assuming the pilot and state assessment systems were administered together prior to the beginning of the pilot, the state could rely on the evidence gathered during the pre-pilot period to establish the comparability of the two systems. However, it is unlikely that the relationship among the two assessments will

persist more than a couple of years, especially if the innovative pilot uses a different approach to instruction and learning than the state system. That said, this model could be combined with one of the other methods after a few years to reestablish the comparability evidence.

5. Random assignment of assessment system to classrooms

No students in common, both measures, concurrent

This option would involve creating experimental design conditions where a sample of students is randomly assigned to either the innovative or standardized assessment conditions. This could be statewide across pilot and non-pilot districts, or states could do the random assignment either within pilot or within non-pilot districts. This would avoid double testing and establish randomly equivalent groups on which to compare performance. However, due to the potential novelty of the innovative assessment system, or perhaps its intentional integration with instruction (e.g., curriculum-embedded performance tasks), the feasibility of this method will be low for many innovative assessment models.

Commentary: Random assignment, as a design principle is typically regarded as the gold standard of causal inference. The quality of these inferences, however, depends on the quality of the sampling design. While the potential of such a design for yielding strong comparability evidence is high, the practicality of such a design may be low. Being able to select an appropriate sample will be the first obstacle, and an additional challenge for this application in particular, is being able to verify that administering novel assessments to students can yield valid information regarding comparability across assessment systems. We suspect this will be very hard to accomplish, so the results from this method will face considerable validity threats. This is especially challenging if the innovative assessment is a full system that is administered throughout the school year. To overcome these obstacles, this option would be most feasible when the two assessment systems are quite similar.

6. Common innovative tasks

All students, some common measures, concurrent

Instead of administering a combination of items drawn from the innovative and statewide assessments to all students (option 3), another option is to administer items from just the innovative assessment. While this option is similar to option 3 provided by ED, this option provides a distinct opportunity to involve all students in the state in the innovative assessment system in some way. For example, the innovative assessment could take the form of a common performance-based assessment that deeply measures a subset of standards and is administered to all students. Another example would be to draw from a randomized performance task bank (Way et al., 2012), which would take advantage of matrix sampling.

Commentary: While technically an embedded item approach, this approach places the innovative system at the center of the comparability inferences. This approach would work if the types of tasks found in the innovative assessment system are at least somewhat familiar to students participating in the state assessment. Further, this

approach could have the advantage of providing information and some practice regarding the innovative assessment for non-participating districts assuming the state is motivated to have the pilot system spread to new districts. Finally, like ED's third option, this option is subject to the same serious threats to comparability inferences as any other common item approach.

7. **Propensity score matching**
   No students in common, some common measures, concurrent
   Districts included in the innovative pilot are required to be demographically similar to the state as a whole. This means it should be feasible to match the pilot schools or students with non-pilot schools or students that are similar in a number of important characteristics (e.g., past performance, demographics, size, geography, etc.). The performance of the matched schools or students could be compared for the first few years of the pilot to evaluate the degree of comparability in results. However, if the innovation is intended to impact the way instruction and learning occurs in the classroom, we would expect to see this type of comparability break down after the first few years of implementation.

   Commentary: There are several approaches that do not rely on common students or perhaps not even common items. As discussed previously, randomly assigning students to an assessment approach has the potential of supporting causal inferences, but requires overcoming some significant hurdles. There are multiple approaches that try to overcome the lack of random assignment that rely on establishing groups matched on key variables such as prior scores and important demographic characteristics of students. Propensity score matching describes a class of methods that uses sophisticated statistical procedures to create the groups so that performance of the two "pseudo-equivalent" groups can be compared on the same or even different assessments. Other than the statistical sophistication needed, these approaches are highly feasible because they do not rely on any double-testing. However, the quality of inferences is dependent on the quality of the matching variables available to use. Further, since common prior scores is a key matching variable, the use of this approach will become less useful within a few years of the beginning of the pilot because the common prior scores would not be viable once the innovation can be presumed to begin to affect the key outcome variable, achievement.

Criteria for Comparability Evidence: How Comparable is Comparable Enough?

How comparable is comparable enough? **We recommend AGAINST setting a standard criterion, or comparability "bar", because the intended uses and the contextual factors surrounding the evaluation of comparability are critical**. However, it is worthwhile to consider what might be reasonable to expect for the amount of variability in proficiency classifications across the two assessment programs. We argue that a reasonable upper bound for comparability across pilot and non-pilot systems is the degree to which comparability is achieved across forms, modes, and years of administration for the statewide, standardized assessment system. This is akin to the axiom that a test cannot correlate any more with another test than it

does with itself (i.e., its reliability). The literature is clear that there are significant effects associated with mode of administration (including paper/computer and across devices), accommodations, and forms across years. Due to the precedence for this type of variation within our current assessment systems, it may be reasonable to expect that the variability across the innovative assessment pilot and non-pilot would be at least as large as levels we see with current state testing programs. Again, when we refer to variability across assessment programs, we are not expecting that pilot and non-pilot districts exhibit the same levels of achievement—because districts are not randomly assigned to the pilot, the systems have potentially different emphases in measuring learning targets, and we hope that the innovation itself will improve achievement—but that the systematic effects of the assessment system on the achievement estimates likely will be larger than the effects of form, mode, device, and year that we see in our current assessment systems.

The unit of analysis for evaluating comparability must be at the school and subgroup levels, given the school accountability purposes of the assessment results. However, because the subgroups may involve small sample sizes, the tolerance for comparability needs to be greater for the subgroup analyses compared to the school level analyses. If school or subgroup differences across systems are detected, the state should evaluate the practical implications of those differences for decision making within the accountability system. Figure 2 presents a series of questions that could determine whether or not the levels of comparability seen are appropriate for the intended purposes:

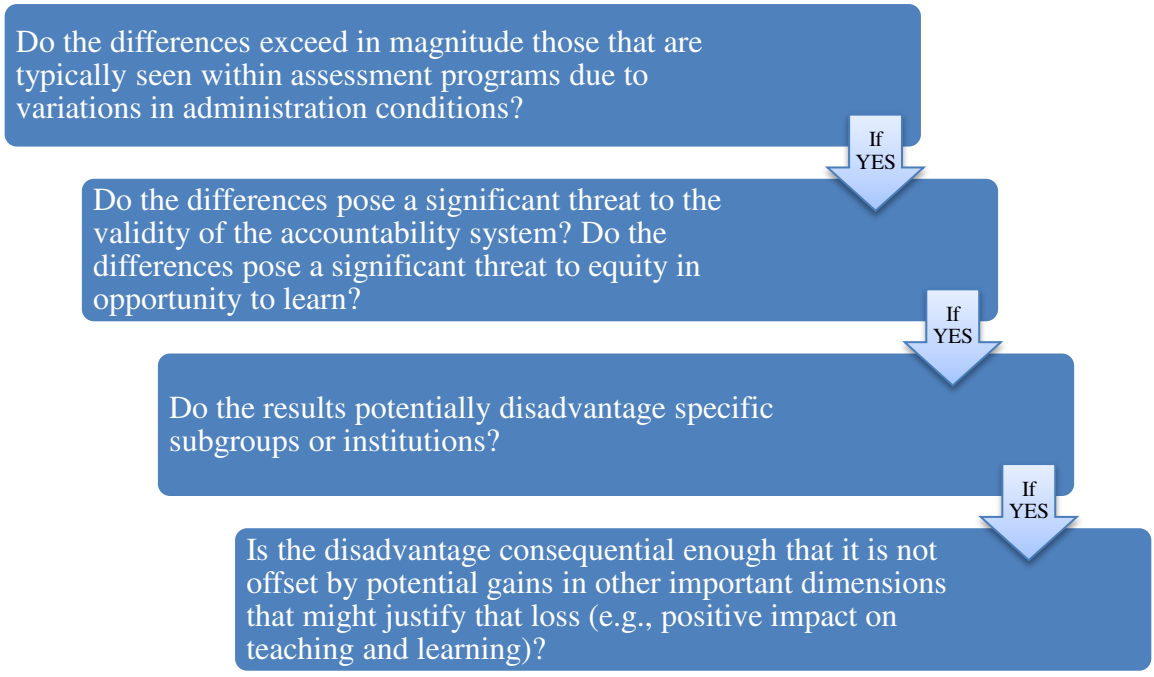

Figure 2. Decision Tree for Determining Degree of Comparability Achieved

If the answer to any of these questions is no, the assessment systems can be considered comparable enough to support their intended uses for the duration of the pilot. However, in the case where all of the answers above are "yes," additional steps will need to be taken to improve the comparability of the achievement classifications to support their use

in the statewide accountability system. To do so, the performance standards on either one of the assessment systems can be shifted or adjusted (such as equipercentile linking) to produce useable results for the duration of the demonstration authority, after which, standards can be re-set. It is worth noting that, if states are using a model that is not qualitatively different from the current state assessment system, scale score equating may be possible in some cases. If this is the case, both the scores and the proficiency classifications resulting from the two assessment systems will be comparable, and there is no need for criteria.

The first few years of the pilot are arguably the most important for demonstrating that results across pilot and non-pilot districts are comparable enough. **As the innovation reaches critical mass and spreads across the state, comparability across the two assessment systems becomes less important than the comparability of results among districts within the innovative system of assessments. Additionally, if the evidence for comparability across the two systems of assessment is strong, comparability need not be re-evaluated every year. Once it has been established, the state should provide evidence that the processes and procedures in place are sufficient for replicating the program across years.**

References

DePascale, C., Dadey, N. & Lyons, S. (2016). *Score comparability across computerized assessment delivery devices.* Washington, DC: The Council of Chief State School Officers.

Gong, B., & DePascale, C. (2013). *Different but the same: Assessment "comparability" in the era of the Common Core State Standards.* Washington, DC: The Council of Chief State School Officers.

Lyons, S., Evans, C., Marion, S.F. (2016, April). *Comparability in balanced assessment systems for state accountability*. Paper presented as part of symposium entitled "Advances in Balanced Assessment Systems: Conceptual framework, informational analysis, application to accountability" at the annual meeting of the National Council on Measurement in Education, Washington, D.C.

Lyons, S., Marion, S.F., Pace, L., & Williams, M. (2016). *Addressing Accountability Issues including Comparability in the Design and Implementation of an Innovative Assessment and Accountability System.* www.Knowledgeworks.org and www.nciea.org.

Way, W., Murphy, D., Powers, S., & Keng, L. (2012, April). *The case for performance-based tasks without equating.* Paper presented at the National Council on Measurement in Education, Vancouver, Canada.

Winter, P. (2010). Comparability and test variations. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 1-11). Washington, DC: Council of Chief State School Officers.

Appendix A: List of Additional Comparability Design Options

1. **Both assessment systems to all students in select grade levels**[5]
   Some students, both measures, concurrent or not concurrent
   1a) Administering full assessments from both the innovative and statewide assessment system to all students enrolled in schools participating in the demonstration authority, such that at least once in any grade span (e.g., 3-5, 6-8, or 9-12) and subject for which there is an innovative assessment, a statewide assessment in the same subject would also be administered to all such students. The methodology to support this design would involve looking at the differences in distributions of performance levels across the two assessment systems for a cohort of students at a single point in time. While any method that assesses the same students on both measures is the gold standard design, there may be significant challenges associated with double-testing students. To increase feasibility, states can consider alternating the grade levels in which each contest area is assessed.

   1b) All students within the pilot districts would participate in the statewide standardized assessment in lieu of the innovative assessment system once per grade span. This would allow for a direct comparison of achievement across years for the same students across taking each of the assessment systems once the pilot is in its second year. This method of gathering comparability evidence would be sustainable throughout the entirety of the innovative pilot.

2. **Both assessment systems to a sample of students in select grade levels**[6]
   Some students, both measures, concurrent
   Administering full assessments from both the innovative and statewide assessment system to a demographically representative sample of students and subgroups of students under section 1111(c)(2) of the Act, from among those students enrolled in schools participating in the demonstration authority, such that at least once in any grade span (e.g., 3-5, 6-8, or 9-12) and subject for which there is an innovative assessment, a statewide assessment in the same subject would also be administered in the same school year to all students included in the sample. The strength of the evidence generated from this design is strong, but may not be feasible due to the requirement of double-testing. In order to increase feasibility, states may consider sampling at the school-level, or alternating the grade levels in which each subject test is administered (e.g., ELA in grade and math in 4th grade).

[5] Option 1 offered by ED

[6] Option 2 offered by ED

3. **Embedded common items across both systems**[7]
   All students, some common measures, concurrent
   Including, as a significant portion of the innovative and statewide assessment systems in each required grade and subject in which both assessments are administered, common items that, at a minimum, have been previously pilot tested or field tested for use in either the statewide or innovative assessment system. This option may be limited in its feasibility if the innovative assessments are substantially different from the standardized assessment system. Our expert panel was very critical of any option that relied on the use of embedded items to evaluate comparability unless the two sets of assessments were designed to measure the same content in the same or very similar ways. In that case, it would be hard to see how the innovative pilot could be very innovative. In most cases, the innovative assessments will be different enough from the state assessment so that any embedded items would be so novel to the students from the "different" system that the results across conditions cannot be validly compared.

4. **"Pre-equating"**
   *Some students, both measures, concurrently administered **in the past***
   This option would be available to those states where the innovative and traditional assessment systems exist simultaneously within the state prior to approval for a demonstration authority. For example, a state that is moving to an interim assessment option that already has a long history of use within the state. Evaluating the degree of comparability across the systems for prior years would be suitable for sustaining a comparability argument for the first one to three years of the innovative pilot. A state that takes advantage of this option would need to provide evidence that the current implementation and scoring processes of the innovative assessment system has not changed over the years, and is therefore likely to continue to produce comparable results.

5. **Random assignment of assessment system to classrooms**
   No students in common, both measures, concurrent
   This option would involve creating experimental design conditions where a sample of students is randomly assigned to either the innovative or standardized assessment conditions. This could be statewide across pilot and non-pilot districts, or states could do the random assignment either within pilot or within non-pilot districts. This would avoid double testing and establish randomly equivalent groups on which to compare performance. However, due to the potential novelty of the innovative assessment system, or perhaps its intentional integration with instruction (e.g., curriculum-embedded performance tasks), the feasibility of this method will be low for many innovative assessment models. Additionally, because there are no students in common, the evidence of comparability is not as strong as the common-student designs.

[7] Option 3 offered by ED

6. **Common innovative tasks**
   All students, some common measures, concurrent
   Instead of administering a combination of items drawn from the innovative and statewide assessments to all students (option 3), another option is to administer items from just the innovative assessment. While this option is similar to option 3 provided by ED, this option provides a distinct opportunity to involve all students in the state in the innovative assessment system in some way. For example, the innovative assessment could take the form of a common performance-based assessment that deeply measures a subset of standards and is administered to all students. Another example would be to draw from a randomized performance task bank (Way et al., 2012), which would take advantage of matrix sampling.

7. **Propensity score matching**
   No students in common, some common measures, concurrent
   Districts included in the innovative pilot are required to be demographically similar to the state as a whole. This means it should be feasible to match the pilot schools or students with non-pilot schools or students that are similar in a number of important characteristics (e.g., past performance, demographics, size, geography, etc.). The performance of the matched schools or students could be compared for the first few years of the pilot to evaluate the degree of comparability in results. However, if the innovation is intended to impact the way instruction and learning occurs in the classroom, we would expect to see this type of comparability break down after the first few years of implementation.

8. **Both assessment systems to a sample of students in every grade level**
   Some students, both measures, concurrent
   Representative or random sample of intact classrooms participate in both assessment systems. The administration would be the same for all of the students within that class. To improve the strength of the evidence, schools could counterbalance the timing of the administration of those assessments within the intact classrooms. Additionally, the sampling could be done by content area so that the double testing is controlled (i.e., you are not taking the whole battery of assessments in any given elementary classroom).

9. **Conditioning on past performance**
   Some students, both measures, not concurrent
   All public schools have over a decade of data on past performance that can be leveraged to provide an indication of the degree of comparability in assessment system results for the first 1 to 2 years of the innovative assessment pilot. This option takes advantage of the notion that true organizational change will likely require more than just one year of pilot implementation. Therefore, past performance for schools can provide a reasonable expectation of current performance for the first couple years of the innovative assessment system. There are a number of analytic methods that could support this design including creating matched groups and running a regression discontinuity analysis.

10. **Leveraging the SLDS for transient/mobile students**

Some students, both measures, not concurrent

Once the pilot grows to assess students in the thousands, it can be expected that there would be enough students moving in and out of the pilot districts each year to provide one source of evidence of comparability in assessment system results. Examining the performance of these students across the two assessment systems across adjacent years will provide substantial insight into the degree of comparability of the results throughout the duration of the innovative assessment system pilot. Though students who are mobile are not likely to be representative of the population in terms of performance and other demographic factors, running these analysis requires relatively little burden because the design is naturally occurring and does not require double testing. While stronger methods for evaluating comparability may be necessary for the first year or two of the pilot, this method may be a sustainable option once comparability has already been established and the number of districts participating in the pilot increases.

11. Common writing task

All students, some common measures, concurrent

Similar to the common innovative task approach discussed above, the common writing task approach will be a relatively non-intrusive approach for evaluating comparability of pilot and non-pilot districts. This approach should be applied only to states that included a stand-alone or essentially stand-alone writing task as part of the statewide assessment. In this case, students in pilot districts would complete one of the major writing tasks included on the statewide, standardized assessment in each grade or in a sample of grades so that the writing performance of the two sets of students could be directly compared. This approach is essentially the inverse of the "common innovative task" approach discussed above and is also limited to writing alone, but could provide another point of comparability.

12. Short form of the state assessment

All students, some common measures, concurrent

This method is distinct from option 3 in that all students participating in the innovative pilot take a short-form version of the state assessment that is intended to contribute to their achievement score. Because all students in the state are administered at least some common items, comparability across the two programs can be evaluated. Additionally, because the short-form assessment is contributing to the scores generated from the innovative assessment system, the annual determinations across the two assessment systems will likely be more consistent than had these items not been counted.

13. Common independent assessment

All or some students, some common measures, concurrent

If all or a large sample students in the both the pilot and non-pilot districts are already taking a third test (e.g., large-scale interim or high school assessment), the scores from that third test can be used to provide evidence on comparability – an "indirect link." This design would allow for the comparison of the distributions of achievement using a number of analytic techniques (e.g., equipercentile, regression, matching, etc.). This option would produce strong evidence of comparability in rigor if the third test is also demonstrated to be aligned to the same learning targets as both the innovative and

standardized assessment system. Additionally, this option would be highly feasible in a state that already has a large number of students participating in an additional assessment program.

14. Relationship to desired external outcome variables

All or some students, some common measures, concurrent

This design involves using a third measure or indicator to show that student performance on the innovative assessment is comparable or better than the state test when it comes to predicting desired outcomes (grades in the following year, performance in college courses, performance on the ACT, etc.). This evidence would support the claim that the assessments are comparable enough to support the intended uses and goals in that to be deemed proficient by the innovative assessment system is consistent with—or even better than—the state test when it comes to predicting the intended outcomes.

15. Judgmental ratings relative to ALDs

Some students in common, no common measures, concurrent

This design would involve having content experts evaluate bodies of work produced by the two assessment systems in order to make judgments about the achievement level to which each body of work best matches. The goal would be to recover the achievement classifications from the assessment. This method rests on the notion of common achievement level descriptors across the two assessment systems. For multiple choice assessments, the bodies of work can include qualitative descriptions of the tasks and information about how the student responded. A key design feature would be to use the same panels of participants to evaluate the two sets of work. This method provides evidence that both assessments can provide for accurate interpretations about what students know and can do using the same achievement level descriptors. An added benefit of this method is that it adds little additional burden to students or schools.

16. Standard setting design

No students in common, no common measures, concurrent

This method would ask that states provide evidence that the standard setting process was developed and implemented specifically to ensure comparability in the performance designations assigned across the two assessments. This could be achieved by:

- Using the same panels of participants
- Using the same performance level descriptors and/or threshold descriptors as were used for the state assessment.
- Incorporating exemplars of student performance at each level based on the state test within the standard setting process (using an item mapping approach).

As with the judgmental ratings design option, this option additionally does not add an additional burden to students or schools.