Considering ESSA's Provisions Allowing Local Options for Assessments

Brian Gong and **Nathan Dadey**
Center for Assessment

Presentation at the
CCSSO TILSA SCASS
October 25, 2016

Center for Assessment

Overview of presentation

- Review ESSA provisions that involve local options for assessments
- Provide framework for thinking about local options
- Discuss high school assessments; Discuss innovative assessment pilot
- Provide guidance on what technical criteria and processes should be used to approve local options for assessments
- Provide guidance on how the state could define and work to create appropriate comparability among state and local assessments for the purpose of making comparisons across districts

ESSA provisions that involve "localized" assessments

Center for Assessment

ESSA "local" assessment provisions

1. Requirement that a state consider a district request to use a "nationally recognized" assessment instead of the state high school assessment
2. Option that the state may choose to use "multiple interim assessments instead of a single state assessment to produce a summative score"
3. Option that the state may apply for a pilot of "innovative assessments" that would replace the state summative assessment for reporting and accountability purposes

ESSA: nationally recognized high school assessment

- ESSA has a provision that a district/LEA may request state approval to use a "locally selected, nationally recognized high school academic assessment" in lieu of the state high school assessment
- State must consider district request
- If an assessment is approved for one district, then any district in state may use that assessment without additional state approval process
- Districts must follow provisions for local notification (e.g., parents) before and during use
- Assessments must pass Peer Review

ESSA interim assessment

- ESSA (*Every Student Succeeds Act*) allows state to consider using "a single summative or multiple interim assessments" to comply with assessment and accountability requirements of ESSA
- Single interim assessment program for the state (e.g., same set of interims in each grade)—not multiple interim assessments selected by districts
- Interim assessment would need to pass Peer Review

ESSA: innovative assessment pilot

- ESSA permits USED to provide demonstration authority to up to seven SEA to pilot an innovative assessment and use it for accountability and reporting purposes before scaling such an assessment statewide
- SEA may propose the innovation, such as performance-based assessments, assessments supporting a competency-based education model, etc.
- SEA must demonstrate quality of the innovative assessments, including comparability across districts and time

A framework for considering ESSA local assessment options

Center for Assessment

Assessments as evidence for claims for uses

- Intended Purpose, Use, and Consequences

School accountability for student learning

- Construct, Reporting, and Claims (Interpretation)

The student is college/career ready in math (Level 3).

- Intended Evidence and Design Conditions

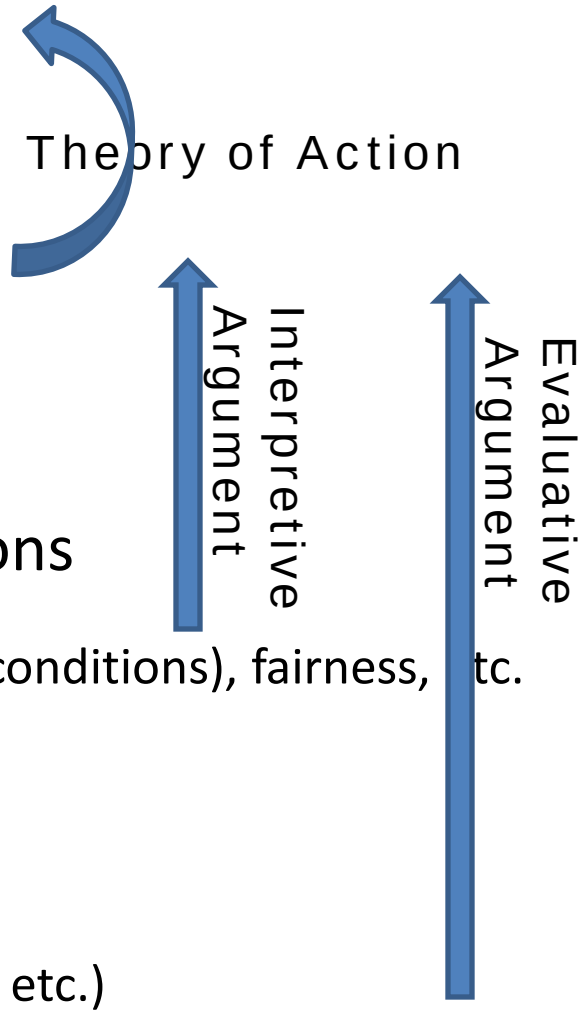
Test measures CCR math knowledge & skills at Level 3

- Quality of Evidence

Alignment, no cheating, etc.

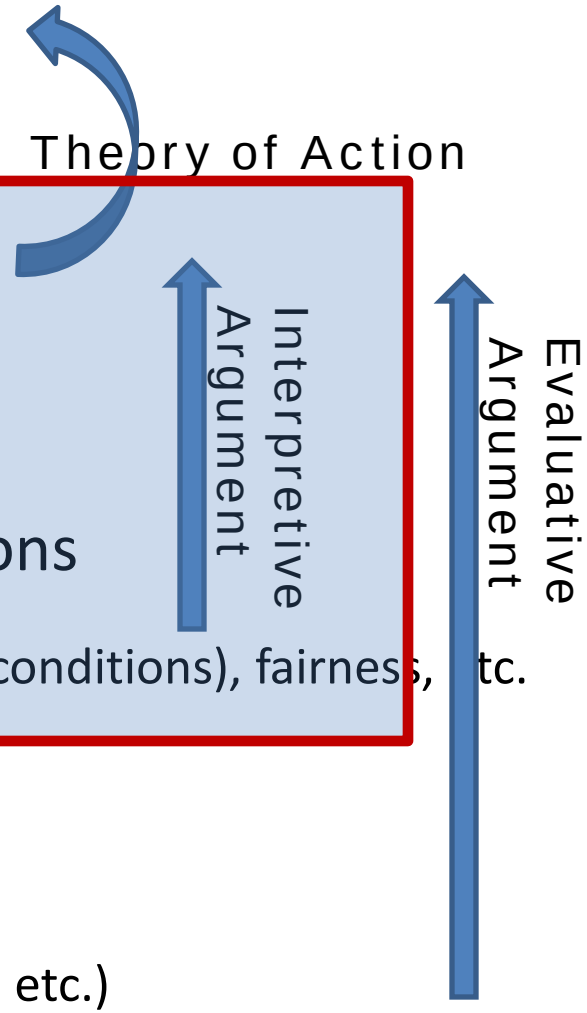
Assessments as evidence for claims for uses

- Intended Purpose, Use, and Consequences
- Construct, Reporting, and Claims (Interpretation)
  - Specificity & Generalization
- Intended Evidence and Design Conditions
  - Content, sampling, comparability (replication conditions), fairness, etc.
- Quality of Evidence and Argument
  - Implementation fidelity
  - Validity/reliability evidence (internal, external, etc.)



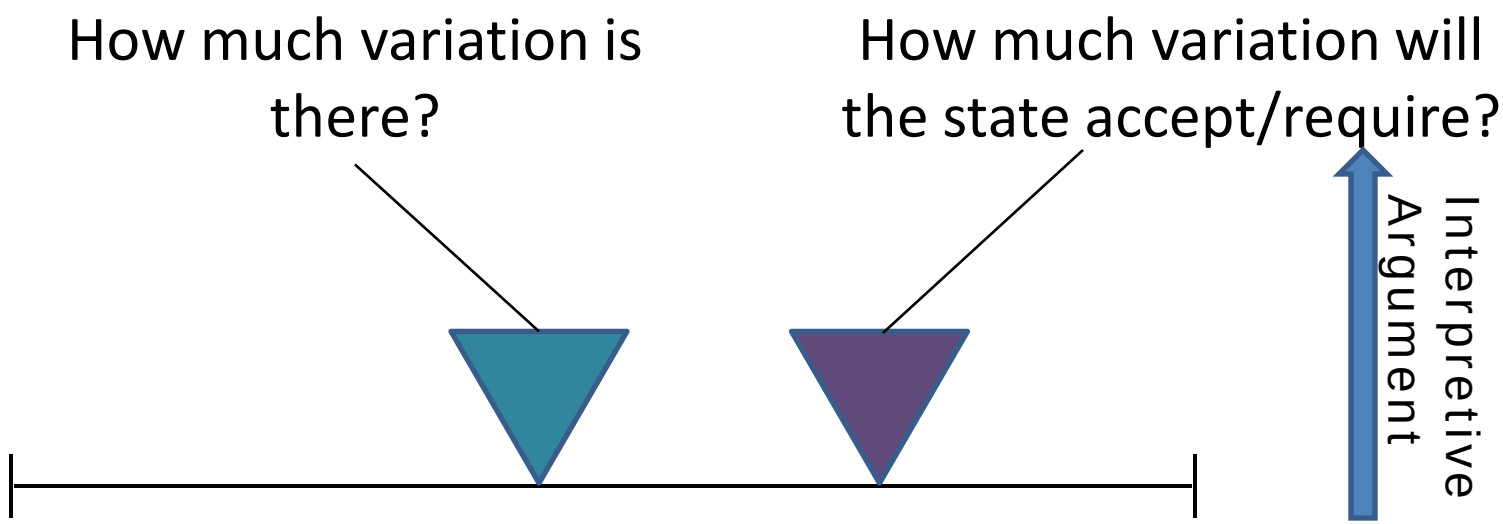
Assessments as evidence for claims for uses

- Intended Purpose, Use, and Consequences
- Construct, Reporting, and Claims (Interpretation)
  - Specificity & Generalization
- Intended Evidence and Design Conditions
  - Content, sampling, comparability (replication conditions), fairness, etc.
- Quality of Evidence and Argument
  - Implementation fidelity
  - Validity/reliability evidence (internal, external, etc.)



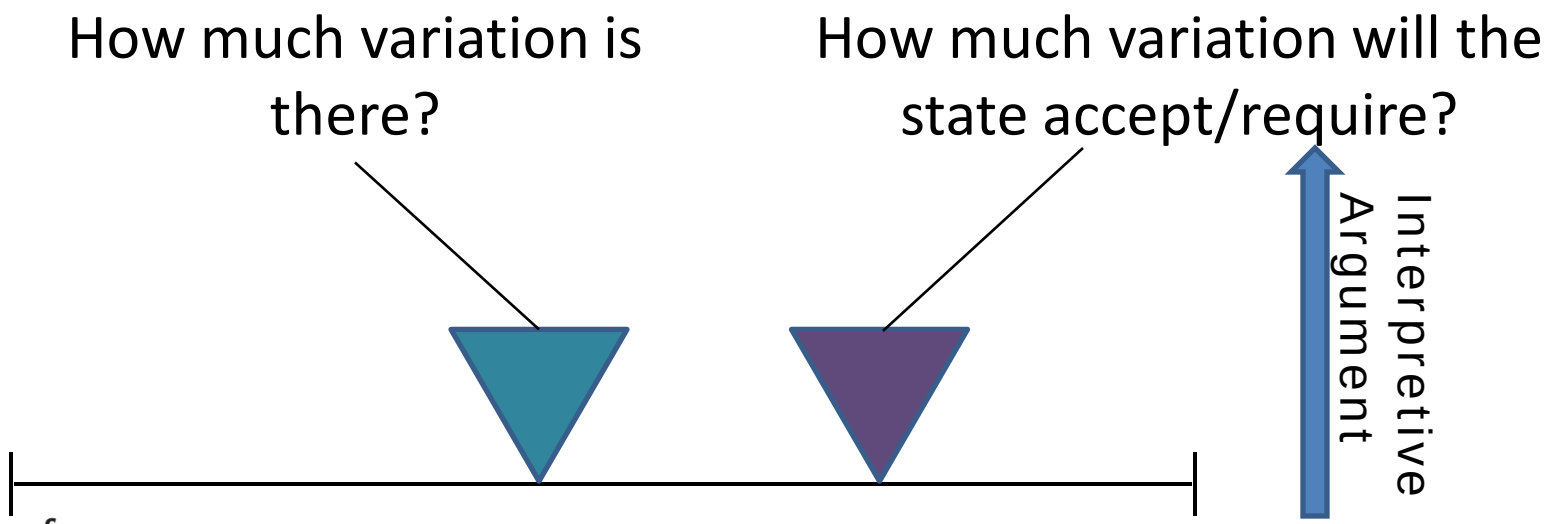
Key questions – a continuum

- **What is the same? What varies? How? Why?**
- **How does that affect the evidence needed?**
- **What is acceptable? Why?**



Key questions and guidance

- Provide guidance on what technical criteria and processes should be used to approve local options for assessments
- Provide guidance on how the state could define and work to create appropriate comparability among state and local assessments for the purpose of **making comparisons across districts**



Locally selected, nationally recognized high school assessments

Center for Assessment

ESSA: nationally recognized high school assessment

- ESSA has a provision that a district/LEA may request state approval to use a "locally selected, nationally recognized high school academic assessment" in lieu of the state high school assessment
- State must consider district request
- If an assessment is approved for one district, then any district in state may use that assessment without additional state approval process
- Districts must follow provisions for local notification (e.g., parents) before and during use
- Assessments must pass Peer Review

ESSA specific provisions

- "Nationally recognized high school academic assessment" means an assessment of high school students' knowledge and skills that is administered in multiple States and is recognized by institutions of higher education in those or other States for the purposes of entrance or placement into courses in postsecondary education or training programs." (Proposed assessment rule, 4/19/16; final regulation not yet posted) – **ACT, SAT, Accuplacer, Smarter Balanced, PARCC, AP, IB, etc. (training programs)?**
- Is aligned with the challenging State academic standards; addresses the depth and breadth of those standards; is equivalent to or more rigorous than the statewide assessments
- Produces valid and reliable data on student academic achievement for all high school students and subgroups that are comparable to those produced by the statewide assessment; are expressed in terms consistent with the State's academic achievement standards; provide for differentiation among schools within the state for accountability
- Has appropriate accommodations
- Submit Peer Review evidence to Secretary
- Applies to all high school students in the LEA except AA-AAS
- LEA provisions (notify parents, etc.) – will state monitor?

This ESSA provision is unprecedented

- **Adoption of College Entrance Exam as statewide test**
  - Several states have used a college entrance exam (i.e., ACT, SAT, augmented/unaugmented) for the state's high school assessment,
  - but almost always had census administration (almost all students were to take a **single** assessment that was reported and used in accountability).
  - Some states allowed a menu of assessments for high school graduation qualification (e.g., state test *or* AP) with established cutscores; but students almost always took the state test, so there was a **single** assessment system for **school** accountability, with a **multiple** assessment system for **student** accountability; variations included using multiple tests in school accountability as optional (e.g., via "bonus points")
- **Adoption of multiple tests, none of which is administered to all students**
  - This provision would result in ***multiple*** high school assessments being administered during the year, so **students would have different test scores** by district.
  - The assessments would have **multiple contractors/administrations** and would need to be **combined for reporting and** use in **accountability**.
  - States should consider implications of a **multiple-disjunctive (non-overlapping) assessment system** for assessment, accountability, credibility, capacity, etc.

State preparation

- We anticipate every state will receive a request from at least one district. Legally, the state has to respond.
- States should:
  - **specify** the technical and operational requirements for an acceptable **multiple-**assessment system to be used for school accountability
  - **anticipate** what it takes to implement,
  - **decide** on appropriate state roles and responsibilities;
  - **establish** application and review processes and criteria to determine whether to approve a district request;
  - plan how to **monitor** implementation
  - put in place a **communication plan**

Addressed in this presentation

- Is the assessment acceptable quality?
- How will the assessment work with the state assessment (e.g., comparability)?
- How will the assessments work with other system components (e.g., accountability)?
- What does the process need to work well (within state, cross-state)?
- What are the state's views/values regarding multiple assessments, especially a college entrance exam?

Specify requirements – single assessment

Technical for Accountability	Fairness
Alignment, including rigor Range Security Data privacy, data ownership Growth Testing time Grade 3-11 cutscore coherence	Fit with state demographics Accommodations Retesting Test preparation support
Relationship with Instruction	**Credibility and Student motivation**
Informing instruction Encouraging high level instruction	

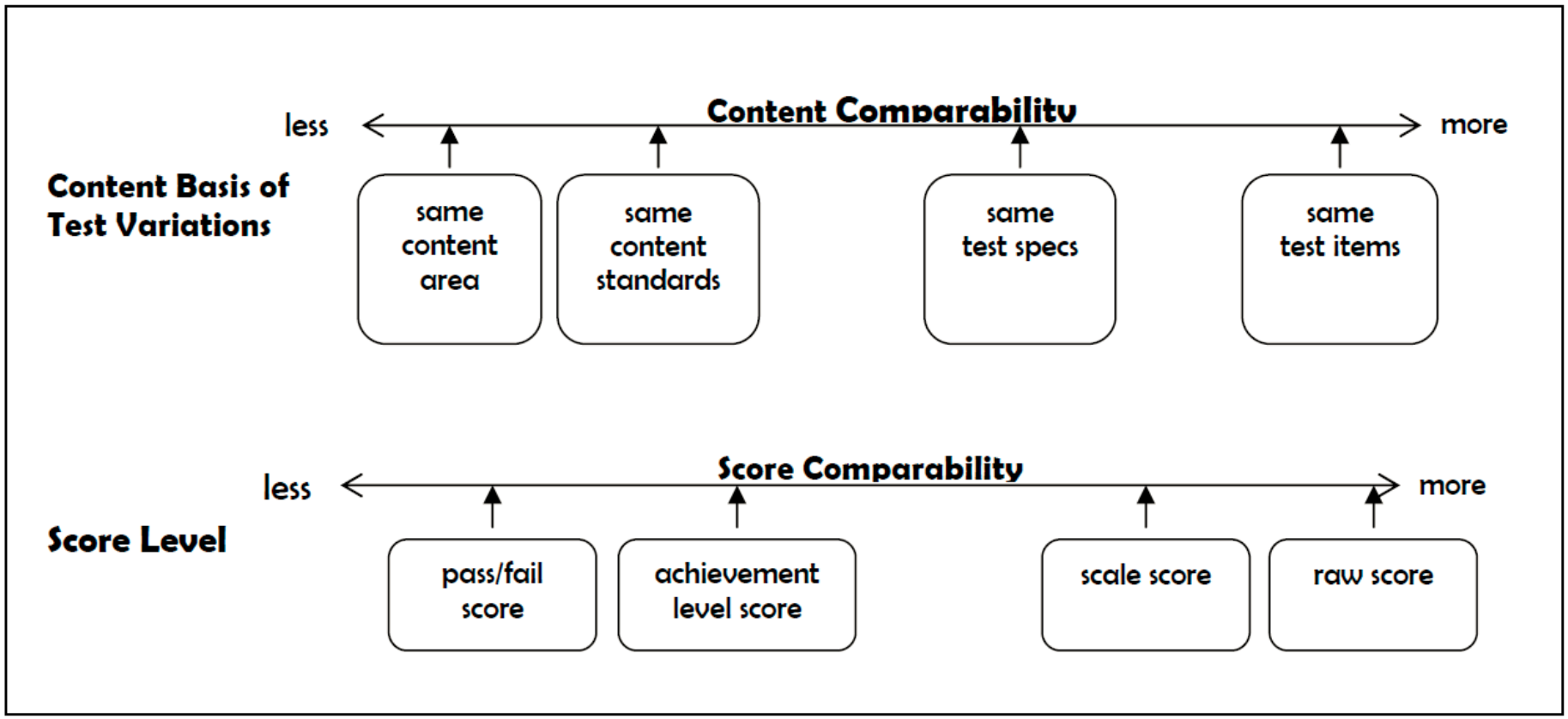
For more detail, see Martineau, Gong, & Zurkowski. (2016). Preparing for ESSA's "Nationally Recognized High School Assessment" Provision. Presentation at the CCSSO National Conference on Student Assessment. June 21, 2016. Philadelphia, PA.
https://ccsso.confex.com/ccsso/2016/webprogram/Session4740.html

Specify requirements – multiple assessments

- Issue is not only “Is assessment high quality?” but “Is it comparable and compatible enough?”

Technical Requirements - comparable	Operational Requirements - compatible
Content: Assessment is aligned to state assessment blueprints and specifications **Scores & ALDs**: Scale scores, ALDs, and cutscores are comparable; allow aggregation for reporting (e.g., avg. scale score, percent proficient), calculation of growth scores, use in making accountability determinations **Accommodations:** Must be equivalent to those allowed on state assessment **Scaling/Equating**: Sound maintenance of score meaning, comparable to state assessment; also, changes in tested populations must not compromise state’s ability to scale and equate state test over time **Precision**: Must be possible to provide metrics of assessment precision for data combined across tests **Reports:** Reports provide at least same information as state assessments (e.g., scoring; student) **Updates/changes**: stay in synch with state tests	**Test Window**: Test windows must be acceptably similar to state assessment test windows to support claims of fairness and comparability **Test Mode/Platform/Training**: Adequate training must be provided for each assessment **Administration Support**: Adequate support must be provided during test administration **Data Processing**: Data must be provided to state in detail, in format, and on time for state processing. **Data Security/Ownership**: Data must be (in)accessible in compliance with state policy for state assessments **Responsive**: Vendor should handle any issues with its testing as responsively as would be done with state testing (e.g., rescoring request). **Transparency/Communication Support**: Vendor should provide at least as much support in materials, expert labor for critical issues as is expected of vendor for state assessment (e.g., released items; anomaly)

Two dimensions of comparability



Comparability Continuum (Winter, 2010, p. 5)

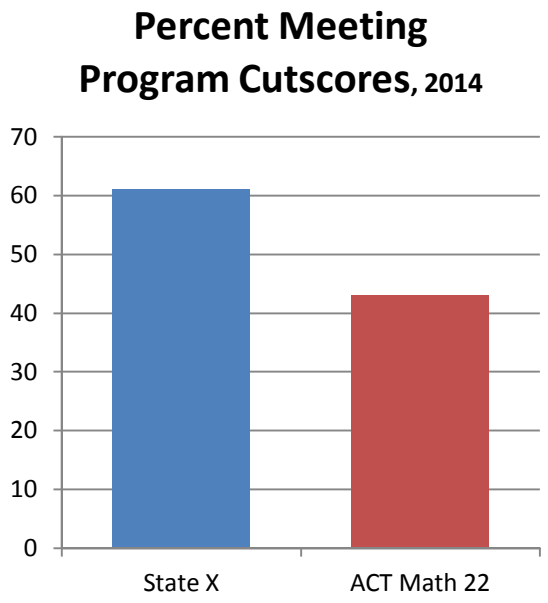
One specific example: Cutscore coherence

- State must create specifications for **"good enough" evidence** that interpretations of performance are **comparable** enough for intended purposes
  - Evaluation of soundness of anchor ALDs, blueprints, standard setting, reported scores (including subscores)
  - Sound content analysis of comparability of Achievement Level Descriptors, Test Blueprints, and reported scores
  - Sound linking study, including considerations of populations, assessment conditions, scale properties for all cutscores; linking study design and execution; credibility
    - Comp'-rable = means the same
    - Com-par'-able = has a systematic (e.g., statistical) relationship
  - Consideration of systemic coherence over grades, uses, time

Example: Systemic coherence over grades

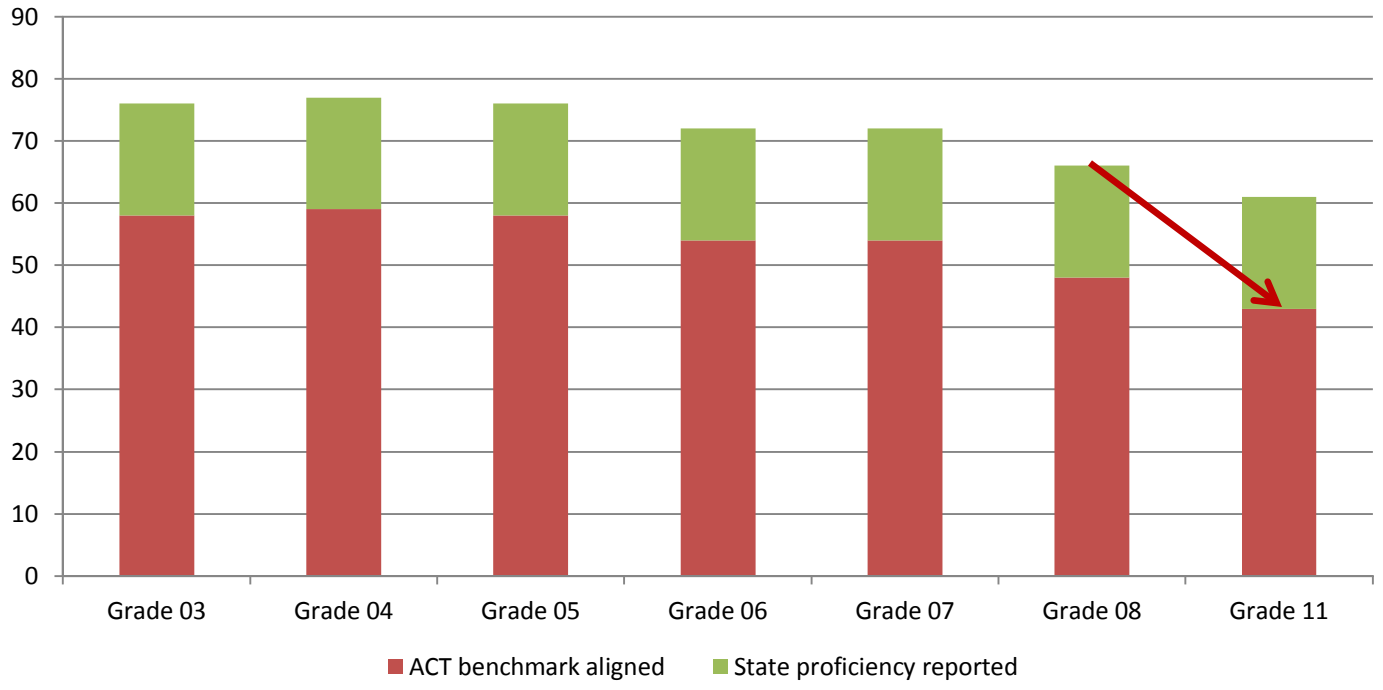
- What ACT score to use in the state for "CCR cutscore? (composite ACT score, 25%ile admitted for IHE. AFQT = Armed Forces Qualifying Test, a subset of the ASVAB. An AFQT score of 31 is required for Army entrance; varies by service branch and area of interest) (Note: ELA may be more difficult; SAT has not produced an official way to combine Reading and Writing.)
- Large difference between percentages of students in state meeting state proficiency cut and ACT national mathematics benchmark in high school (approx. 84% of graduating students had an ACT score)

State X's IHE/Career	Min. ACT score
State X selective 4-yr	22
Other State X 4-yr	20
AFQT (31-49)	15-16



Example: Systemic coherence - 2

- Illustrative adjustments to grades 3-11 achievement to correspond with ACT “percent meeting benchmark”

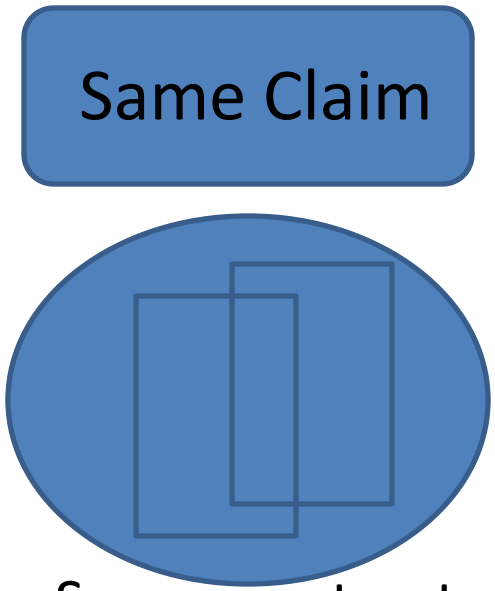


Other systemic implications

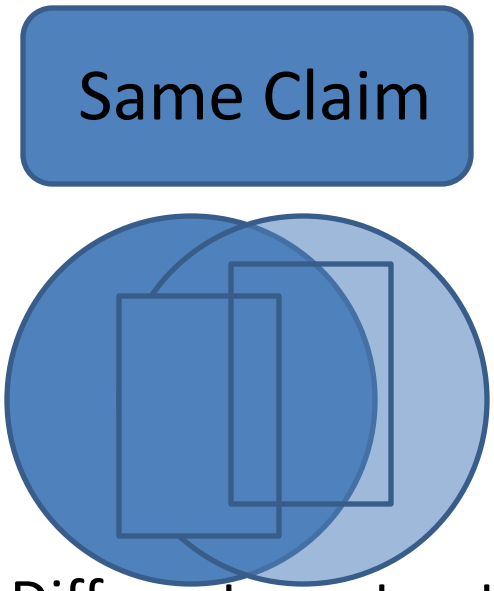
- Consider implications of multiple high school assessments for rest of educational system:
  - Grades 3-8 ALDs and cutscores
  - Accountability system
  - "Career readiness" efforts
  - School/district support
  - Public support
  - Assessment contract planning (e.g., volume pricing)
  - State department of education capacity (e.g., management support and communication)

Claims, evidence, and comparability

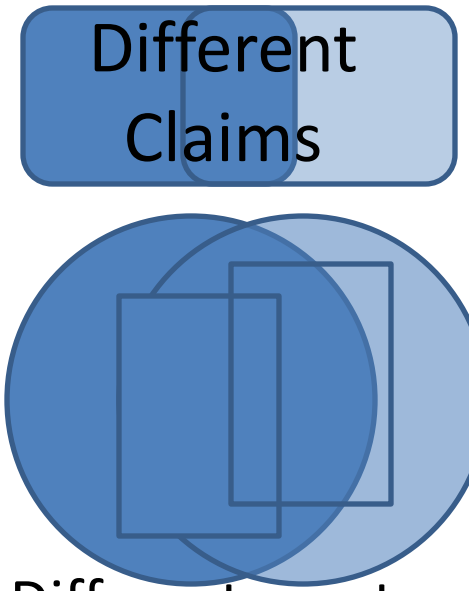
- How much difference is there between the NRHS assessment and the state assessment? Is it at the Claim level or Intended Evidence Level?



Same construct, different evidence



Different construct, different evidence



Different construct, different evidence

“Good enough” criteria for comparability

- How much alignment is “close enough”?
- How much “concordance table” is “close enough”?

- How much variation is there in the state assessment?
  - E.g., items, forms, administration conditions, persons, trend, etc.
- How much difference does it make that variations would be systematic by district, and irregular by time?

Some very specific questions – for each state

- What if a state has end-of course tests?
- What if a state's high school test is in grade 10 (since ACT, SAT are generally viewed as assessing grade 11 or grade 12 achievement)?
- Could a district use a math test from one nationally recognized test and an ELA test from a different test?
- What if the evidence of appropriateness is quite variable across possible adopting districts?
- What if a district wanted to switch assessments year after year—how would that affect accountability? Other ways this could be used to "game" assessment or accountability?
- Does the first district need to wait until the assessment is approved through federal Peer Review before it can administer the test?
- How to forecast/manage participation, volume, and cost?
- Etc.

Some questions – for all states

- What if one state approves a HS test; are subsequent states under pressure to also approve the same test?
  - Should states using a multi-state test consider coordinating or communicating with each other about this review?
  - Should any/all states consider communicating with each other about their review criteria? (See CCSSO work group)
- How might this provision interact with Demonstration Authority states?
- Etc.

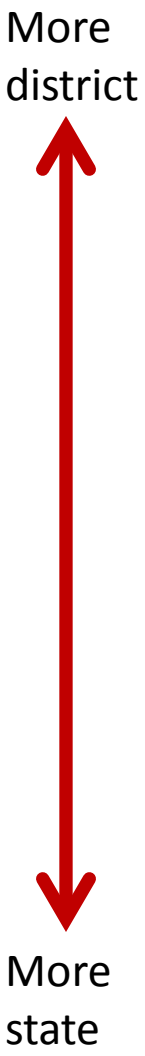
Some key considerations for Review

- State's requirements for (multiple) assessments
  - What does state require **technically** and **operationally** for a credible school accountability system?
    - How comparable must assessment results be, given intended uses?
      - What is acceptable evidence?
    - How much control must state have to responsibly certify use of assessment results for accountability and carry out sustainable program?
- State's and others' responsibilities and resources
  - What would it take for the state to be assured the technical and operational requirements have been met?
    - What is a reasonable process? This is *procurement.*
  - Who is responsible for doing which parts, when? States will probably need a *scope of work* and a *contract* as for other state assessments (or at least an MOU with districts).

Decisions for Assessment Review Process

- What are technical and operational criteria? What is acceptable evidence? How to evaluate practical impacts of dealing with multiple assessments?
- When will reviews be available? (Starting what year?) When will reviews be conducted (annually)?
- How many likely applications to review?
- What will review process be? How independent? How elaborate? Advisory? Who will generate the specific review protocols/forms? What will the appeal process be? How is evaluation review related to state's Peer Review?
- Who will reviewers be? What qualifications? How recruited? Paid?
- Who generates information for adoption/evaluation review?
- Who pays for what? State or district? Who negotiates with assessment provider?
- How will state formulate policies and communicate? How will state update policies?
- How will state ensure needed capacity for dealing with additional complexity?

Example: state roles in review process



- State establishes criteria, sets up application process. State says districts/publishers are largely responsible for assembling evidence for application, generating acceptable assessment data for state use, and paying for the additional assessment option
- State establishes criteria, sets up application process. State takes large role in gathering evidence and implementing tests.
- State reviews and adopts multiple assessments without district request. State takes responsibility for gathering evidence, negotiating contracts, paying for the multiple assessments, and implementing tests.

State values, approach

- If a state has previously considered and rejected a college entrance exam (CEE) as the state exam in high school, and the state continues to view a CEE as unacceptable, then the state should make sure its criteria are very clear, the process robust, and both are defensible.
- If a state thinks a CEE is probably acceptable, then the state should be proactive in working out the operational, logistical, contractual, etc. details.
- If a state is not sure (including disagreement among key stakeholders), then the state should start by laying the foundations of what would be acceptable goals and evidence. For example, "Are CEE aligned to state content standards comparably with the state test?" versus "Do other strengths of a CEE compensate for lower alignment?"

Possible next steps

- Possible next steps for states (and perhaps CCSSO)
  - Look for final regulations from USED
  - Learn more (e.g., probe USED about tendency to approve)
  - Generate specific evaluation materials
  - Generate specific multiple-assessment contract/operations materials
  - Generate specific LEA (review) materials about LEA actions (e.g., notification)
  - Communicate about Review criteria?
  - Coordinate Peer Review submissions?

Summary

- Be **clear** about state's values, uses, resources
- Specify the criteria **both** for an acceptable high school assessment *and* for acceptable **multiple** assessments, considering systemic coherence
- The criteria must address not only technical quality but probably **all operational aspects** in state's assessment contracts.
- The process is not only technical review but **procurement**.
- **Communicate** before, during, and after adoption/review process
- The state will need **specific** criteria, evidence evaluation guidelines (rubrics), policies, and programs.
- The state should be **proactive**!

Using System of Innovative/Interim Assessments Instead of Summative

Center for Assessment

The ESSA Provision and Guidance

Center for Assessment

ESSA interim assessment provision

- ESSA allows state to consider using "a single summative or multiple interim assessments" to comply with assessment and accountability requirements of ESSA
- Single interim assessment program for the state (e.g., same set of interims in each grade)—not multiple interim assessments selected by districts
- Interim assessment would need to pass Peer Review

ESSA: innovative assessment pilot

- ESSA permits USED to provide demonstration authority to up to seven states to pilot an innovative assessment system and use it for accountability and reporting purposes while scaling such an assessment system statewide
- SEA may propose the innovation, such as performance-based assessments, assessments supporting a competency-based education model, etc.
  - See NH's PACE Project under ESEA Waiver (state competencies, local performance assessments, common tasks, common state assessments in select grades)
- SEA must demonstrate quality of the innovative assessments, including comparability across districts and time

Why a state may be interested

- **Time.** Overall reduction in testing time (if state and districts already administer summative and interims)
- **Relevance.** Have state-sponsored test(s) provide more information useful for improving instruction and curriculum. May fit instructional model better.
- **Depth.** May provide ways to allow more complex, extensive assessments for summative use (e.g., performance and/or curriculum-embedded assessments prior to end-of-year summative)
- **Support.** Capitalize on support for interim/ performance assessments and their features.

Why a state may be wary

- **Technical Quality**. Adequacy of technical characteristics of innovative/interim assessments for high stakes uses?
- **Purpose**. Repurposing innovative/interim assessments for summative purposes?
- **Support**. System will require additional resources, as well as support of, and for, educators, administrators and policy makers.
- **Unknown**. Such systems of innovative/interim assessments for summative purposes have not been implemented before, or have been criticized/dropped.

Evaluation Guidance

Some key questions to guide evaluation of the use of interim assessment results to produce summative results

Center for Assessment

Key questions

1. How similar are the constructs, claims, scores, and uses of the innovative/interim and summative assessments?
2. How would a score/inference based on the innovative/interim assessments be generated?
3. Are the innovative/interim assessment administrative conditions appropriate for summative use?
4. Are the innovative/interim assessments' technical characteristics appropriate for summative use (e.g., fairness, scale stability, reliability/precision, documentation)?
5. Are the practical aspects feasible for the state (e.g., cost, support, data responsibility, control)?

1

Constructs, claims, scores, and uses

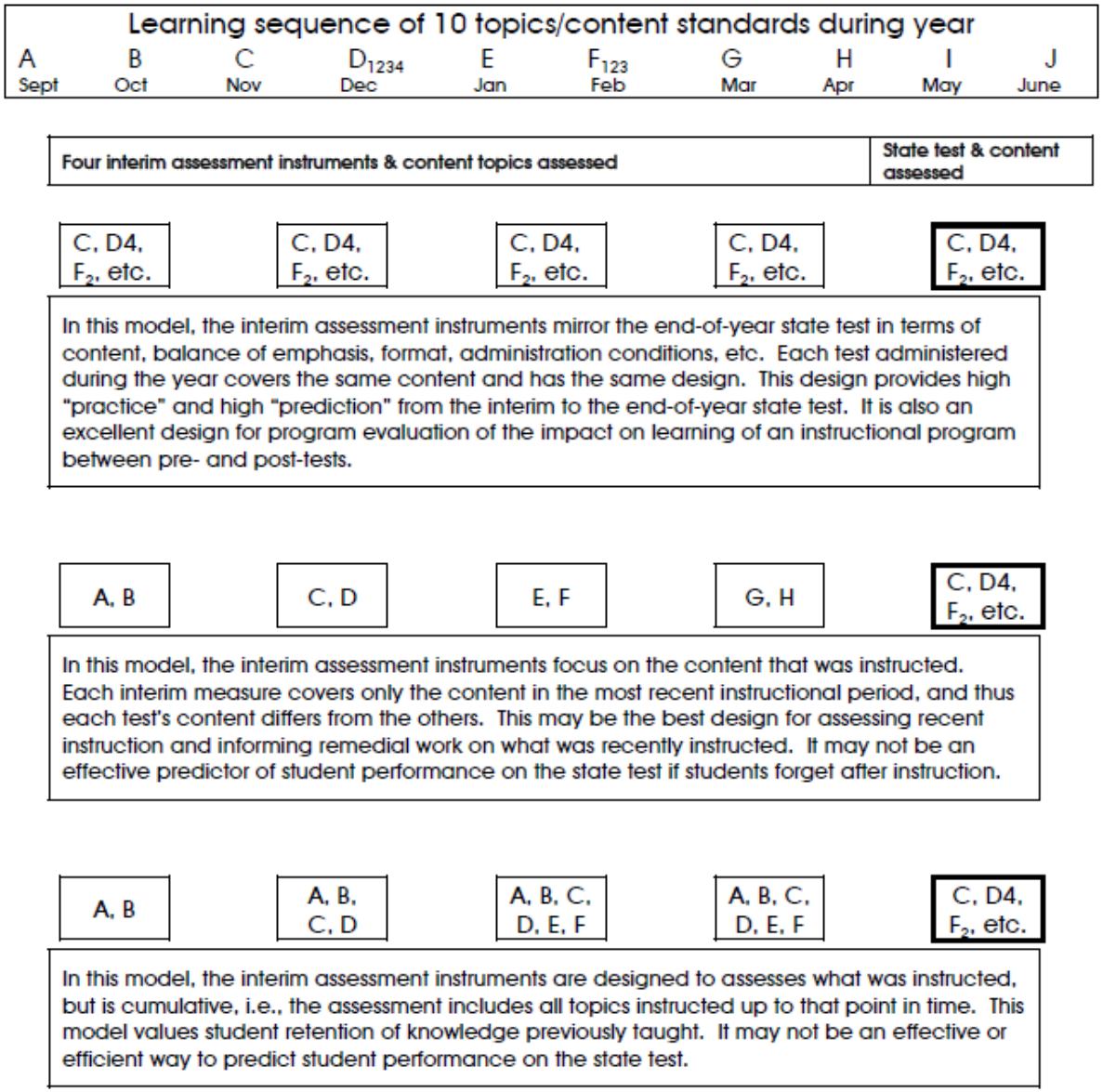
Center for Assessment

What is the state's summative claim?

"Proficient"/"Ready" means:

- Student is ready for next stage (close in time to when student will enter next stage) because student has learned the requisite knowledge/skills
- Student has learned the requisite knowledge/skills at some time (which may be subject to forgetting or additional learning)
- Student is likely ready for the next stage based on predictive performance, but no claim that the student has learned full set of requisite knowledge/skills

What information is provided by innovative/interim assessment?



Center for Assessment

How is interim assessment information used?

- **Inform growth:** Interim assessment information is used to inform instruction intended to improve the learning of the same students who took the interim assessment within the same grade/year in which the interim assessment was administered.
- **Document performance:** Interim assessment information is used to document student learning at the end of instructional units, as the likely "high point" of learning close to instruction; the student performance at that time informs decisions (e.g., grades) or actions (e.g., competency "move on").

2

Single Summative Student Score

Center for Assessment

The results from each interim must be combined into a “single summative score that provides valid, reliable, and transparent information on student achievement or growth.”

Summative Score: Key Questions

- Within each grade, are the interim assessment blueprints the *same* or *different*?
- Is the single summative score meant to capture what students know and can do *throughout the year* or at the *end of the year* (e.g., how does the claim account for time)?
- Does "single summative score" translate as a *scale score*, *proficiency level classification*, or both?

Same Blueprint

- Allows for the computation of growth.
  - Unless different modes of assessment are taken, e.g., one assessment made up of multiple choice and constructed response and a second made up of extended performance tasks.
- However, creating a score to represent status may run into conceptual trouble – i.e., status at what point?
- Some potential claims about status include:
  - *Average* achievement during the year.
  - *Best* achievement during the year.
  - Others, e.g. a *composite* in which achievement on each assessment is weighted empirically or based on value-judgment.

Same Blueprint: Student Score

- Wise (2011) simulated data four for quarterly assessments and examined three approaches to creating a single summative score that are applicable when the interim blueprints are the same.
- Assuming that the score is meant to capture end of the year achievement:
  - A simple average generally under-estimates end-of-the year achievement.
  - The maximum score over-estimates end-of-the year achievement.
  - A weighted model* generally captured end-of-the year achievement well.

* Wise uses a weighting scheme of 4, 8, 12, and 17 for the scores from quarters 1 to 4, respectively. This scheme is meant to reflect the amount of instructional time before each assessment.

Same Blueprint: Standards Setting

- Given an approach to creating a score like those previously mentioned, standards setting could proceed using common methods (e.g., the bookmark method).
  - Then, for example, the average score would be used in lieu of the score from a traditional summative.
  - Alternatively, each test could produce a classification (using the same cut points) and a final classification could be produced using a decision rule.
- However, what data should be used for standards setting is an open question.
  - E.g., item orderings for the bookmark method may change depending on when the assessment is administered.

Different Blueprints

- Results must be combined not only to meet the ESSA provision, but also to represent the full set of content standards.
- Potentially, each assessment could be treated as “if they were different sections of the same test” (Wise, 2011).

Different Blueprints: An Aside

- State content standards represent a consensus about what students should know and be able to do, but are agnostic about *when* student should demonstrate their knowledge and abilities.
  - Implicitly, students should master the standards by the end of the year.

Different Blueprints: An Aside

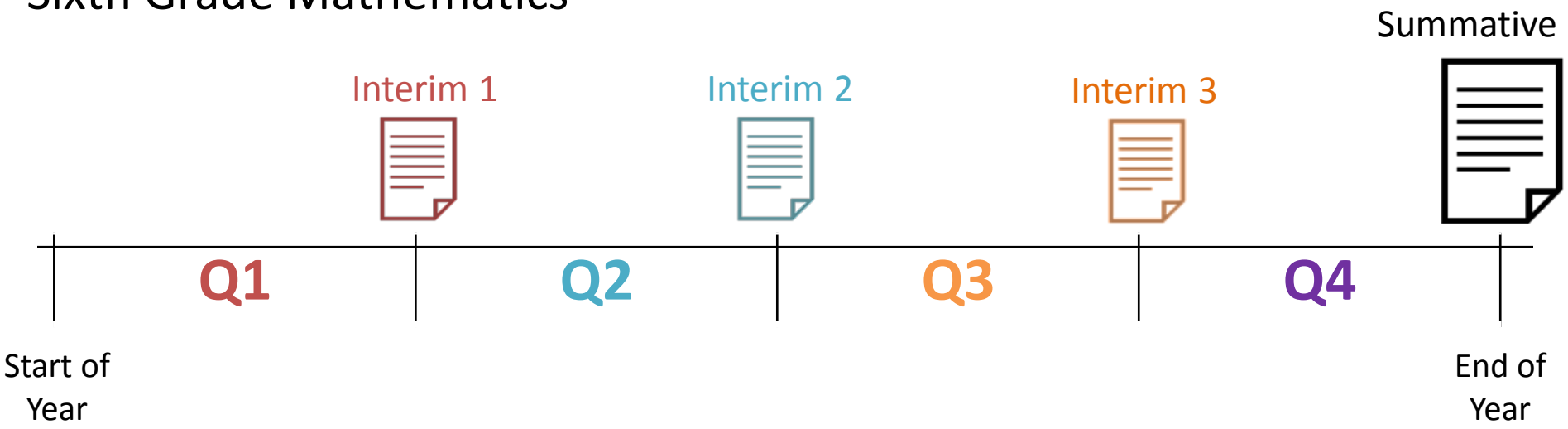
- Developing such a consensus does not require a state-wide common or shared curriculum.
- However, there does need to be
  - Some agreement on what content standards should be on each assessment,
  - Very wide and flexible administration windows, or
  - Both.

Different Blueprints: Student Score

- Based on our empirical examinations, the method (i.e., average, maximum or weighted) used to combine scores does little to change the association between the resulting summative score and performance on an end-of-year summative assessment.
  - That is, the ordering and magnitude of differences among students changes little from assessment to assessment.
- Data for this investigation stems from a district's three quarterly interim assessments and end-of-the year summative assessment in sixth grade mathematics.

Different Blueprints: Student Score

Sixth Grade Mathematics



- 3 interims with 30 items each & end-of-year summative (approx. 50 items)
- Interim items generally aligned to instruction in prior quarter
- Approximately 5,000 students

Correlations with Summative

	Scale Score	Achievement Level	Proficiency
Average (or sum)	0.83	0.81	0.71
Maximum	0.81	0.78	0.68
Weighted	0.83	0.81	0.71

Different Blueprints: Standards Setting

- Options:
  - Treat the set of interims as a single test and set standards using the entire set of interim items.
  - Set standards on each assessment, then use a decision rule to create a final classification.

- The question of what data to use arises here as well. Specifically, should data be pooled for each interim (ignoring differences in administration order and timing) or should only a subset be used?

Tentative Conclusions

- When the blueprint for each interim is the *same*, and the claim is about end-of-year achievement, then the summative student score varies by aggregation method.
- When the blueprint for each interim is the *different*, and the claim is about end-of-year achievement, then the summative student score does not vary much by aggregation method.

Caveats

- We've assumed that each assessment is scaled independently and the resulting scores need to be combined.
  - An alternative is to use a measurement model that combines the results in one step, e.g., latent growth curve models, longitudinal IRT model.
- Properties of summative scores (e.g., reliability, classification accuracy & consistency, standard error of measurement) require additional exploration.
- The distinction between blueprints that are the same and different is not as clear-cut as we have shown here.
  - There could have substantial overlap, which would allow for linking.

3

Administration

Center for Assessment

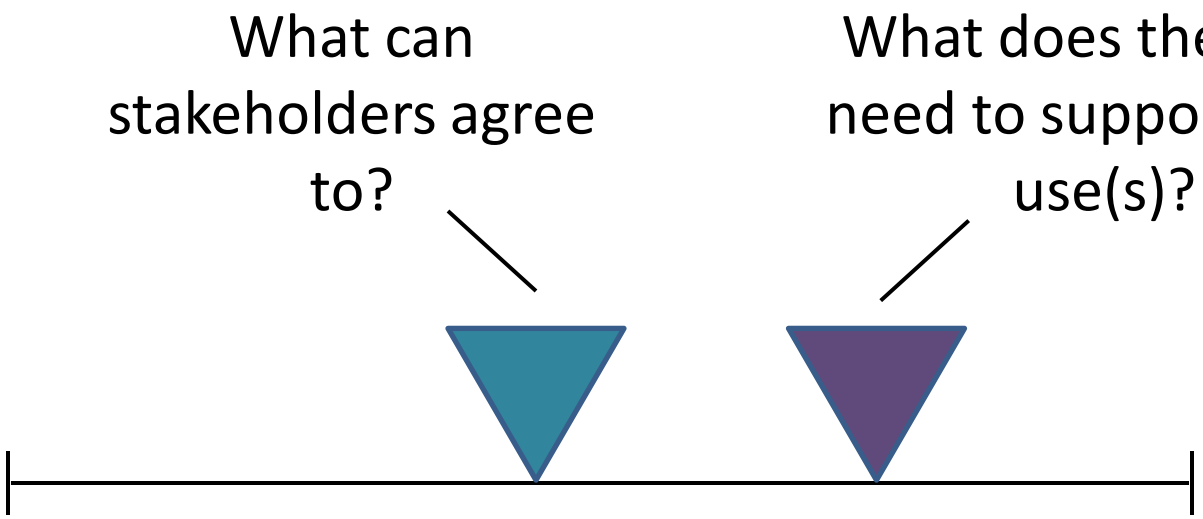
What administrative conditions are required?

Condition	Summative	Interim
Administration & scoring	Highly standardized	Flexible in timing (usually); varies across districts
Administration Security	High	High (district high stakes) to low (classroom use)
Curriculum-specific	Low (grade-level)	Low for commercial; high for district-custom
Accommodations	State mandated	Variable; rarely validated
Transparency (e.g., public, reviews, documented)	High (typically)	Low (typically)
Control	High (if state custom)	Low state control if interim controlled by district
Participation	High for accountability	Low for individual student; Med-high for district eval.

Developing Agreement

- In the prior section, we suggested developing agreement around the blueprints of the interims.
- Similarly, a state may be well served by developing stake holder agreement on assessment administration.
- As an example, consider the *order* and *duration* of the administration windows. Some questions are:
  - If the blueprints are different, can the assessments be given in any order?
  - How long are the assessment windows? Can they overlap?

Administration Flexibility As a Continuum



Any test administered at any point in time without a specific security protocol.

Each test administered within a short window in a prescribed order with stringent security.

4

Technical Characteristics

Center for Assessment

Fairness

- In addition to concerns that arise with traditional end of the year summative assessments (e.g., accessibility, accommodations), administering multiple interims poses unique challenges in terms of opportunity to learn.
  - Variability in curriculum and instruction interacts with the timing and order of assessment administration.

Reliability

- The reliability of each interim assessment, as well as the single summative score, are of concern.
- How to define the reliability of the single summative score is an open question.
  - Much research has been done on the reliability of composite scores, but we are not aware of research investigating the reliability of composite scores where the scores are defined at different time points
  - However, when the test blueprint is the same (e.g., repeated measures), methods from the longitudinal research literature may be drawn upon

Maintaining Year to Year Comparability

- The proper equating design given a particular set of interim assessments is unclear.
  - That is, should, and if so how, differences in administration be accounted for in the equating? Should data from a particular window be chosen to define the parameters?

5

Practicality

Center for Assessment

Cost

- While some efficiencies can be gained, many expenses increase linearly with the number of assessments administered (e.g., help desks).
- However, many costs are dependent on the design (same/different blueprints) of the interim assessments.
  - E.g., with interims that have different blueprints, the total number of items needed across assessments is likely to exceed the number of items needed for a traditional end-of-year summative.

Support

- Implementing and maintaining such a system is will require additional training and support of educators and administrators.
- Long term viability will depend on the support of educators, administrators and other stakeholders, including the public at large.
  - By providing additional utility, reducing burden, or both.

Data sharing

- If the interim assessments were not owned by the state (e.g., a commercial test or one developed/published by a district), then the state would need an agreement to be able to access the data needed for reporting, accountability, and monitoring of test quality.
  - Least amount of sharing at student level: Achievement level

Control

- A state typically has a high degree of control over most aspects of the tests used to generate state assessment scores and for use in the state accountability system. There may be less control, especially for use of commercial tests.
  - Control over scale stability, test blueprint, administration policies, public statements about assessment quality and results, etc.
- The unintended negative consequences in accountability may be different for interim assessments than for a single summative.
  - Possible effect on instruction for a student who is far behind/below on the interim assessments partway through the year.

Comparability of systems

- Unit of comparability: Across students within district, across districts, across years
- Means of establishing/evaluating comparability:
  - Conceptual analysis of claims,
  - Content review and analysis
  - Analysis of student/school/district performance

Some possible approaches

- Consistency of achievement classifications (Lyons & Marion, 2016)
- 16 design options, organized by Common Students (All, Some, None) and Common Measures (Both, Some, Third Measure, Other)
- Caveat: If truly innovative, then should not expect strict comparability

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 http://www.nciea.org/publication_PDFs/Center%20for%20Assessment_Comparability%20Recommendations%20for%20Section%201204_090716.pdf

Same measures in both units (e.g., districts)

	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

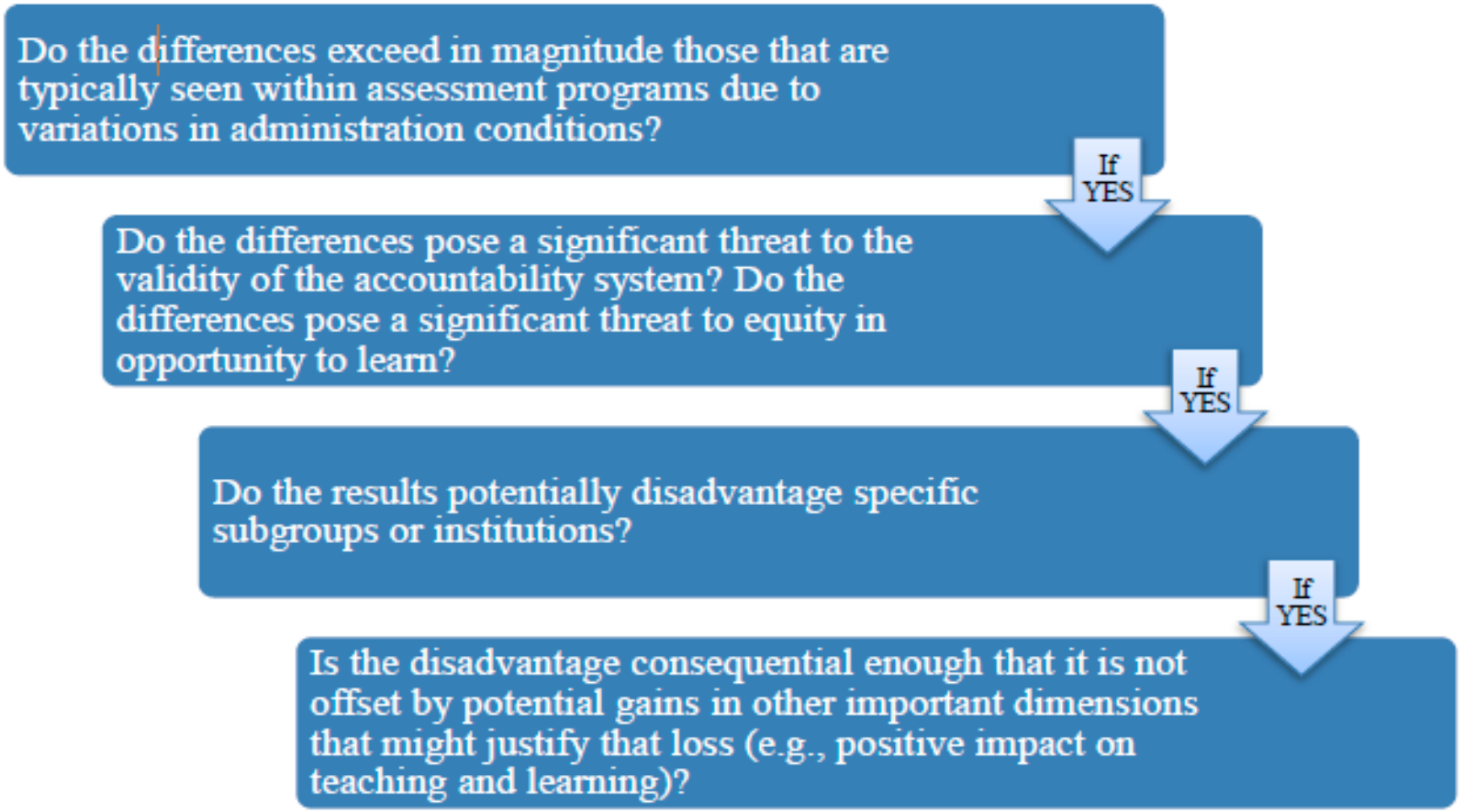
Some measures in both units (e.g., districts)

	All Students	Some Students	No Students in Common
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 both units, Other (e.g., districts)

	All Students	Some Students	No Students in Common
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

“Good enough” criteria – decision tree



Lyons & Marion, 2016

Extreme comparisons

- What if the construct is inherently unreliable, but valid? – For example, “best work portfolio”
- What if timing is supposed to be non-standardized? – For example, “assess when ready”
- What if content standards are not fixed by grade level? – For example, Learning progression, individual progress
- What if content standards are not common to students? – For example, student pathways to different goals
- What if assessment evidence is deep but not representative of the full construct, and varies by student? – For example, student choice
- What if assessment evidence comes from non-score system? - For example, content sequence

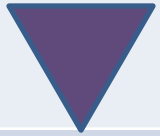
Final Remarks

Center for Assessment

General Conclusions

- Implementing such a system of interim assessments:
  - will require a sustained, multiyear effort that goes above and beyond the effort currently involved in typical summative assessment programs
  - may require developing agreement among stakeholders on a number of issues not often addressed in typical summative assessment systems
- Most commercially available assessments would likely require additional documentation, development, or both to meet this ESSA option.

A self-evaluation rubric

	Level 1 (little)	Level 2	Level 3	Level 4 (a lot)
Tolerance for Innovation Risk				
Amount of State Control			 Can get	 Need
Congruence between interim and summative				
Administrative feasibility				
Technical quality				
Stakeholder support				

For more information:

Center for Assessment

www.nciea.org

Brian Gong

Nathan Dadey

bgong@nciea.org

ndadey@nciea.org