Session 3D: Addressing Comparability in IADA

Panelists: *Scott Marion, Carla Evans*

[*Sessions 3D and 3G Addressing Comparability in IADA slides 8–18*]
The session started off with small discussions about comparability. Dr. Evans then shared a definition of comparability: the degree to which the results of assessments intended to measure the same learning targets produce the same or similar inferences. In the case of IADA, in which states are running two or more assessments, we want to know that the achievement levels are comparable across the tests. The inferences made about what students can do should be similar or the same. ESSA's Section 200.105(b)(4)(ii) requires that states' innovative assessment systems generate results, including annual summative determinations, that are valid, reliable, and comparable for all students and for each subgroup of students among participating schools and LEAs, which an SEA must annually determine as part of its evaluation plan described in Section 200.106I (proposed Section 200.78[e]). Section 200.105(b)(4) has been revised to clarify that determinations of the comparability between the innovative system and the statewide assessment system must be based on results—including annual summative determinations, as defined in Section 200.105(b)(7)—that are generated for all students and for each subgroup of students. Comparability must be reevaluated every year per the final regulations for IADA (https://www.gpo.gov/fdsys/pkg/FR-2016-12-08/pdf/2016-29126.pdf).

Although "comparable" is not defined in ESSA, to show comparability between innovative tests and traditional tests, ED asks SEAs to determine comparability of assessment results in one of five defined ways.

1. Double-testing at least once per grade span in the same subject.
2. Double-testing using a demographically representative sample of all students and subgroups at least once per grade span in the same subject.
3. Linking or anchor sets from state to innovative (items or performance tasks).
4. Linking or anchor sets from innovative to state (items or performance tasks).
5. Using an alternative method. ("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 statewide assessment, including for each subgroup of students.").

The basic requirement for comparability is that (1) the same students take two different tests or (2) different students answer the same test questions. There is a fundamental tension between innovation/problem-solving and comparability to a legacy assessment program. It is important to note that a state is free to establish new achievement standards for its new assessment once it has implemented the test statewide. The comparability requirement is in effect only as long as the state continues to use the legacy assessment in non–IADA pilot schools. The comparability requirement is in place only while the two tests are happening. That is why Massachusetts is rapidly prototyping and will then switch the innovative test to become the state system.

Dr. Mbella asked whether everyone still takes the statewide assessment in Massachusetts and whether the innovative tests are delivered as local pilots. A participant responded that the legacy Massachusetts Comprehensive Assessment System (MCAS) involves two days of testing on parallel forms; Massachusetts took the Day 1 form for students in the pilot study and took out the application section and field test items to make room for additional items, and that is now the form it reports on (the Mini MCAS). For Day 2, students take half the IADA assessment. Dr. Evans added that this is for the science assessment in two grades.

Dr. Marion discussed threats to real innovation and said there are legitimate reasons for noncomparability, including (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 deeply, and (4) to measure targets more completely (e.g., listening, speaking, extended research, and scientific investigations). Dr. Marion clarified that moving testing time to allow students more time to learn could affect comparability but said not to be shy if that is what your state wants to do. He also explained that a state can currently test learning targets more deeply but said the state needs to be prepared to meet peer review standards. He added a quotation from Robert Brennan, Founding Director of the Center for Advanced Studies in Measurement and Assessment: "Perfect agreement would be an indication of failure." Dr. Marion said that if you are happy with your current system, there is no need for IADA. If you want to make a change for any of the reasons discussed, you will most likely encounter a threat to comparability.

[*Sessions 3D and 3G Addressing Comparability in IADA slides 19–22*]
Dr. Marion stated that *Evaluating the Comparability of Scores from Achievement Test Variations*, a 2010 report by the Council of Chief State School Officers, offers a good examination of comparability and that there are two dimensions, content comparability and score comparability. Before one applies advanced statistics, the content needs to be representative in the same sort of way. You also want your equating set to be a representative set of items from your test blueprint. Dr. Marion shared an example of how reading and writing are often equated only through the reading items and that it is difficult to assess differential improvements in writing. IADA gives states a chance to try prioritizing standards in different ways, which all need to be measured, but states should be aware that they might start to threaten content comparability and think about what they might expect.

The challenge of comparability goes beyond IADA. Producing "comparable annual determinations" is a key ESSA requirement. Dr. Marion recommended the National Academy of Education's report titled *Comparability of Large-Scale Educational Assessments: Issues and Recommendations* for further reading on the topic. Dr. Marion offered that challenges to accommodations include online versus paper and pencil, test accommodations, computer adaptive tests, translations for English learners, and alternate achievement standards. Dr. Marion urged attendees to remember that correlation does not mean comparability.

Dr. Evans shared a submitted question: Could you talk more about the methods of establishing comparability? For example, what do people report out as evidence, and are any of those correlations? Dr. Marion replied that if you think students will perform differentially—for example, performing better on IADA than the traditional test—that's going to affect correlation.

If you have the same content standards but a different design, look at comparable items, students who are comparable, expected trends, and a baseline. One will always see variation, but how much should one expect? How much variability is too much? One can look at the known variability in a state's assessments over the years and see the distribution of variation and what the acceptable bounds of variability are to determine whether one is in the ballpark. Dr. Marion offered that if you can't solve the problem in front of you, try to find the problem that you can solve.

An attendee from Massachusetts commented that states are making different claims from the traditional assessment and designing new assessments but are asked to have comparable results. They need comparable results for ESSA. But the attendee said the team felt frustration when trying to make two goals meet, which is why Massachusetts went a different route and made the standard test shorter so that the only thing that needs to go through IADA approval is the mini assessments. States need to show that the linking works and that they are giving comparable results. Massachusetts is building and trying another test out, and no student is being double-tested because all students sit for the same amount of time. But Massachusetts is gathering the data and will be able to use the results instantly. A third option of sampling might also work. Dr. Marion replied that sampling is one of the four methods.

A representative from a research organization asked how states can say they have a different claim—i.e., "I expect different things. I want a different interpretation. But I want to meet ESSA requirements." A representative from another research organization, replied that studying correlations might be helpful in this case to provide some of the evidence that you are studying something different. Similar to New Hampshire, an innovative approach is to look at the correlation of students who took the regular assessment from one year to the next and then the correlation of those students who the next year took the innovative assessment. If the correlations are very high or almost exactly the same, you have to question whether you are doing something innovative. If the correlations are very high but your categorizations are different, you may just have different levels of expectations for students. But examining the categorizations and the correlations could help you build the case that you are measuring something that is substantively different and you should expect lower levels of comparability. Or it could be that you are measuring something that is mostly similar and, if you don't get a similar level of comparability, maybe you are advantaging or disadvantaging one program or the other. Doing things such as that could help build substantial evidence to show the level of comparability needed.

Dr. Marion added that one can inform a priori expectations with small-scale studies. If you held think-aloud tasks or cognitive labs with students and saw them interacting with the material differently, you could add items to the innovative test or standard test without full field testing; it's a way to gain expectations. For IADA, you don't need to go in with comparability established; you need a plan for comparability. An assessment peer reviewer once said during a peer review panel that the ED panel leader said that "we just have to think about the preponderance of evidence". A representative at a research organization raised an issue about the fairness of showing comparability on the aggregate level when you know it is not there on the individual level. Dr. Evans said that’s the case you make when you write your argument. The researcher reiterated that it may be comparable enough for accountability, but there’s a problem when we know that we will disadvantage certain schools based on the way the test is designed. Dr. Marion replied that if you are going for full interchangeability at the performance level and at the student level, it is a higher bar if we expect these reasons for noncomparability. If you had an assessment that would help high-performing students perform even better, that would be a harder sell, but if you had an assessment that would improve the engagement and performance of students who are traditionally less advantaged, people would be unlikely to complain. Dr. Marion went on to share Louisiana’s innovation to provide access to the ELA text through the years so that it is more of an access test than an IQ test.

Dr. Peasley was asked whether one could provide a decision tree on the IADA so that one doesn’t have to establish comparability while the two systems are running. Dr. Peasley replied that he likes to think that if it is a well-designed plan with a strong rationale, ED would have to consider it.

Dr. Marion provided a brief overview of IADA states’ approaches to comparability [Slide 21]. Regarding the need for comparability, he recommended that people point at the same standards and provide evidence that although they may be measuring the intended learning outcomes differently, they are doing so faithfully with a legitimate and defensible design. Dr. Marion reported that Dr. Peasley added a provision that there would be similar classifications of achievement levels. Dr. Marion disagreed with the provision, stating that he believes there could be different degrees of achievement in innovative assessment systems. The base of the case could be documenting, with high-quality alignment studies, that your innovative design measures the standards as proposed in the design and in a way that meets federal law. Performance shouldn’t be orthogonal to the base test, but it should be allowed to be different.

Questions and Comments

Dr. Mbella asked how to address expectations that the tests will be equal. Dr. Marion said education and communication are needed to set expectations and let people know that the relationship between the tests will be evaluated and though it is anticipated that they won’t be exactly the same, that’s OK because the innovation should be better. Dr. Evans shared a question about what the parameters are for sampling size, inclusion requirements, and other characteristics, if you are looking for common items. Dr. Marion said that there one can fall back on the professional literature in the field of educational assessment.

Session 3G: Addressing Comparability of IADA (Repeat of Session 3D)

Panelists: *Scott Marion, Carla Evans*

[*3D and 3G Addressing Comparability of IADA slides*]
As the presentation was the same for sessions 3D and 3G, only the questions and comments unique to Session 3G are provided here.

Questions and Comments

An attendee asked whether ED had replied to the RFI comments. Patrick Rooney replied that the information is being used to inform ED's next steps and that there is no requirement for official responses to individual comments from RFIs. Dr. Banks of ED then clarified that achievement level alignment does not need to be exact, but they should be similar. For example, a student could be at Level 3 on the traditional assessment and at Level 4 on the innovative assessment. Dr. Evans suggested that variability in statewide assessments from year to year could be used as an a priori threshold on what is acceptable and comparable.

Dr. Marion stated that if you are happy with your current state assessment, you may not want to innovate. If you want to make some minor improvements to the assessment, you will want the results on the innovative and traditional assessments to be close. If you want to do something different, why would anyone expect the same results on both assessments? The language of the law mentions the concept of competency-based personalized learning. With this concept, you measure when students are ready to demonstrate mastery. If one student is ready to demonstrate mastery in November and another is ready in February, we should not expect the results to be comparable. Dr. Marion said that the same goes for modular design when the same material is tested but can be taken in a different order. He added that others might disagree with these opinions.

Dr. Marion stated that the desire to measure targets deeply could also lead to noncomparability, seeing as most state assessments superficially cover the standards because there are a lot of standards and there is not a lot of time. If one shifts the makeup of the test to gain deeper and richer information on a subset of the domain—as long as one can prove that the other standards were covered in some way—one would not expect the same results.

An audience member asked about sample testing in science. Dr. Marion replied that many states are conducting sample testing in science and aiming to make deliberate choices about what they are testing, as opposed to randomly selecting items. Dr. Marion said that IADA applicants could say that they think these selected practices are the most critical for students at a certain grade and that they will focus more on these practices, without ignoring the others, and develop assessments that provide multiple opportunities for representations of these practices. Dr. Marion explained that the reason to do this is the school system cares that students have particular knowledge and skills. He offered writing skills as an example. Most ELAs test heavily on reading. The results of an innovative assessment with a focus on writing would most likely not be comparable to the traditional assessment.

Dr. Marion said that all programs should be expected to explain what they think will happen with the innovation and what the influence on comparability will be. In the most radical sense, one could say that the results will be completely unrelated to the legacy test, although that's a statement most would be unwilling to make. Dr. Marion added that most professionals in test development make these types of predictions based on students' reactions during cognitive labs or think-aloud tasks.

Dr. Marion discussed anchoring comparability to standards. Instead of worrying about comparing the innovative test score, whether at the achievement level or at the scale score level, or the legacy test score, consider the comparability of assessing the standards. Both traditional and innovative assessments can be judged independently in different but legitimate ways; both will have different results, but one can document that the results were derived in ways that represent the standards. Mr. Rooney of ED noted that the statute says that comparable results should be generated and that just focusing on alignment of content is insufficient. Mr. Rooney continued that "comparable" does not mean "equivalent." "Comparable" provides room to work; it does not mean exact scale-to-scale scores or the same exact percentage of students passing both tests. Dr. Banks added that IADA is flexible enough for the state to come up with the parameters for what comparability means; states need to make a reasonable argument and show the evidence for it. Dr. Marion stressed the importance of understanding that states do not need to have established comparability for their applications; they need to show they have a plan to do so.

An attendee asked about the comparability requirements when standards change radically but the standardized assessment stays the same. Dr. Marion suggested looking at the PLDs or ALDs, which should have different meanings. If any of the content described shows up in items around the cut score, a new standard setting could be needed. Dr. Banks added that such a change in standards could trigger a peer review.

A question was asked about how to standardize assessment for a subject that is locally based and involves hands-on learning. How do you offer flexibility yet ensure standardized comparability? Dr. Marion replied that you would have to keep some aspects the same. For example, Massachusetts keeps some aspects standardized but also allows for innovation. Dr. Marion mentioned that Queensland, Australia, offers the flexibility of locally driven curricula. However, emulating this example would be difficult in a state with a large number of schools. Dr. Evans cautioned that it can be a problem when IADA states try to merge instructional and assessment purposes. Dr. Banks added that there can be variability in test items that offers teachers flexibility with instruction, but some parameters must be set to show comparability.