Validity Considerations When Measuring Growth

Scott Marion
Center for Assessment
The 6th Annual Reidy Interaction Lecture Series
October 7-8, 2004
smarion@nciea.org

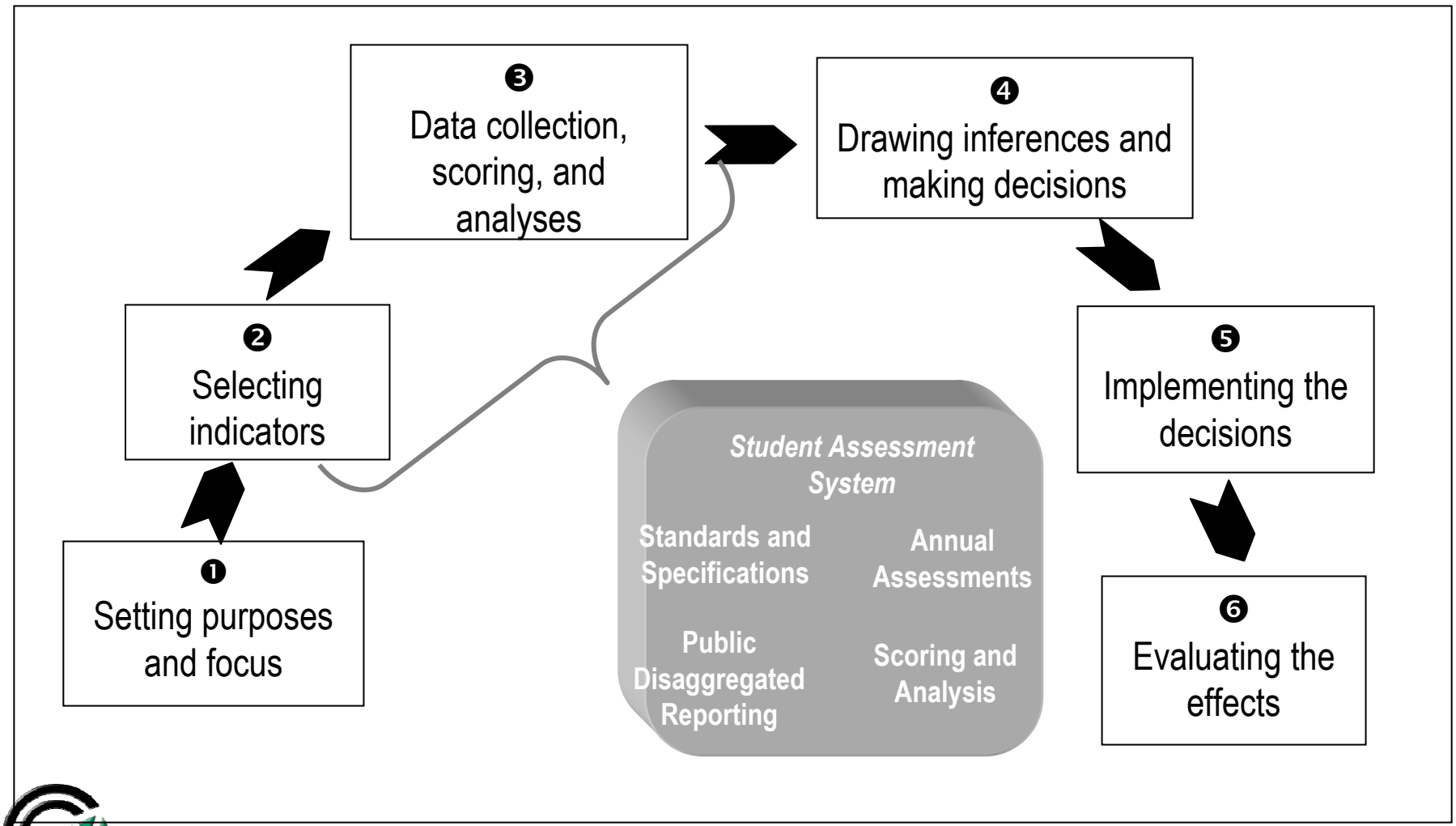
Purpose of this Session

- To introduce some ideas that will be covered in more depth in subsequent sessions.
- To briefly discuss other issues to consider when evaluating the validity of student growth assessment and accountability systems.

Validity of what?

- Inferences about students from assessments
- Inferences about students, teachers, and schools from measurement model
- Inferences about schools (or teachers) from accountability model

State Accountability System



From Carlson, 2002

Challenges of Growth

- While Dale's schematic seems daunting, consider the added challenge of validating the 2nd, 3rd, and 4th boxes when we are measuring student growth between two assessment events.

It all starts with the standards

- In order to make valid inferences about student growth, we need to be able to talk about this growth in the context of the content and cognitive demands.
- We have been concerned about the quality of state standards within grades, let alone across grades.
- Without these content anchors, we can only measure growth normatively—
  - Even with these anchors, we might need some normative help
- Laurie will provide an overview of criteria for developing vertically-articulated standards and Karin will discuss the lessons learned from creating developmentally aligned grade-level expectations.

Grades and standards

- We are constrained by our traditional model of schooling because of the need to fit growth into the typical school calendar and grade structure
- Imagine if we could create the same types of developmental benchmarks throughout the entire grade span that we use in early reading or writing?
- Our measurement model could then be based on the attainment of these concrete benchmarks.
- The model for the development of GLEs is a step in this direction.

Test Specifications

- Translating standards/GLEs into test items requires clear and coherent test specifications.
- Test specifications tell us:
  - What type of items can be written
  - The limits of such items (what's allowed)
  - How the content should be represented in the items
- These test specs and derived performances should lead to the measurement of knowledge in a developmentally coherent way across grades.
  - We should question whether or to what extent this has been done
- Stanley will talk more about the challenges of creating test specs when we are interested in measuring growth.

Vertical Scaling

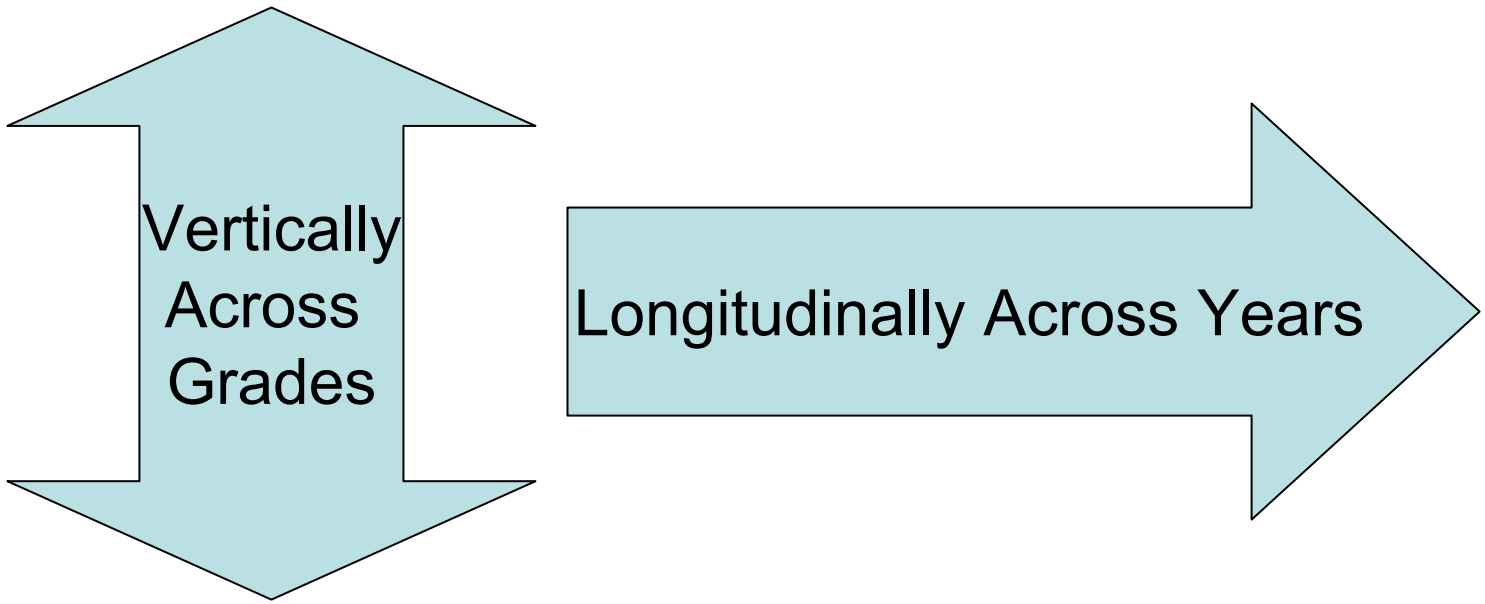
- Most current growth modeling or value-added models proceed as if there is an equal-interval scale across grades.
- When/if creating a vertical scale, we need to keep asking if our intended inferences are:
  - Across adjacent grades only
  - Across a significant grade span (3-8)
- The answer will help us focus on the validity of the construct interpretation across the intended span of inference

Vertical Scaling-2

- Vertical scales are intended to connect scores (or score interpretations) across multiple grades.
- We should question how our current, generally compensatory, scoring models affect the validity of our interpretations of movement on the scale
  - Same profile across years, but more correct responses
  - Different profiles across year, but not necessarily more correct response in specific areas
- Laurie will talk about these and other challenges, considerations, and techniques when creating (or deciding not to create) vertical scales.

A word or two about equating

- Vertical scales are based on equating forms across grades.
- But, we are also interested in maintaining our year-to-year equating, both within and across grades



Equating

- We and others (e.g., Michaelides & Mislevy, 2004; Skorupski, Jodoin, Keller, & Swaminathan, 2003) have become increasingly concerned that many "across year" equating designs are not adequate for capturing change in performance.
- The validity of the equating must be established both across years and across grades (within years).
- It could also be argued that the vertical equating across grades needs to be validated (e.g., does a 100 point gain between 4th and 5th grade mean the same thing in each of 2 years?)
- Many growth models based on NRTs have avoided part of this problem because the test remains stable for many years, but…

Instructional Sensitivity

- Most state leaders and others reject the notion of measuring growth between high school biology and chemistry or between algebra and geometry.
- The same people often embrace the measurement of growth between 4th and 5th grade mathematics.
- Why the difference?

Instructional Sensitivity-2

- If people argue that there are not the kinds of curricular/instructional differences across elementary grades as we find across high school grade, does that mean we aren't doing what we should be doing in elementary mathematics?
- If we get much better at designing GLEs and building instructionally sensitive tests, won't we run into the same concerns in elementary school that we have for high school?

The Measurement Model

- “It all depends on the question”
- This is true for all validity arguments.
- Once certain very big assumptions are met, the models that Pete will be talking about offer a very powerful to examine the effects of a multitude of variables on changes in student and school test scores.
- These models can be judged for their internal validity in terms of the way they can explain variance, be replicated, and make sense.

The Measurement Model-2

- Yeah, but which one is the best model?
- Like any good validity investigation, it depends on the particular question(s).
- Therefore, it is incumbent on state leaders to get very clear and specific about their accountability/evaluation questions.
- Yeah, but which is best….
- We can't really tell without some ground truthing.

The Accountability Model

- The measurement model just gives us some numbers, what we do with them is dependent upon our accountability model.
- Many validity questions need to be addressed regarding accountability models.
  - Previous work by Carlson, Gong, Marion, Forte-Fast, and others have outlined validity questions and concerns for accountability systems.

Accountability questions

- State accountability systems must reflect the values and intentions of key stakeholders.
- The extent to which an accountability system leads to improvements in teaching and learning is a key validity component of accountability systems.
- When evaluating the different measurement models and accountability proposals discussed during the next two days, state/district leaders should consider the two previous points.

Accountability Purposes

- Remember, validity arguments for assessments and/or accountability systems must always consider the purposes in the context of specific uses.
- Brian will talk about how the results of the different types of measurement models can be put to different valid uses.

Accountability Models

- Using different models or changing components within models (Rich's value tables) can lead to very different evaluations of schools.
- Brian, Rich, and Pete will talk about "conditional" models in terms of data-driven and policy-driven approaches. This distinction will be very important to consider in terms of the validity of the various models.

Conclusions

- Much of what I've said thus far appears to be critical of growth and value-added models.
- But, growth is the most important way to judge the effectiveness of schools.
- Measuring and holding schools accountable for growth is an important issues for important stakeholders.
- We need to keep asking these questions to help us find the most appropriate ways to capture student progress through schools.
- Yeah, but how do we know which one is "right"?
- Again, we need to find better ways to check these results on the ground.