Incorporating Measures of Student Growth in Educator Evaluation

SCOTT MARION
CENTER FOR ASSESSMENT

APRIL 17, 2012

Overview of Presentation

- Very Brief Review of Student Growth Percentiles (SGP) in school and educator accountability
- Discussion of Shared Attribution
  - Theory of Action
- Combining Multiple Student Growth Measures

Growth for Educator Accountability

- While criterion-based growth can be very important for school accountability—although it is not part of Utah's Comprehensive Accountability System—we we are very concerned that it is not fair to base educator accountability on criterion-based growth
  - Highly correlated with socioeconomic status
- Therefore, we recommend using normative information for educator evaluations because it is more fair to all educators than a criterion-based approach

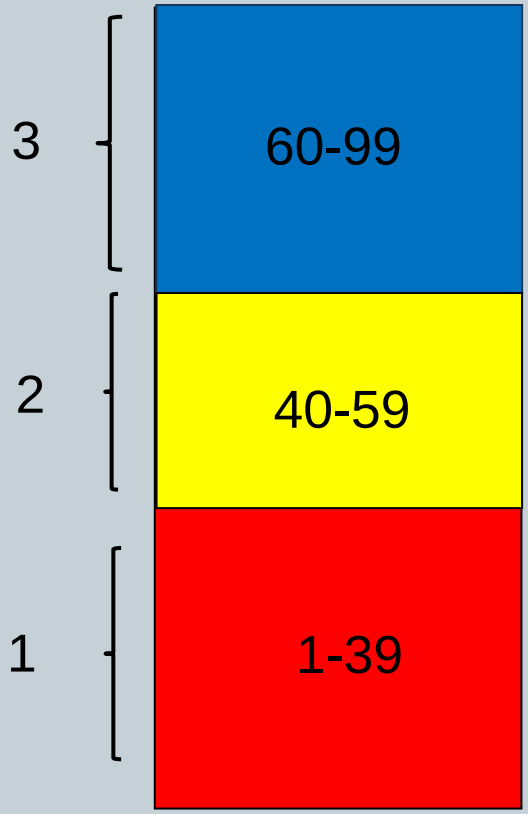
How many categories

- Most states using SGPs (or VAM) for educator evaluations are categorizing growth into three categories:
  - High
  - Typical/Average
  - Low
- Why not more?
- Given the number of students included in SGP calculations for each teacher, it is doubtful that we can reliably distinguish among more than these categories

An Example of Potential Categories of MGP

- The specific median SGP cuts will have to depend on empirical analyses, but several states are using:
  - MGP<40 = Low
  - 40<MGP<60 = Typical
  - MGP>60= High

- Potential MGP rubric for educator evaluation



Shared Attribution

- Is the approach where median SGP or other (e.g., SLO) results are “shared” among more than the educator most closely associated with the SGP results
- Can be shared among all educators:
  - In the school
  - At a grade level
  - In a content area grouping (e.g., math department)
  - Other?

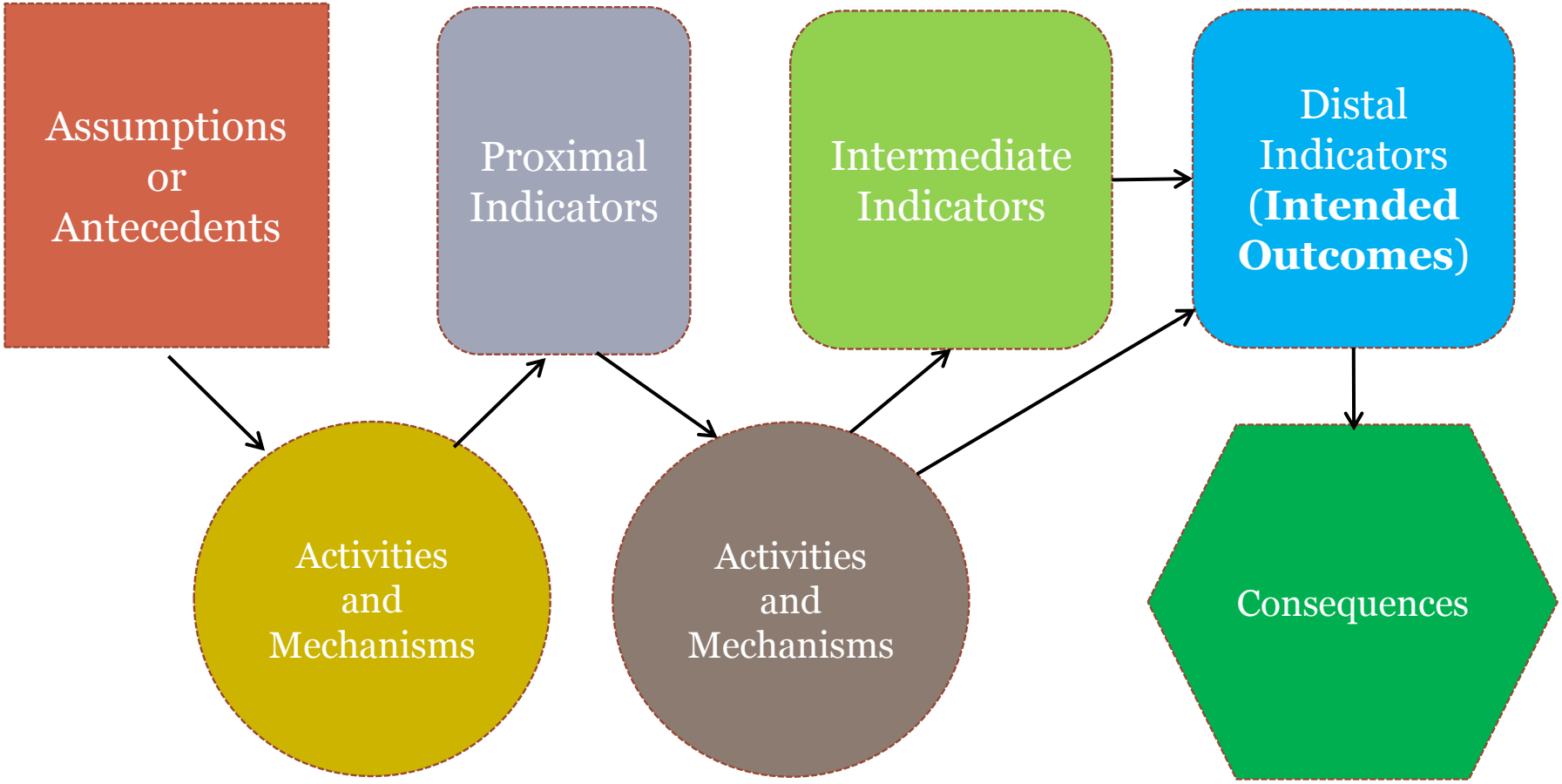
Tradeoffs of Shared Attribution

Advantages	Disadvantages
• Larger sample sizes can lead to more reliable inferences • Promotes collaboration among colleagues • Avoids “isolating” or creating a hyper-focus on reading and math teachers	• Educators are held accountable for results for which they may have little to no control • Masks true variability in educator quality

Theory of Action/Improvement

- Shared attribution should be based on more than just reliability concerns, but should be tied to your theory of improvement
  - For example, if the focus of improvement activities is the grade level team, that suggests attribution should be shared among educators at that grade
- Remember, we should talk about shared attribution for both SGP and SLOs

Basic Structure of a Theory of Action



Theory of Action

- What is your school's locus of improvement actions (e.g., grade level teams, content area departments, whole school)?
- Which subjects are shared and with whom? Does the team share both math and ELA results or just one subject?
- Who should share SLO results?
- **What should the state require, if anything, regarding shared attribution for districts?**

Student Growth/Performance

- Now that we've talked about SLOs, SGP, and shared attribution…
- We need to create an initial plan for how we will combine these multiple measures as we incorporate student performance results in educator evaluations

Aggregation & Combination

- There are many layers to the overall system that we are proposing:
  - Student Growth
    - Multiple measures of student growth
    - SGP, SLO, Shared attribution
  - Educator Practices
    - Multiple Domains
      - Multiple elements within each standard
        - Multiple measures of each domain and element (sub-domain)

- For now, we are focusing only on Student Growth

Aggregating What?

- Before talking about aggregating information to help lead toward a summative judgment, we need to talk about the nature of the data
- Student Growth
  - SGP results will have to be converted into some sort of rating system as described previously.
  - "Non-tested" using SLOs and additional measures will have to be converted into some sort of rating or rubric score
  - Shared attribution will likely rely on the same scales used for SGPs and SLOs, but we will have to determine how the results get attributed and whether shared attribution is a required or optional component of local systems

Approaches for aggregation

- Once we have these scores (ratings) for the various measures of student performance, we still need to have a way of combining these scores to produce a summary at the next level of aggregation (e.g., Student Growth)
- People appear to like and feel comfortable with simple or even weighted averages of numerical scores
  - Might not be the best approach, but it could be
  - If we want to be more explicit, panel approaches are a more transparent and value-driven way to combine scores

Four general approaches for combining multiple measures

- **Conjunctive**—must score above the criterion on **each** measure in order to meet the overall criterion
- **Disjunctive**—must score above the criterion on **any one** measure in order to meet the overall criterion
- **Compensatory**—higher performance can offset lower performance on other measures as long as some combination (e.g., average) of scores is above the overall criterion
- **Profile**—similar to a compensatory approach, a profile approach goes further by identifying specific combinations of scores (or score ranges) that must be present to meet the overall criterion
- **Hybrid**—any two or more of the above may be used in combination

Implications for scoring and interpretation

- Knowing about the relationships of the indicators—conceptually and empirically—will help inform the scoring models. For example (these are simplified):
  - With non-overlapping indicators of equal importance, a *conjunctive* decision model may be relevant
  - If the indicators are ordered, such as performance on indicator B implies a certain quality of performance on indicator A, then one might add the results.
  - With overlapping indicators, a *compensatory* decision model may be relevant

- What is your opinion of the nature of the student performance indicators in this system?

Transforming Scores into Ratings

- A score of “3” does not automatically equal “effective,” for example.
- Scores get converted into ratings by way of some sort of deliberative process. This is one of the reasons why it makes sense to have scores that do NOT map neatly onto performance categories, particularly at the fine-grained level of the data (e.g., use a scoring range of 5 points instead of 4)

At what level do we want to rate?

- We do not want to call teachers, for example, "effective" for SGP results or the results of a single SLO, etc.
- Rather, we should reserve the effective/ineffective ratings for the overall determination or perhaps at the two major subcategories.
- *Do we want to make such a recommendation?*

Some examples

- I present a few examples on the following slides to illustrate some approaches for combining multiple measures.
- For the overall rating, I recommend using a panel approach which is a hybrid of a profile and conjunctive approach. An example of such a decision table follows…
- You don't need to decide on this now; it is just for context

Panel Approach for Combining Measures (using 3 categories of student performance)

Practice Domain Score	**4**	No Rating	Effective	Highly Effective
	3	Approaching Effective	Effective	Highly Effective
	2	Ineffective	Approaching Effective	Effective
	1	Ineffective	Approaching Effective	No Rating
		1	**2**	**3**
		Student Performance Score		

Aggregating below the overall rating

- We can use the same approach for aggregating the various sources of information within each major component (practice domains and growth)
- The following slide shows how we might combine an SGP rating with an SLO rating or two SLO ratings…

Combining across SLOs or SLOs and SGPs to arrive at an overall student growth rating (four performance outcomes)

SLO Rating	**Exceeded**	2	3	4
	Met	2	3	3
	Did Not Meet	1	2	2
		Did Not Meet	**Met**	**Exceeded**
		SLO or SGP rating		

More than two measures

- It is easy to see how a decision table works with two indicators, but what about 3, 4, or 5?
- It can still be done, but it is much harder to represent it in a visually understandable way
- With multiple SLO, we can decide among:
  - Profile
  - Compensatory
  - Conjunctive
  - Hybrid

Examples of approaches using fictional SLO scores

- Key: 1=did not meet SLO; 2=met SLO; 3=Exceeded SLO
- Assuming 4 SLO per teacher and assuming all SLO are of "equal value and worth"

SLO 1	SLO 2	SLO 3	SLO 4	Compensatory Rating	Conjunctive Rating	Profile Rating
1	1	1	1	1	1	1
3	3	3	3	3	3	3
1	2	3	4	2.5 (exceed?)	1	2?
2	2	3	3	2.5	2	2 or 3?
2	3	3	3	2.75	2	3?

Discussion

- What questions do you have about combination approaches?
- Do you feel like you know enough to make a recommendation?
- Will most districts will just use a compensatory approach because that is familiar?
- Can you see why that might not be the most appropriate approach given the examples presented?
  - With homogeneous set of scores, almost any method will yield the same results, but with heterogeneous scores some important differences may emerge.
- How should we treat SGP results? As a separate category of indicator or just like an additional SLO? What about shared attribution results?
- Other issues and concerns?