Assessment and Analytic Approaches for use in Non-Tested Subjects and Grades

SCOTT MARION & CHRIS DOMALESKI
CENTER FOR ASSESSMENT
www.nciea.org

UTAH STUDENT GROWTH WORKGROUP MEETING
DECEMBER 7, 2011

Advance Organizer

- A brief overview of assessment, analysis, and accountability
- Measurement tools and approaches
- Analytic approaches
- Measurement & analytic recommendations

Some background

- There is often considerable confusion regarding the distinction between assessment, analytic approaches and accountability
- This leads to misunderstandings in the public realm about the quality and usefulness of the assessment system and the components of the school and/or educator accountability system

Assessment

- Is the process of collecting data about some set of knowledge, skills, and/or behaviors
- It can range from highly structured and formal to "in-the-moment" observations of student group work for example
- Assessment is typically classified as:
  - Summative
  - Interim
  - Formative

Analytical approaches

- Are the collection of methods-often statistical—that are used to transform or summarize the assessment results in some way
- These methods include psychometric techniques such as scaling (turning raw scores into scale scores), linking (being able to compare scores across occasions), and standard setting
- They also include statistical techniques such as a variety of growth and status calculations

Accountability

- Is the set of policies, rules and decisions that determine:
  - Which indicators (data) are collected,
  - How they are weighted and combined (if they are combined),
  - What counts as "good enough" on each indicator (perhaps) and/or on some overall composite,
  - How the results are used and reported, and
  - If there are any consequences and/rewards and how they are applied.

More about assessment

- Let's talk a little bit more about assessment types to help contextualize our discussion about the types of assessment tools that might be used for educator evaluation…

Summative Assessment

- *Assessments administered at the end of some period of instruction to evaluate students' knowledge and skills relative to a specific set of academic or other goals. The results are used for various levels of documentation and/or accountability ranging from grades to formal accountability systems.*

Interim Assessment

- *Assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level. The specific interim assessment designs are driven by the purposes and intended uses, but the results of any interim assessment must be reported in a manner allowing* ***aggregation across students, occasions, or concepts*** *(Perie, Marion, & Gong, 2009).*

Formative Assessment

- *Formative assessment is a* ***process*** *used by teachers and students* ***during instruction*** *that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes* (FAST SCASS, 2006, emphases added).

Tiers of a Comprehensive Assessment System

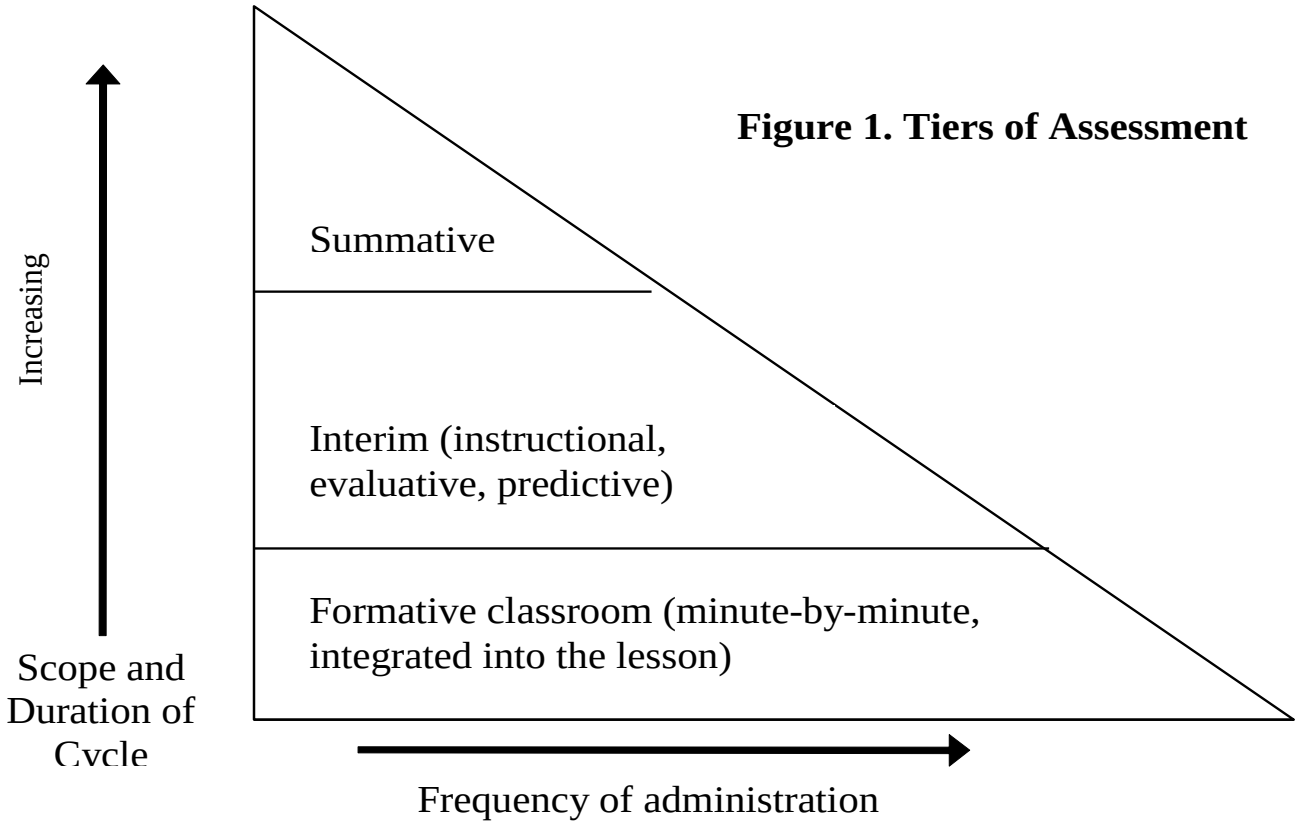


Figure 1. Tiers of Assessment

Continua

- We argue that formative, interim, and summative assessment are not as distinct as some advocate (or as certain picture portray), but can be thought of as being on a continuum or multiple continua such as:
  - Intended and actual uses
  - Timing (related to curriculum and instruction)
  - Types of items/tasks (designed to provide summary information or insight into student learning)
  - Form of the results and feedback (summaries, descriptions

Back to Use

- Accountability (including educator evaluation) is about judgments, actions, and perhaps consequences
- Assessment is about collecting information that can be used in accountability systems (or not)
- Many people who complain about large-scale assessment systems are really just "shooting the messenger"
  - The concern is really with the accountability uses
- So as we go forward with design decisions, let's keep these distinctions in mind

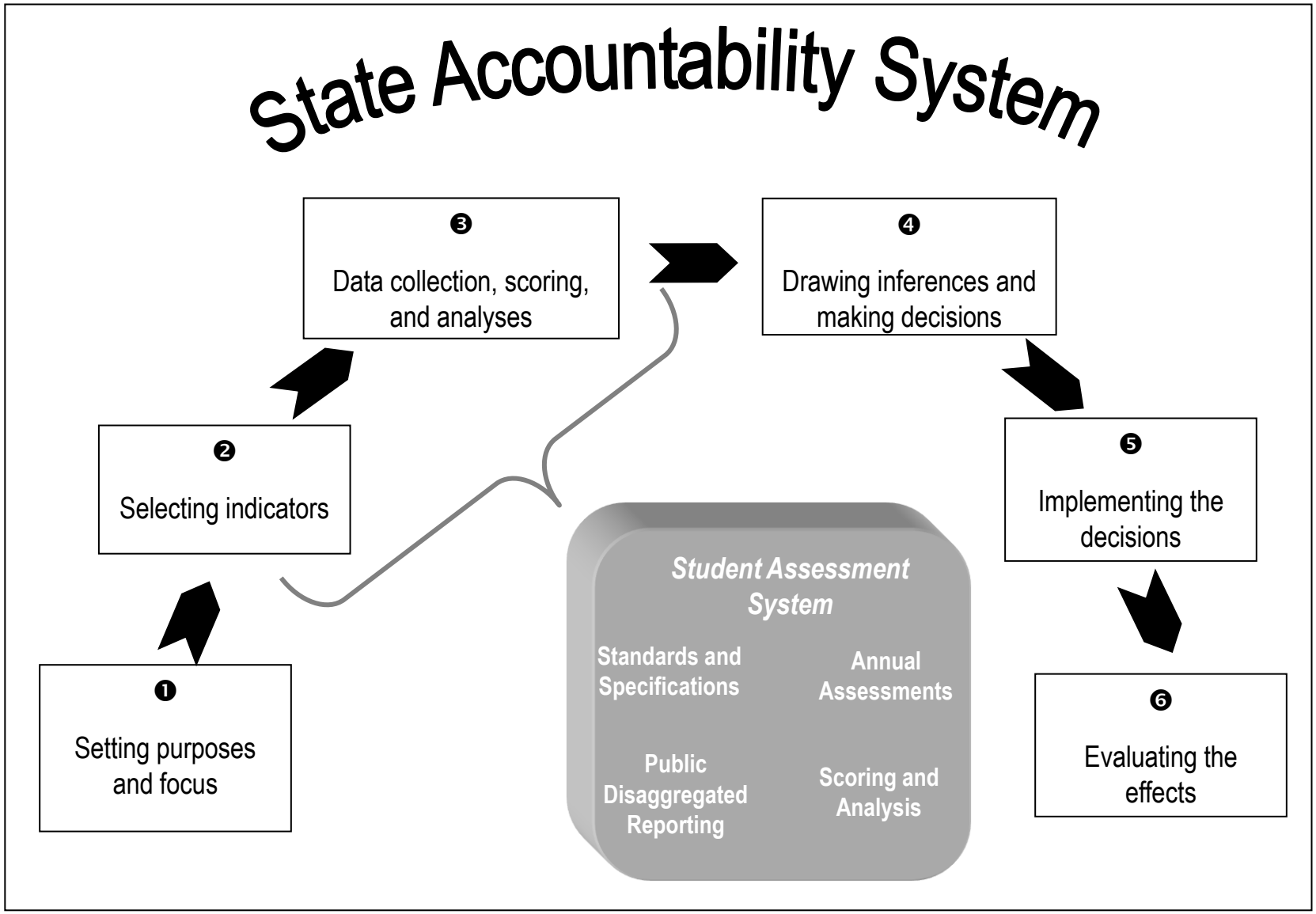
Relationship Between Assessment & Accountability

- A high quality and technically defensible assessment system is a necessary condition for the development of a valid accountability system…
- but it is not sufficient!

- Accountability (including educator evaluation) systems include more data and many decisions that go beyond the assessment information.

- Dale Carlson created the following diagram about 10 years ago to help explain this intertwined relationship

The relationship of assessment & accountability validity



Assessment & Analytics

- In the following slides, we will go into considerable detail about various types of assessment and analytical approaches for educator evaluation in NTSG
- The system can only be valid if BOTH the assessment AND analytic approaches are valid
- As hard as it is to get the assessment piece right, the analytics invite even more opportunities for invalid inferences

MEASUREMENT APPROACHES FOR NON-TESTED SUBJECTS AND GRADES

Starting with Claims about NTSG

- We present several approaches that states and others are using to both measure student achievement and calculate "growth"
- We also present several general opportunities and challenges associated with each approach
- In order to weigh the specific opportunities and challenges for your context, you need to consider these approaches in light of the claims your state wants to be able to make, such as:
  - Measures of student performance will allow for normative comparisons educators within each school district, and/or
  - Measures of student performance will allow for evaluations of educators against a specific statewide criterion

Comparability

- What do we mean by comparability in this context?
  - Educators within the units of analysis are held to similar levels of expectations, at least in some relative sense
  - For example, it would be a threat to the system if the teachers in grades 4-8 reading and math received noticeably lower ratings than the rest of the teachers (NTSG) in the school
- At what levels is comparability important?
  - Within schools? Clearly yes.
  - Within districts? Probably yes.
  - Across districts? It would be nice, but it might be too high of a bar right now.

What Measurement Approaches Are Being Proposed?

1. Norm-referenced tests (NRTs)
2. Commercial interim assessments
3. State or district created end-of-course exams (both externally and locally developed)
   1. Includes new assessment development in places like DE, CO, Hillsborough, FL
4. School or teacher-developed measures of student performance

*Note: 1 & 2 rarely cover courses beyond the four core content areas and even then, not well in HS.

Relying on/adding NRTs

- Opportunities:
  - Allows for reasonable quality assessments beyond most commonly assessed grades (3-8; 11)
  - Multiple options for pre-post designs
  - Many are built using a vertical scale that <u>may</u> support some growth determinations
- Challenges:
  - Not often well aligned to the state content standards—will threaten validity of score inferences and send mixed instructional signals
  - Limited transparency of development, equating and quality assurance processes
  - Can be costly to implement, but perhaps not as costly as other options

Using Commercial Interim Assessments

- Opportunities:
  - Multiple assessment windows—could allow for pre/post within the same course
  - Depending on vendor, could have high levels of security
- Challenges:
  - Problems with technical quality and lack of technical information for most options
  - Purportedly designed for purposes other than summative accountability (e.g., Perie, et al, 2009: predictive, evaluative, instructional)
  - Not often well aligned to the state content standards—will threaten validity of score inferences and send mixed instructional signals
  - Can be costly to implement

State or district created tests

- Opportunities:
  - Could create high quality assessments—aligned to relevant standards—in many subject areas
  - Could provide a PD vehicle for developing assessment literacy if done as a "grassroots" project (e.g., NYC, CO)
- Challenges:
  - If done with a vendor, it will cost an **enormous amount of money**
  - If done without a vendor, it will require high levels of in-state capacity and quality could be a problem
  - Must avoid an item bank approach
  - Adding significant amounts of external testing may be seen as a burden by schools and districts
  - Potentially very corruptible
  - Most of the models we've seen do not include ongoing development and if they do, linking designs are weak or non-existent

School/teacher created assessments

- Opportunities:
  - Could provide as a PD vehicle for developing assessment literacy
  - Should allow for assessments better aligned with specific curriculum & instruction
  - Could permit the measurement of more complex performances (e.g., NYC)
  - Might not require "additional tests"
- Challenges:
  - Quality of educator-created assessments is often very low
  - Comparability will be a significant concern with this approach
  - Potentially even more corruptible than other approaches
  - What about the 2nd and 3rd years (ongoing development)?

What about classroom assessment?

- It is <u>not</u> correct to assume that classroom assessment = formative assessment. Most often that is not the case.
- A classroom assessment <u>system</u> should serve multiple purposes such as formative and summative (e.g., awarding grades)
- As such, classroom assessments can play an accountability role…for student accountability. We have much less experience and many more worries when using classroom assessments for educator accountability.

More thoughts about "adding tests"

- The issues of cost and quality are much more obvious to you than to policy audiences
- We must consider massive extensions of state testing
- If the measures are locally developed, it raises concerns about comparability, quality, and corruptibility.
- I'm not saying that we should avoid adding any new tests, but it will be very difficult to test our way out of the problem
- We should always consider having the right tests for the right purposes (back to our claims)

ANALYTIC APPROACHES FOR DOCUMENTING STUDENT "GROWTH" IN NON-TESTED SUBJECTS AND GRADES

Analytical Approaches

- If you thought the measurement/assessment issue was daunting….
- It pales in comparison to the analytic challenges (i.e., how growth is calculated at local levels)

- Remember, using the most sophisticated VAM models **with high quality state test data** has been rightfully questioned based on challenges with causal inferences, unreliability (year-to-year), and other technical issues (e.g., EPI report, Braun, et al., 2010, Rothstein, 2009 & 2010)

What Approaches Are Being Proposed for NTSG?

1. Growth models using pre and post test from the same subject
2. Value-added models
   a. Pre and post test score in the same subject
   b. Conditioned on data other than pretest from same content area as posttest
3. Student Growth Percentiles
4. Shared attribution of aggregate growth/VAM results
5. Student learning objectives (SLO)

Definitions

- **Growth** refers to measures of performance for the same students at two or more points in time and requires a common, often vertical, scale to evaluate the magnitude of change. **Only true growth** model here.
- **VAM**: Generally describes multivariate models that include certain variables to produce to an expectation against which actual performance is evaluated.
- **Student Growth Percentiles (SGP)** is a regression based measure of growth that works by evaluating current achievement based on prior achievement and describing performance (using percentiles) relative to other students with the "same" prior achievement histories.
- **Student Learning Objectives (SLO)** is a general approach (often called Student Growth Objectives) whereby educators establish goals for individual or groups of students (often in conjunction with administrators) and then evaluating the extent to which the goals have been achieved.

"Growth" models

- Opportunities:
  - Intuitively appealing because it appears to be simple and familiar
  - Appears to be a fair measure of how learning has occurred throughout the instructional period

- Challenges:
  - While intuitively appealing, hard to interpret without context (e.g., conditioned on prior scores, similar students, and other factors)
  - Requires a vertical scale, pseudo vertical scale, or the same scale (e.g., using alternate forms of the same test as pre-post) and these are unlikely to be available for most non-tested areas
  - Assumes interval vertical scale, which few tests have, unless only concerned with ordinal change

VAM—same content area prior scores

- Opportunities:
  - Seen as producing more fair and defensible inferences than simple growth models because student scores are conditioned on where students start as well as other influences on posttest scores

- Challenges:
  - Much less reliable when only one prior score is used, which will likely be the case with many NTSG
  - When demographic factors are included, need to explain that different standards are established for teachers with different groups of students
  - VAM is often considered to permit causal attribution, but much evidence suggests that such **causal claims are unwarranted**
  - Generally large sample sizes are required to obtain stable and meaningful inferences (e.g., referent for "center" of distribution)
  - High quality assessments with good scale properties are necessary for appropriate inferences

VAM—conditioned on "other" information

- A subset of VAM, but used when only a common posttest (e.g., EOC) is available, but no true pretest. Conditioning is based on other information such as tests in other subjects (e.g., reading/math), demographic factors, or other influences on posttest scores

- Opportunities:
  - Similar to VAM, can provide information about how students performed relative to prediction

- Challenges:
  - Same challenges as VAM with the added challenges:
    - Concerns about weaker statistical relationship among "priors" and posttest
    - Concerns about "face validity" among educators of being held accountable without a "true prior" and perhaps lack of common priors

Student Growth Percentiles

Opportunities:

- Provides a descriptive tool for understanding growth levels achieved by students sharing same baseline starting point
- Allows for normative comparisons to be made across various units of aggregation (e.g., schools and districts)

Challenges:

- Student growth percentiles require large samples in order to appropriately establish "academic peers"
- Unless districts are quite large (e.g., several thousand students/grade), SGPs might not be the most appropriate tool for calculating student growth for NTSG
- Similar to point made with VAM, less reliable when only one prior score available

Shared Attribution

- Opportunities:
  - Encourages collaboration among groups of educators in school
  - Based on statewide standardized data and generally most technically defensible analytical methods
  - **Might make much more sense for grade-level teams, for example, than full school shared attribution**

- Challenges:
  - Masks true variability among educators
  - Requires accountability decisions on certain educators when they may have limited influence over student test results
  - Has not been verified for high stakes personnel evaluations

Student Learning Objectives

- Opportunities:
  - Can be inclusive of all educators
  - Can incentivize appropriate educational behaviors
    - Setting meaningful goals, providing opportunities for feedback and collaboration across educators sharing goals, monitoring progress toward those goals, and evaluating the extent to which those goals are achieved
- Challenges:
  - Will require significant PD and oversight to establish meaningful and *comparable* goals
  - Potential for corruption on both goals and measures
  - Still requires high quality measures, at least for posttest
  - Using as a "growth" measure suffers from the same limitations as all of the previous

MEASUREMENT & EVALUATION RECOMMENDATIONS

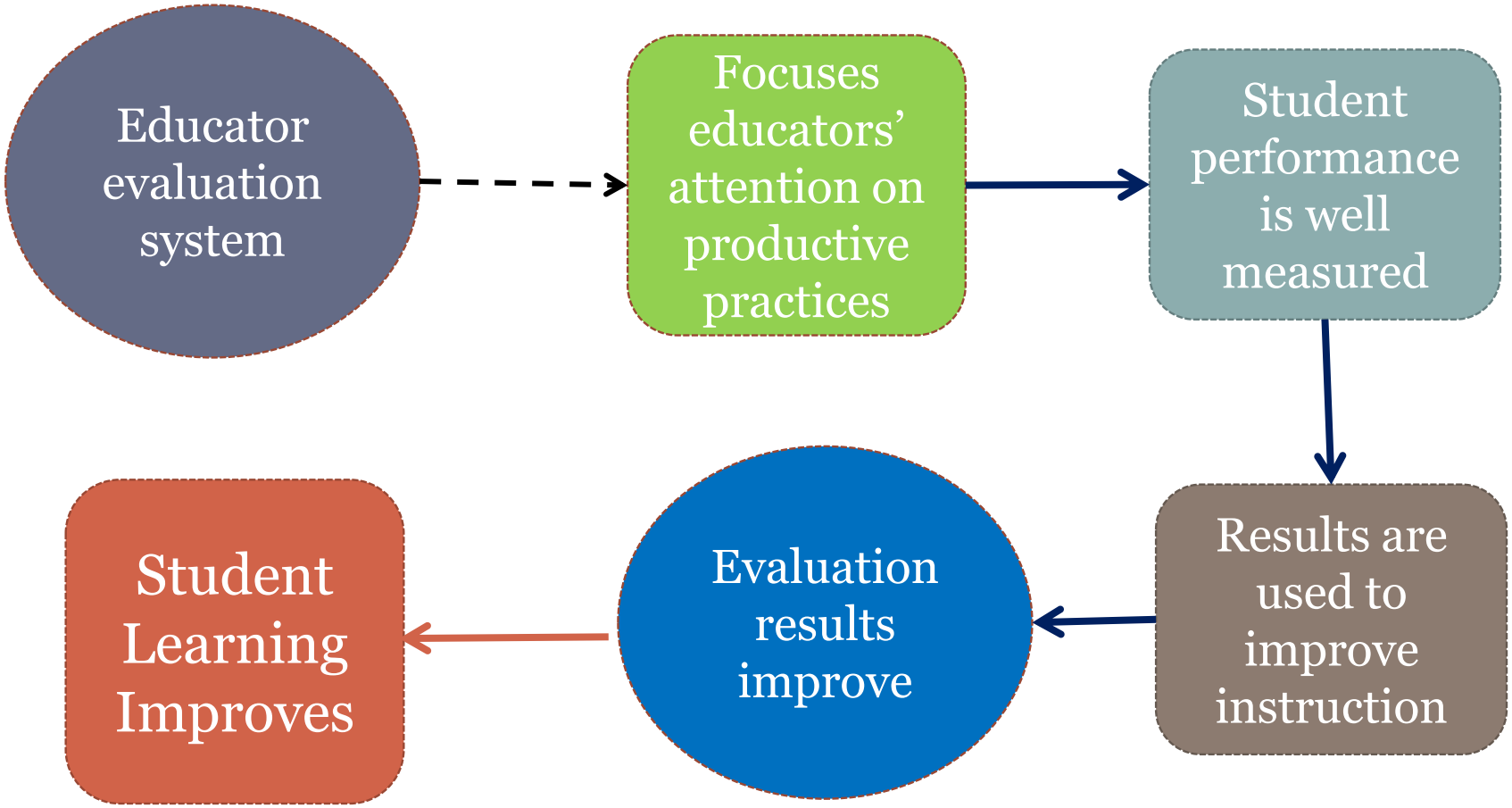
Measurement & Evaluation Recommendations

- So what do we recommend given all of these opportunities and challenges
  - Theory of Action
  - A Student Learning Objective Framework
  - Evaluation & Continuous Improvement

Theory of Action

- I often recommend developing a theory of action for many reasons including…
  - clarifying the intended uses and making clear the mechanisms required for such uses to be fulfilled,
  - developing a validity argument and validity evaluation plan,
  - revealing opportunities for corruption, and
  - illuminating potential irresolvable conflicts in proposed uses.

An oversimplified theory of action



Thinking Through a Theory of Action

- Policy makers should have to very explicitly say why and **how** implementing test-based approaches to support educator effectiveness for these grades and subjects will lead to improved educational opportunities for students
  - For example, one might postulate that holding teachers accountable for increases in student test scores on classroom-based assessments will lead to the development of both better assessments and improvements in student learning.
- What are the specific mechanism(s) by which the intended outcomes will occur?
  - E.g., targeted instruction, better PD, and/or more appropriate curricular materials?

Campbell's Law

- "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and **corrupt the social processes it is intended to monitor**." (emphases added)

http://en.wikipedia.org/wiki/Campbell%27s_Law

- **Educator accountability systems will invite significantly more implicit and explicit corruption than has been seen with school accountability**

Theory of Action and Campbell

- Too often theories of action turn out to be pretty pictures of how systems <u>should</u> work
- A theory of action should also be used to identify ways in which the system inadvertently incentivizes or invites corruption
- For instance, the earlier TOA example indicated that, "*results will be used to improve instruction*," but we should also be alerted to possibility that educators will look to improve evaluation results via unseemly means
  - A theory of action should anticipate and try to specify these possibilities

Student Learning Objectives as a Framework

- As we work in this area, we need to strive toward building a comprehensive and thoughtful approach that includes the tested subjects/grades, the "non-tested" content area teachers, and other licensed professionals

- "Tested" and "non-tested" subjects and grades can then be viewed as special cases of the comprehensive framework

Claims to evaluate SLOs

- A theory of action is useful for crafting a validity argument
- More formally, we can create claims for SLOs and then consider the challenges and support for these claims
- We present some examples to illustrate this point, but this would have to be done in more detail prior to implementing such an approach on a large scale

Claims, challenges, opportunities

- Claim: *Teachers have the knowledge, skills, and attitudes (& ethics) to set meaningful, ambitious, and fair goals for students*
- Challenge: Who will guide, monitor, and/or evaluate the quality of these goals?
  - This adds an extra (or at least different) significant validation requirement beyond test-based approaches
  - Places principals into the role of instructional leaders and many might not have the skills (but this could be an opportunity if successful)
- Opportunity: Teaching quality would likely improve if teachers, working with good leaders, were supported in the way they use data to establish goals for their students.

Claims, challenges, opportunities

- Claim: *Teachers have the knowledge and skills to tailor learning opportunities for their students*
- Challenge: Will there be a temptation to limit the range/variability of the goals to maximize efficiency?
- Opportunity: If teachers were really expected to focus on the needs of individual students, learning opportunities could very well improve. Would using group instead of individual goals limit this opportunity?
  - Undoubtedly high school teachers will have to set group goals (150 vs. 30 students)

Claims, challenges and opportunities

- Claim: *Teachers and/or others have measurement or evaluation procedures sufficient for judging whether students have reached the intended goals*
- Challenge 1: Are classroom assessment tools capable of validly measuring ambitious goals?
- Challenge 2: If external assessments are used, would that lead to narrow goals to match the more limited tools (tail wagging the dog)?
- Opportunity: Could this be a lever for improving the quality of classroom assessment and evaluation tools and processes?

SLOs as a framework

- Paraphrasing Churchill…
- SLOs are the worst possible approach for NTSG,
- …except for all of the others…

- This would put all educators in the same boat, while those in "tested" subjects could use data from the state test to evaluate their goals
- We still need a lot work in this area—e.g., how are goals set, who approves them?—but it appears to be a promising starting point

Evaluation & Improvement

- Given the magnitude and novelty of what we are trying to do, it is critical to conduct a comprehensive evaluation of the system and policy.
  - **Formative evaluation** (Scriven) must be employed to help collect data on implementation so early course corrections can be made before things get derailed
  - A thorough **summative evaluation** should be designed, based in large part on the theory of action, to evaluate the full system. It makes sense to use a different evaluator than the one conducting the formative evaluation
- For now, we suggest one approach for beginning to evaluate NTSG systems

Evaluation and Improvement

- A recent report from the Brookings Institution offers a useful framework for conceptualizing how one might evaluate educator evaluation systems, particularly for those educators in non-tested subjects and grades
  - Glazerman, et al. (2011, April). *Passing Muster: Evaluating Teacher Evaluation Systems*. Washington, DC: The Brookings Brown Center Task Group on Teacher Quality. Retrieved on May 22, 2011 from: http://www.brookings.edu/reports/2011/0426_evaluating_teachers.aspx

Evaluation & Improvement

- We question much of the report's specifics, but their conceptualization of an evaluation makes sense…
- Aggregate VAM scores in subsequent years can be used as one criterion for evaluating the capacity of educator evaluation systems to "predict" teacher effectiveness
  - Of course we can and should question whether VAM is the right criterion
- Employing "non-tested approaches" in "tested" subjects & grades can allow us to compare the results of VAM/SGP scores with SLO results

Technical Quality

- Many states and others are beginning to examine the technical quality of measures used in student growth determinations
- Most of what we have seen focuses on traditional aspects of assessment quality
- This is a good start because **you can't measure growth if you can't measure status well**
- But it is not enough! We must also evaluate how these measures hold up for calculating "growth" as well as evaluating the analytic methods themselves
  - We are currently working on such an evaluation system

Next Meeting

- We will begin examining a framework for evaluating the technical quality of assessments and analytic approaches used in NTSG
- We will also propose a process for conducting and supporting such evaluations

Final Thoughts

- If anyone tells you this is easy, they are either trying to sell you something or they don't understand the challenges
- We have to be honest and humble in what we tell policy makers what we can do
- We should always and only focus on approaches in this area that have a chance of supporting teaching and learning