



# Formative Assessment: A Critical Review

*Randy Bennett*

*ETS*

*[rbennett@ets.org](mailto:rbennett@ets.org)*

Staff seminar at RAMA, Tel Aviv, Israel, October 25, 2009

*Listening.  
Learning.  
Leading.*

## Outline

- Six issues
  - The definitional issue
  - The effectiveness issue
  - The domain issue
  - The measurement issue
  - The professional development issue
  - The system issue
- Summary



## What's "Formative Assessment?"

- **Scriven (1967)**
  - **Summative evaluation**
    - Provided information to judge the overall value of an educational program
  - **Formative evaluation**
    - Targeted at facilitating program improvement



## What's "Formative Assessment?"

- Bloom (1969)
  - Summative evaluation
    - Judge what the learner had achieved at the end of a course or program
  - Formative evaluation
    - “... *to provide feedback and correctives at each stage in the teaching-learning process.*”

B. Bloom,  
*Educational Evaluation*,  
1969, p. 48

## “Test Industry Split Over 'Formative' Assessment”



“Testing expert Richard J. Stiggins [of ETS] says he has stopped using the term *formative assessment*.”

*Education Week*, 28(4),  
Sept 17, 2008



## One Side of the Split

- It's an instrument
  - A diagnostic test
  - An “interim” assessment
  - An item bank



## Item Banks

[Overview](#)

[Item Development](#)

[Frameworks and Blueprints](#)


[Frequently Asked Questions](#)

[Contact Us](#)



### Related Links

[K-12 Learning & Development](#)

 [Rate our website.](#)

## ETS Formative Assessment Item Bank

This item bank contains more than 50,000 items that measure mathematics, reading and writing for kindergarten through grade 12 and science for grades 3 through high school.

The majority of items are multiple-choice, but the bank also contains some short and extended constructed-response items. The items are aligned across multiple states and were checked to match standards for the states of:

- |            |               |
|------------|---------------|
| California | New Jersey    |
| Florida    | Nevada        |
| Hawaii     | New York      |
| Indiana    | Texas         |
| Ohio       | Virginia      |
| Maryland   | West Virginia |

For states without specific state alignments the items have been aligned to a [National Framework](#) developed by ETS.



## The Other Side

*“... formative assessment is not a test but a process...”*

W. J. Popham,  
*Transformative Assessment*,  
2008, p.6





*“Such assessment becomes formative assessment when the [results are] actually used to adapt the teaching to meet student needs.”*

P. Black & D. Wiliam,  
*Phi Delta Kappan*, 1998, p. 2

*“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.”*

S. McManus,

*Attributes of Effective Formative Assessment,*  
2008, p. 3



## Popular Rendition

- It's a process
  - As long as the results are used to change instruction, *any* instrument may be used formatively, regardless of its original intended purpose

# The Definitional Issue

- Each position is an oversimplification
  - It's an instrument
    - The most carefully constructed, scientifically supported instrument is unlikely to be effective instructionally if the process surrounding its use is flawed
  - It's a process
    - The most carefully constructed process is unlikely to be effective instructionally if the “instrumentation” is not well-suited for the purpose

## Alternative Terminology

- *Assessment for learning = formative assessment*
- *Assessment of learning = summative assessment*



# Terminology

- **Summative assessment**
  - Primary purpose: documenting what students know and can do
  - Secondary purpose: Supporting learning



# Terminology

- Formative assessment
  - Primary purpose: suggesting how instruction should be modified
  - Secondary purpose: suggesting what students know and can do

	Assessment <i>of</i> Learning	Assessment <i>for</i> Learning
Summative	X	
Formative		X



## By Careful Design

	Assessment <i>of</i> Learning	Assessment <i>for</i> Learning
Summative	<i>X</i>	<i>x</i>
Formative		<i>X</i>

## By Careful Design

	Assessment <i>of</i> Learning	Assessment <i>for</i> Learning
Summative	<i>X</i>	<i>x</i>
Formative	<i>x</i>	<i>X</i>

Note. X = primary purpose; x = secondary purpose.

## Moving Toward Definition

- Definition is important
  - If we can't clearly define it, we can't:
    - Document its effectiveness
    - Meaningfully summarize across effectiveness studies
    - Transport it to our own context



## Moving Toward Definition

- Definition presumes:
  - A conceptual framework
  - An action theory
  - A concrete instantiation



## *Keeping Learning on Track<sup>®</sup>* Program (KLT)

- An attempt to define formative assessment
  - Conceptual framework
    - One “big idea” and five key strategies
      - Big idea: *students and teachers using evidence ...to adapt teaching and learning to meet immediate learning needs minute-to-minute and day-by-day.*

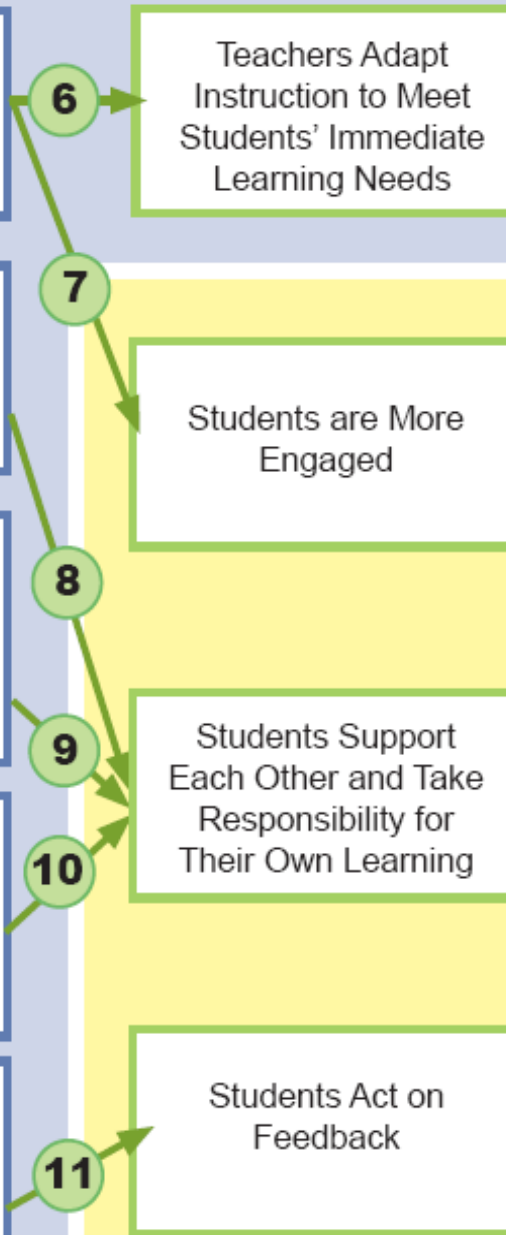
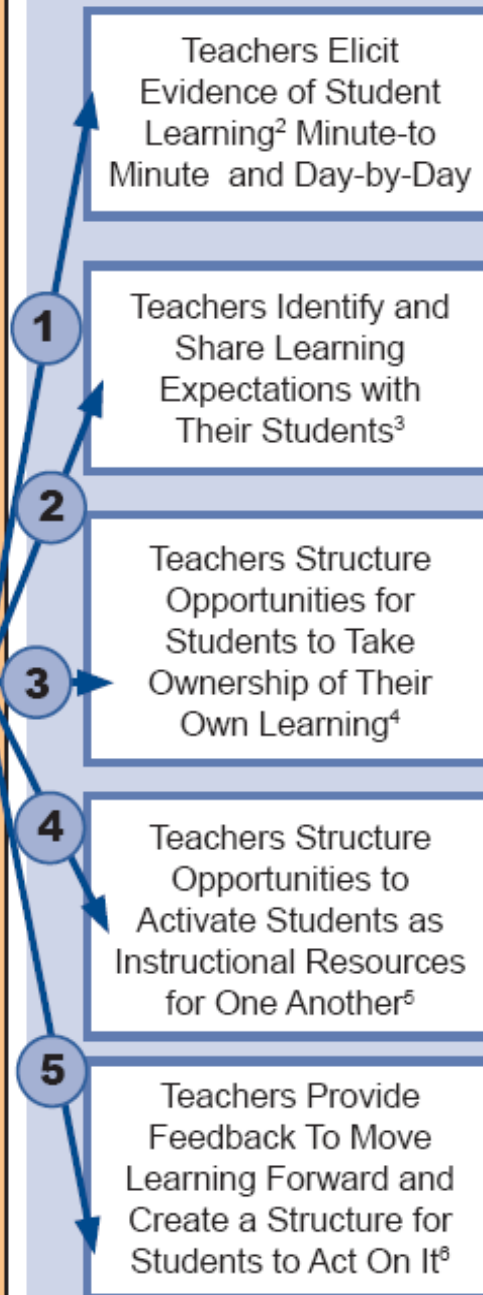
# *KLT*: Five Key Strategies

- **Sharing Learning Expectations**
  - Clarifying and sharing learning intentions and criteria for success
- **Questioning**
  - Engineering effective classroom discussions, questions and learning tasks that elicit evidence of learning
- **Feedback**
  - Providing feedback that moves learners forward
- **Self Assessment**
  - Activating students as the owners of their own learning
- **Peer Assessment**
  - Activating students as instructional resources for one another

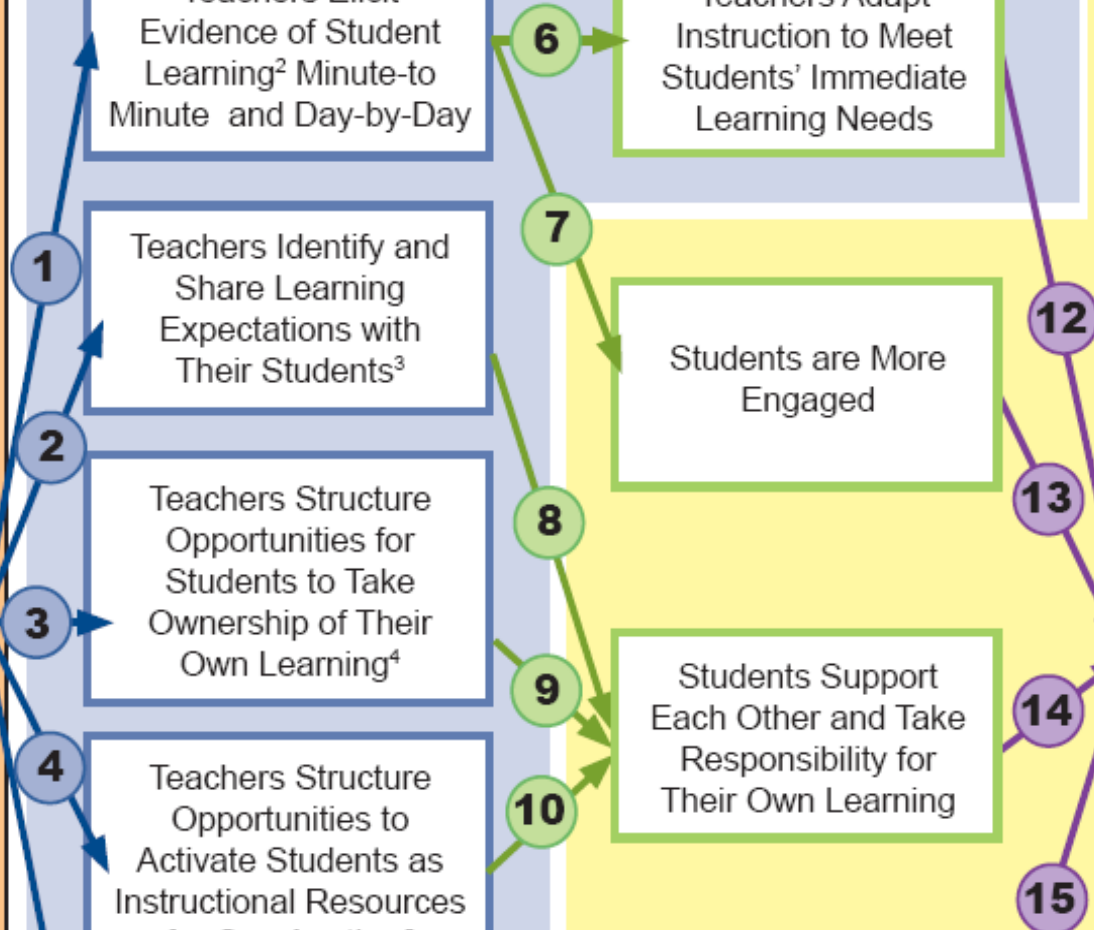
## KLT Components

- Introductory professional development in AfL, including:
  - ✓ the research basis,
  - ✓ the theoretical framework,
  - ✓ practical strategies and techniques for implementation, and
  - ✓ a process for planning changes to current practice
- On-going monthly TLC meetings, focused on AfL including time and structure to:
  - ✓ Report on progress
  - ✓ Trouble shoot with colleagues
  - ✓ Plan for future changes
- On-going support including:
  - ✓ On-going dialogue with ETS consultants
  - ✓ Year 2 workshops
  - ✓ Ancillary materials

## Teacher Outcomes



## Student Outcomes





## The Concrete Instantiation

- Some of KLT's components
  - Teacher Learning Community (TLC)  
Leaders Workshop
  - 16 Modules that form a 2-year curriculum for TLCs
  - Participant workbooks
  - Guidebook for TLC Leaders



## Outline

- Six issues
  - The definitional issue
  - **The effectiveness issue**
  - The domain issue
  - The measurement issue
  - The professional development issue
  - The system issue
- Summary

## Claims: Example 1

*“Based on their meta-analysis, Black and Wiliam [1998] report effect sizes of between .4 and .7 in favor of students taught in classrooms where formative assessment was employed.”*

W.J. Popham,  
*Transformative Assessment,*  
2008, p.19.



## Claims: Example 2

*“English researchers Paul Black and Dylan Wiliam recently published the results of a comprehensive meta-analysis and synthesis of more than 40 controlled studies of the impact of improved classroom assessment on student success ...”*

R. J. Stiggins,  
*Phi Delta Kappan*, 1999



## Claims: Example 3

*“Black and Wiliam, in their 1998 watershed research review of more than 250 studies from around the world on the effect of classroom assessment, report gains of a half to a full standard deviation.”*

R. J. Stiggins,  
*EDge*, 2006, p. 15



## Claims: Example 3 (con't)

*“Bloom and his students (1984) made extensive use of classroom assessment ... for learning ... [and] reported subsequent gains in student test performance of one to two standard deviations.”*

R. J. Stiggins,  
*EDge*, 2006, p. 15

# Research on Effects of Formative Assessment on Student Learning

■ Bloom (1984)	1.0 to 2.0 *
■ Black & Wiliam (1998)	.5 to 1.0 **
■ Meisels, et. a. (2003)	.7 to 1.5
■ Rodriquez (2004)	.5 to 1.8 **

\* Rivals one-on-one tutoring

\*\* Largest gains for low achievers

# Research on Effects of Formative Assessment on Student Learning

■ Bloom (1984)	1.0 to 2.0 *
■ Black & William (1998)	.5 to 1.0 **
■ Meisels, et. a. (2003)	.7 to 1.5
■ Rodriguez (2004)	.5 to 1.8 **

\* Rivals one-on-one tutoring

\*\* Largest gains for low achievers

## The Effectiveness Claims

- Empirical research proves “formative assessment” causes medium-to-large achievement gains
- These results come from trustworthy sources:
  - Rigorous meta-analyses
  - Noteworthy individual studies



## Meta-Analysis

- A pooling of results from a set of comparable studies that yields one or more summary statistics
  - Effect size: the difference between the treatment-group and control-group means, divided by the standard deviation

# Meta-Analysis

- The results of meta-analysis should be considered suspect when:
  - Studies are too disparate in topic to make summarization meaningful
  - Multiple effects too often come from the same study or authors
  - Study characteristics are not considered
  - The meta-analysis itself is not published so that the methods used are unavailable for critical review

## The Black and William Review

- The research covered in the *Assessment in Education* article is too disparate to be summarized meaningfully through meta-analysis
  - Includes studies:
    - Related to feedback, student goal orientation, self-perception, peer assessment, self assessment, teacher choice of assessment task, teacher questioning behavior, teacher use of tests, and mastery learning systems
    - Too diverse to be sensibly combined and summarized by a single effect-size statistic

*“No Meta-analysis*

*It might be seen desirable... for a review of this type to attempt a meta-analysis of the quantitative studies that have been reported... Individual quantitative studies which look at formative assessment as a whole do exist..., although the number with adequate and comparable quantitative rigour would be of the order of 20 at most. However, whilst these [studies] are rigorous within their own frameworks and purposes, ... the underlying differences between the studies are such that any amalgamations of their results would have little meaning.”*

P. Black and D. Wiliam,  
*Assessment in Education*, 1998, p. 53



## Further Sources for Effectiveness Claims

- Bloom, B. S. (1984). "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring." *Educational Researcher* 13(6), 4–16.
  - Based largely on dissertations by Bloom's students



## Further Sources for Effectiveness Claims

*“Bloom’s claim that mastery learning can improve achievement by more than 1 sigma is based on brief, small, artificial studies that provided additional instructional time to the experimental classes [and not to controls]. In longer term and larger studies with experimenter-made measures, effects of group-based mastery learning are much closer to 1/4 sigma, and in studies with standardized measures there is no indication of any positive effect at all. [The]1-sigma claim is misleading ... and potentially damaging ... as it may lead researchers to belittle true, replicable, and generalizable achievement effects in the more realistic range of 20-50% of [a] standard deviation.”*

R. J. Slavin,  
*Review of Educational Research*,  
1987, p. 207



## Further Sources for Effectiveness Claims

- Nyquist, J. B. (2003). *The Benefits of Reconstructing Feedback as a Larger System of Formative Assessment: A Meta-analysis*. Nashville, TN: Vanderbilt University.
  - College-level students
  - Unpublished master's thesis



## Further Sources for Effectiveness Claims

- Meisels, S. J., Atkins-Burnett, S., Xue, Y., Bickel, D. D., & Son, S. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives, 11(9)*.
  - Used a volunteer treatment group
  - Collected data in classrooms that may have been simultaneously implementing other curricular innovations



## Further Sources for Effectiveness Claims

- Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education, 17*, 1-24.
  - No clear interpretation possible regarding a cause-effect relationship between formative assessment and student achievement
  - *Negative* relation between the use of teacher made tests and student achievement

## Further Sources for Effectiveness Claims

- Kluger, A. N., & DeNisi, A. (1996). The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-analysis, and a Preliminary Feedback Intervention Theory. *Psychological Bulletin* 119, 254–284.
  - A (real) meta-analysis of a large number of studies in a *very* high-quality journal focused on *one* topic relevant to formative assessment
  - Mean effect size = .41
  - 38% of effects were *negative*



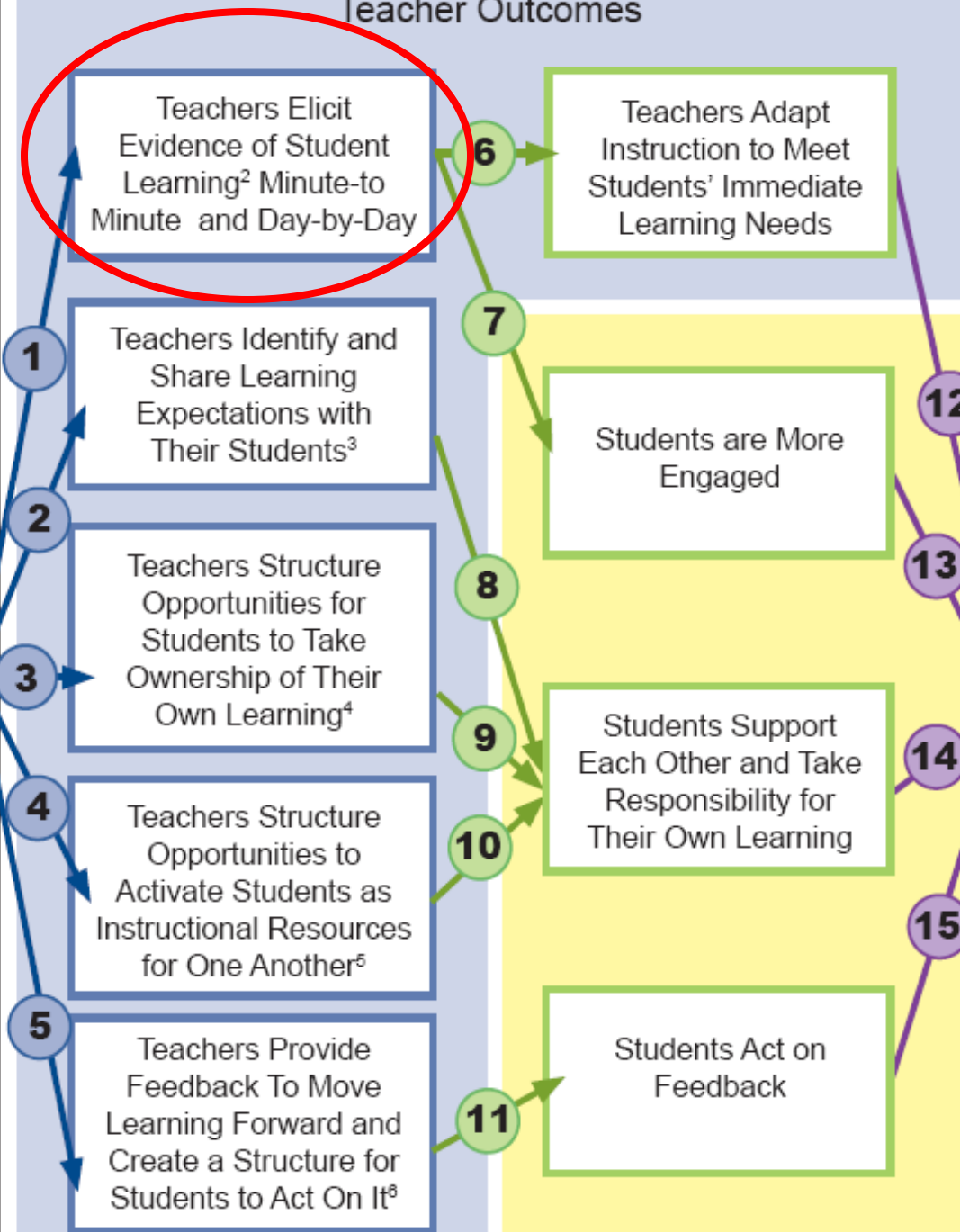
## Improving Our Claims

- Without the action theory, we can't meaningfully evaluate the underlying mechanisms that are supposed to cause the intended effects

## KLT Components

- Introductory professional development in AfL, including:
  - ✓ the research basis,
  - ✓ the theoretical framework,
  - ✓ practical strategies and techniques for implementation, and
  - ✓ a process for planning changes to current practice
- On-going monthly TLC meetings, focused on AfL including time and structure to:
  - ✓ Report on progress
  - ✓ Trouble shoot with colleagues
  - ✓ Plan for future changes
- On-going support including:
  - ✓ On-going dialogue with ETS consultants
  - ✓ Year 2 workshops
  - ✓ Ancillary materials

## Teacher Outcomes



## Student Outcomes



## Improving Our Claims

- If the inferences about students resulting from formative assessment are wrong, the basis for adjusting instruction is undermined
- If the inferences are correct but instruction is adjusted inappropriately, learning is less likely to occur

## Two Arguments

- Formative assessment requires:
  - A *Validity Argument* to support the quality of inferences about students and the adjustments to their instruction
  - An *Efficacy Argument* to support the impact of the inferences and adjustments
- Each argument requires backing, both logical and empirical



# The Validity Argument

- Assert that formative assessment facilitates:
  - Inferences about student strengths and weaknesses
  - Related instructional adjustments
- Offers backing for the reasonableness of the inferences and adjustments
  - Resulting inferences and adjustments are similar to those that an expert teacher would make

# The Efficacy Argument

- Asserts that the use of formative assessment improves students' knowledge and skills
  - This improvement is caused by actions the teacher (or student) takes based on assessment inferences
- Offers backing for knowledge and skill gains
  - Empirical research comparing formative assessment to some alternative treatment



## Outline

- Six issues
  - The definitional issue
  - The effectiveness issue
  - **The domain issue**
  - The measurement issue
  - The professional development issue
  - The system issue
- Summary

## The Domain Issue

- General and specialized knowledge function in close partnership (Salomon & Perkins, 1989)
  - Domain-independent strategies are broadly useful but weak
  - Domain-specific knowledge is powerful but brittle

## The Domain Issue

- To be maximally effective, formative assessment requires the interaction of:
  - General principles, strategies, and techniques *with*
  - Deep cognitive-domain understanding
    - Processes, strategies and knowledge important for proficiency
    - Habits of mind that characterize the community of practice
    - Features of tasks that engage those elements

## Two Implications

- A teacher with weak cognitive-domain understanding is less likely to know:
  - What questions to ask
  - What to look for
  - What inferences to make
  - What actions to take
- The specifics of formative assessment may differ significantly from one domain to the next

## A Possible Approach

- Conceptualize and instantiate formative assessment within the context of specific domains
  - Cognitive-domain model to guide the substance of formative assessment
  - Learning progressions to indicate steps toward mastery
  - Tasks to provide evidence of student standing
  - Techniques tuned to the substantive area
  - Process suited to the materials and domain

# An Example in Reading

- Cognitive-domain model
  - Ability to use and understand text conventions
- Hypothesized learning progression for literary text
  - (1) determine the basic idea of plot,
  - (2) identify key plot elements (e.g., climax, resolution)
  - (3) understand how events advance the author's goals
- Tasks
  - Examples of literary text
  - Questions that tentatively place each student
- Domain-specific techniques
  - Graphic organizers for identifying plot elements to be completed by students for literary text the teacher assigns

## Outline

- Six issues
  - The definitional issue
  - The effectiveness issue
  - The domain issue
  - **The measurement issue**
  - The professional development issue
  - The system issue
- Summary

## The Measurement Issue

- Educational measurement involves:
  - Designing opportunities to gather evidence
  - Collecting the evidence
  - Interpreting the evidence
  - Acting on interpretations





## Formative Assessment as an Inferential Process

- We can't *know* what understanding exists inside a student's head
- We can only make conjectures
- Backing for the meaning of our conjectures is stronger to the extent we observe reasonable consistency
- Each teacher-student interaction is an opportunity for posing and refining our conjectures

## “Formative Hypothesis”

*“I see a strong connection between ... formative assessment practices... and my training as a clinician when I used observations to form a tentative hypothesis, gathered additional information to confirm or revise, and planned an intervention (itself a working hypothesis).”*

L. A. Shepard,

*Educational Measurement*, 2006, p. 642

## “Formative Hypothesis”

*“By examining ... student work..., the teacher can form hypotheses about the student’s competencies and about gaps in ... understanding ... If a particular set of conjectures ... does account for the student’s pattern of performance (including mistakes), and no plausible alternative hypothesis does as well, the proposed conjectures can be accepted as a reasonable conclusion about the student.”*

M. T. Kane,  
*Educational Measurement*, 2006, p. 49



## Errors, Slips, Misconceptions, and Lack of Understanding

- **Error:**
  - What we *observe* students make--  
some difference between a desired  
response and what a student provides

## Errors, Slips, Misconceptions, and Lack of Understanding

- Underlying Causes of Error
  - *Slip*: a careless procedural mistake
  - *Misconception*: a persistent conceptual or procedural confusion
  - *Lack of understanding*: missing bit of conceptual or procedural knowledge without any persistent confusion



## Errors, Slips, Misconceptions, and Lack of Understanding

- *Any* attribution of underlying cause is an inference, a “formative hypothesis,” that can be tested through further assessment
  - Asking the student's explanation
  - Administering more tasks
  - Relating the error to other examples



## Principled Formative Assessment

- Our characterizations of students are inferences
- Inferences are uncertain
- We can tolerate more uncertainty when the consequences of error are low and decisions are reversible
- The more certain we are, the more effectively we can adjust instruction
- Uncertainty can be decreased through multiple sources, occasions, and contexts

## Outline

- Six issues
  - The definitional issue
  - The effectiveness issue
  - The domain issue
  - The measurement issue
  - The professional development issue
  - The system issue
- Summary



## The Professional Development Issue

- Effective formative assessment requires:
  - Pedagogical knowledge
  - Deep domain understanding
  - Measurement fundamentals
- A subset is unlikely to work!

## Developing Teachers' Formative Assessment Practice

- Can the components be effectively addressed semi-independently?
  - *KLT* focuses on the *pedagogical-knowledge* aspect of formative assessment
  - Formative-assessment pedagogical knowledge is connected to domain understanding through the TLCs
  - Measurement fundamentals presumably come from some other source



## The Professional Development Issue

- Time to learn to use or adapt purposefully constructed, *domain-based*, formative-assessment materials
  - Items
  - Integrated task sets
  - Projects
  - Diagnostic tests
  - Observational and interpretive guides

## Outline

- Six issues
  - The definitional issue
  - The effectiveness issue
  - The domain issue
  - The measurement issue
  - The professional development issue
  - **The system issue**
- Summary

## The System Issue

- Formative assessment exists within a larger educational context
- The components of that context must be:
  - Internally coherent
    - Formative and summative assessments are aligned with one another
  - Externally coherent
    - Formative and summative assessments are consistent with accepted theories of learning, as well as with socially valued learning outcomes

## A Common Reality

- For practical reasons, summative tests are relatively short and predominantly take the M-C format
- Those tests measure a subset of the intended curriculum
- Classroom instruction and formative assessment will be aligned to that subset
- The potential of formative assessment to effect deeper change will be significantly reduced

## The System Issue

- The effectiveness of formative assessment will be limited by the nature of the larger system
- We have to change the system

## Outline

- Six issues
  - The definitional issue
  - The effectiveness issue
  - The domain issue
  - The measurement issue
  - The professional development issue
  - The system issue
- **Summary**



## Summary

- The term, *formative assessment*, does not yet represent a well-defined set of artifacts or practices
  - A meaningful definition requires a conceptual framework, action theory, and concrete instantiation
- *KLT* has moved us toward such a definition

## Summary

- The practices associated with formative assessment can, *under the right conditions*, facilitate learning
- The benefits may vary widely from:
  - One implementation of formative assessment to the next
  - One subpopulation of students to the next

## Summary

- Commonly made quantitative claims for the efficacy of formative assessment are suspect
  - The effect-size claim of .4 - .7 SD is *not* meaningful, nor traceable to *any* inspectible empirical source
  - Other empirical sources are dated, unpublished, critically flawed, or show smaller effects than advocates cite
  - The validity argument, and backing to support it, are generally absent
- We need to be more responsible in our claims

## Summary

- Rooting formative assessment in pedagogical skills alone is insufficient
- Formative assessment should be conceptualized and instantiated within specific domains
  - *Foundational Approaches in Science Teaching* (Shavelson, 2008)
  - CBAL

## Summary

- Formative assessment is *assessment*
- If it's *assessment*, relevant measurement principles should figure centrally in the conceptualization and instantiation

## Summary

- Teachers need *substantial*:
  - Knowledge to implement formative assessment effectively in classrooms
  - Time and support to develop it
  - Materials that model the integration of pedagogical, domain, and measurement knowledge

## Summary

- Formative assessment exists in an educational context
- Ultimately, we have to rethink assessment as a coherent *system*

*“After five years of work, our euphoria devolved into a reality that formative assessment, like so many other education reforms, has a long way to go before it can be wielded masterfully by a majority of teachers to positive ends.”*

R. J. Shavelson,  
*Applied Measurement in Education*  
2008, p. 294





# Formative Assessment: A Critical Review

*Randy Bennett*

*ETS*

[\*rbennett@ets.org\*](mailto:rbennett@ets.org)

Staff seminar at RAMA, Tel Aviv, Israel, October 25, 2009

*Listening.  
Learning.  
Leading.*