Assessing Students: An Integrated, Probabilistic Approach

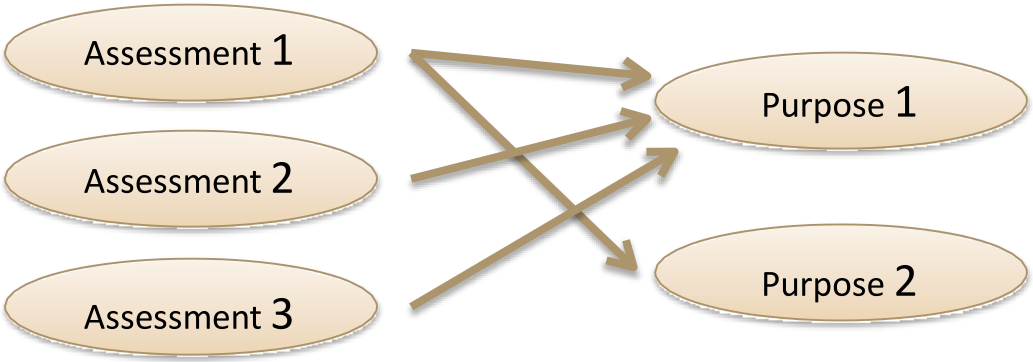
Huan Wang
PhD. Sr. Research Scientist, DRC
hwang@datarecognitioncorp.com

Motivation

If formative assessments are more native to instruction, more likely to adapt to student profiles & needs, gather richer performance data, provide more immediate and targeted feedback, and **occur much more frequently than summative assessments, which means more information**,

- ☒ why can't we utilize data from formative assessments in a principled way to inform accountability decisions typically only served by summative assessment?
- ☒ why can't we use data from formative and summative assessments together to provide evidence on student achievement, progress, etc.?

If single assessment cannot serve multiple purposes well, how about multiple assessments for multiple purposes?



Key IPA Components and Relationship to Previous Work **(new ideas/expansion in red)**

1. Definition of a common set of assessment objectives (e.g., standards).

2. Principled design and development of tasks and explicit linkage of tasks to targeted assessment objectives, e.g., based on Evidence-Centered Design *(ECD; Mislevy, Steinberg, & Almond, 2003; Mislevy, Almond, & Lucas, 2003)*.
   - See related discussion in *Mislevy & Durán, 2014* on
     - Learning progression
     - Identification of important contextual variables (e.g., student demographics, teacher effect, etc.)

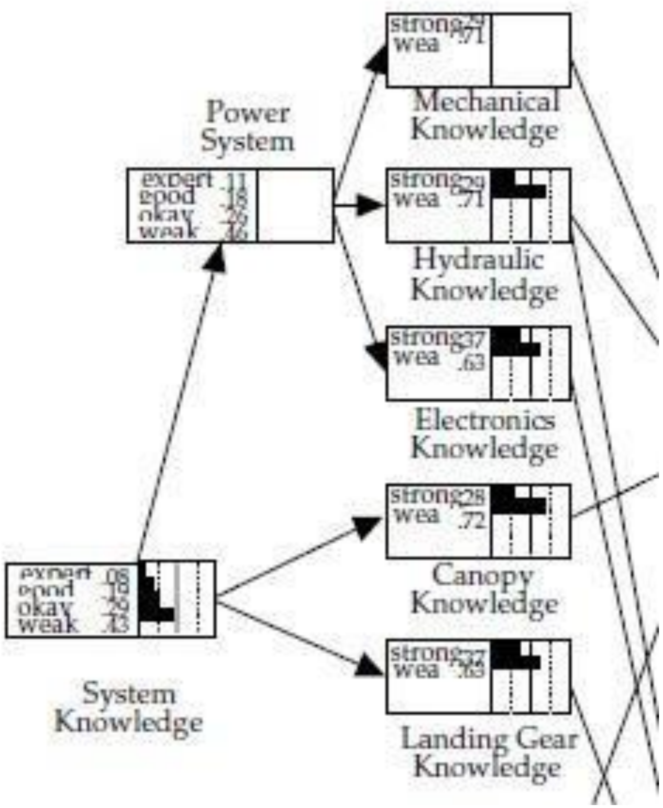
3. Collection of task and performance data (see single-assessment examples in Mislevy, 2008) from local assessments
   - Possibly automated data parsing and collection via web-based assessments to reduce teacher burden and to support data integrity

(to be cont'd)

Key IPA Components and Relationship to Previous Work (new ideas/expansion in red)

4. Ongoing updated estimation of student's probability of mastery on each assessment objective over time using data continuously accumulated from multiple assessment events.
   - Building and learning a probabilistic graphical model *(see Pearl, 1988; Korb & Nicholson, 2010; Koller & Friedman, 2009; better computation efficiency, better handling of complex variables and missing data)*
   - Utilization of PGM in game-based assessment
5. Auditing and calibrating with data from summative assessments.
   - Audits and other formal strategies for gauging local evaluation *(Linn, 1993; Mislevy, 2008; Resnick, 1997)*
6. Empirical estimation and monitoring of measurement error
7. Tracking and reporting the mastery status of each student *(see single-assessment examples in Mislevy, 2008)* along with structured compilation of supporting evidence.
8. Reporting and interpreting composite scores with evidence available.

Reporting example from Mislevy, 2008



Mapping to Peer Review Guidance

Content Standards

1. Definition of a common set of assessment objectives.

Statewide Assessment Systems; Technical Quality; Alignment; Inclusion

2. Principled design and development of tasks and explicit linkage of tasks to targeted assessment objectives.
3. Collection of task and performance data from local assessments.
4. Ongoing updated estimation of student's probability of mastery on each assessment objective over time using data continuously accumulated from multiple assessment events.
5. Auditing and calibrating with data from summative assessments.
6. Empirical estimation and monitoring of measurement error

Academic Achievement Standards; Assessment Reports

7. Tracking and reporting the mastery status of each student and structured compilation of supporting evidence.
8. Reporting and interpreting composite scores with evidence available.

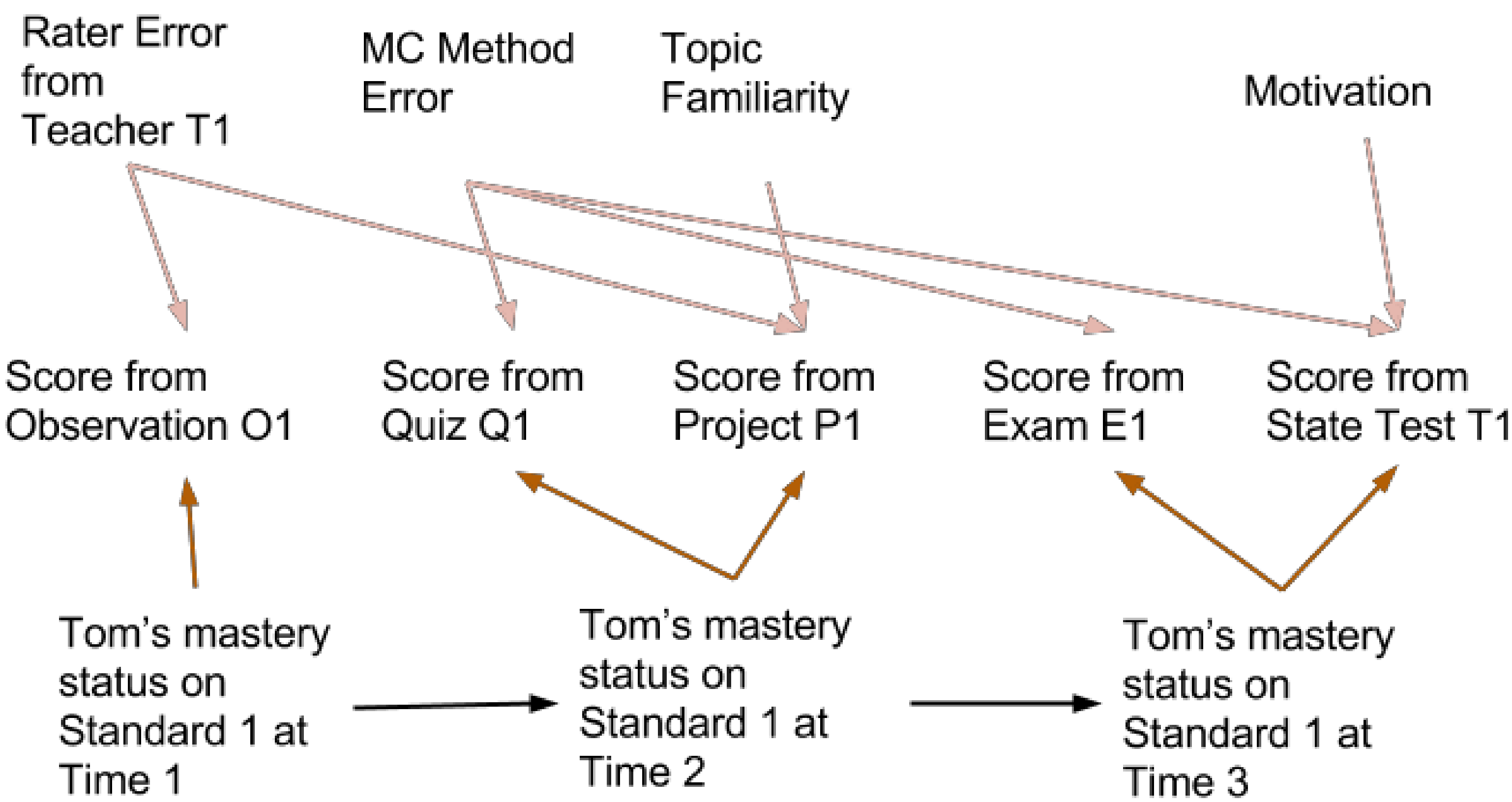
Benefits and Challenges

- ***Benefits***
  - More individualized & intuitive assessment experience for students
  - More instructionally meaningful reporting information for students with diverse profiles and students with missing evidence on assessment objectives.
  - Better tracking and monitoring of sources of measurement error
  - Better correspondence btw standards, instruction, assessment, and professional development (PD)
  - Principled, efficient use of assessment data for multiple purposes (e.g., instructional feedback, growth, mastery status, accountability)

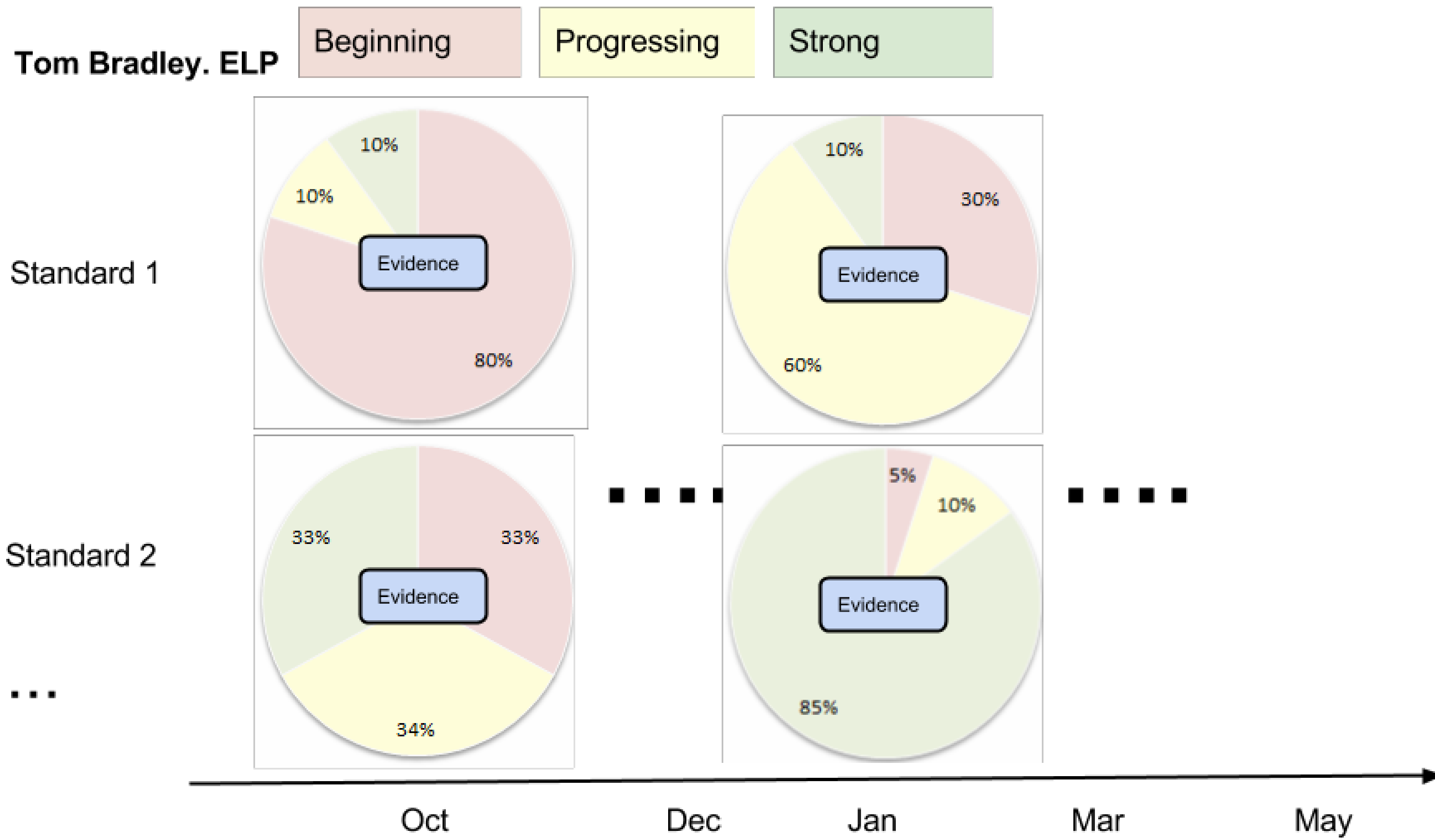
- ***Major Challenges***
  - PD (task design and development)
  - Technology infrastructure
  - Data sharing and security

An Example (Simplified)

Probabilistic Graphical Model (PGM)



Student Report Mock-Up



"the 3rd Generation of Assessment"

The third generation of assessments is one of reinvention occurring on multiple fronts simultaneously (Bennett, 1998, 2010b). It is in this third generation that what was, at first, an evolution driven primarily by technology becomes driven by substance. For one, these assessments serve both institutional and individual-learning purposes. Second, they are designed from cognitive principles and theory-based domain models. Third, the assessments use complex simulations and other interactive performance tasks that replicate important features of real environments, allow more natural interaction with computers, and assess new skills in more sophisticated ways. Finally, the assessments are more integrated with instruction, sampling performance repeatedly over time.

(Bennett, 2015, p. 372)

Reference

Bennett, R. E. (2015). Chapter 10: The Changing Nature of Educational Assessment. *Review of Research in Education, 39*, 370-407.

Korb, K. B., & Nicholson, A. E. (2010). Bayesian Artificial Intelligence (2nd ed.). London, UK: Chapman & Hall/CRC.

Koller, D. & Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

Mislevy, R. J. (2008). Issues of structure and issues of scale in assessment from a situative/sociocultural perspective. In P.A. Moss, D. Pullin, E. H. Haertel, J. P.Gee, & L. J. Young (Eds.), Assessment, equity, and opportunity to learn (pp. 259–294). New York, NY: Cambridge University Press.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (ETS Research Report RR-03-16). Retrieved from http://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Mislevey, R. J., & Duran, R. P. (2014). A sociocognitive perspective on assessing EL students in the age of Common Core and Next Generation Science Standards. *TESOL Quarterly*, *48* (3), 560-585.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems : networks of plausible inference.

Resnick, L. B. (1997). Student performance portfolios. In H. J. Walberg & G. D. Haertel (Eds.), *Psychology and educational practice* (pp. 158-175). Berkeley, CA: McCutchan.