Rethinking our Approach to Testing and Reporting: Special Student Groups and Beyond

Presenters:
Michael Middleton, WA OSPI
Becky Huang, University of Texas, San Antonio
Huan Wang, McGraw-Hill Education CTB

Discussant:
Richard Durán, UCSB

NCSA June 24, 2015
INTRODUCTION

Huan Wang
PhD. Sr. Research Scientist, CTB/McGraw-Hill
huan.wang@ctb.com
## Scoring Schema (assuming ten items per domain)

- **Table scoring**: Ray will receive the same Math score as Tom.
- **Pattern scoring**: Ray may have a higher/lower score than Tom.
- **Computer adaptive testing (starting from easy items)**: Ray may be assigned a lower score than Tom.
- **In all three cases**, Ray may have larger measurement error & person misfit issue than others.

## Conclusion

- Score values and measurement precision for students who deviate from the mainstream performance pattern may be impacted by such deviation.
**Low-Incidence Disability Students on ELP Assessment**

*Example: Grade 6 ELP Test*

<table>
<thead>
<tr>
<th>Scale Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tom (lost hearing and hasn’t learned ASL)</strong></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>500</td>
</tr>
<tr>
<td>480</td>
</tr>
<tr>
<td>?</td>
</tr>
</tbody>
</table>

1. Speaking  
2. Listening  
3. Reading  
4. Writing

Evidence missing on oral communication skills  
Evidence missing on written communication skills

Copyright © 2014 CTB/McGraw-Hill LLC
In this Symposium

Presentations
• WA state alternate assessment: challenges and approach  
  *(Michael Middleton)*
• Assessing young ELL learners, paper and online  
  *(Becky Huang)*
• An alternative assessment approach  
  *(Huan Wang)*

Discussion
• Implications for formative and summative assessments  
  *(Richard Durán)*

Moderator
*Joanna Tomkowitz*
Testing Students on Alternate Standards:
Looking at Achievement Through a Different Lens

Michael Middleton, Director – Select Assessments
National Conference on Student Assessment
San Diego, CA – June 21-24, 2015
Topics

• What is an “Alternate Assessment”?
  • How Washington addressed elements for assessing
• What is the design behind WA-AIM?
• How does the approach to achievement look different?
• When will WA know about meeting Goals?
Alternate Assessments

• Assessing academic proficiency of students with significant cognitive disabilities based on alternate standards.
  • Washington’s post-secondary (aka, college-&-career) readiness standards correspond with Common Core State Standards (CCSS) then apply linkages through the Essential Elements\(^1\) in mathematics and English language arts
  • Washington’s state standards in science are basis for deriving alternate linkages

• Alternate standards provide students with access to the general learning standards through alteration of the depth, breadth, and complexity
  • Washington’s framework establishes varying levels of content complexity to support students with diverse levels of cognitive development who participate in the alternate assessment

• Common assessment instrument may be collections of student work (i.e., portfolios) or performance-based tasking
  • Washington has moved to a performance based model

\(^1\) Dynamic Learning Maps Consortium (2013). Dynamic Learning Maps Essential Elements for English Language Arts/Literacy & Mathematics. Lawrence, KS: University of Kansas.
## Access to Instruction & Measurement
### Alternate Assessment Test Blueprint for Mathematics

<table>
<thead>
<tr>
<th>Domain</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
<th>Grade 6</th>
<th>Grade 7</th>
<th>Grade 8</th>
<th>High School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operations and Algebraic Thinking (OA)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number and Operations in Base Ten (NBT)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number and Operations – Fractions (NF)</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measurement and Data (MD)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geometry (G)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Ratios and Proportional Relationships (RP)</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Number System (NS)</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Expressions and Equations (EE)</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functions (F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Statistics and Probability (SP)</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>The Real Number System (N-RN)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Creating Equations (A-CED)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Reasoning with Equations and Inequalities (A-REI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Interpreting Functions (F-IF)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpreting Categorical and Quantitative Data (S-ID)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Access Point Frameworks:

• Access Points (AP) define the knowledge and skills the WA-AIM is designed to measure.

• For each standard a continuum of three Access Points was developed, representing three levels of content complexity:
  • (M) more complex, (I) intermediate complexity, (L) less complex.
Every standard has an AP framework with three levels of content complexity aligned to the Common Core Essential Element & CCSS.
Performance Task Requirements

**Mathematics**

**Domain: Measurement and Data**

**Cluster: Solve problems involving measurement and estimation of intervals of time, liquid volumes, and masses of objects.**

### Access Points Built on Three Levels of Complexity

<table>
<thead>
<tr>
<th>More Complex</th>
<th>Less Complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student will measure the length of an object to the nearest whole unit.</td>
<td>Student will identify tools that can be used to measure length.</td>
</tr>
</tbody>
</table>

---

**Performance Task Example 1:**

**Performance Task:**

- **Student:** Measure the length of the classroom.

**Directions:**

- Use a ruler or a tape measure to measure the length of the classroom.

**Examples:**

- Classroom is 40 feet long.

**Performance Task Example 2:**

**Performance Task:**

- **Student:** Measure the length of the table.

**Directions:**

- Use a ruler or a tape measure to measure the length of the table.

**Examples:**

- Table is 60 inches long.

---

**Performance Task Example 3:**

**Performance Task:**

- **Student:** Measure the length of the notebook.

**Directions:**

- Use a ruler or a tape measure to measure the length of the notebook.

**Examples:**

- Notebook is 12 inches long.
The Access Point level for assessment is determined on a standard-by-standard basis. Therefore a student's complete math assessment might have standards from varying levels of complexity.

<table>
<thead>
<tr>
<th>Standards</th>
<th>AP Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometry</td>
<td>I</td>
</tr>
<tr>
<td>Measurement &amp; Data</td>
<td>M</td>
</tr>
<tr>
<td>Numbers Operations Fractions</td>
<td>I</td>
</tr>
<tr>
<td>Operations and Algebraic Thinking</td>
<td>L</td>
</tr>
<tr>
<td>Numbers and Operations in Base Ten</td>
<td>M</td>
</tr>
</tbody>
</table>

L = lowest complexity level  
I = intermediate complexity level  
M = maximum complexity level
Teacher Achievement Level Study

- The Teacher Achievement Level Study is a modification of Contrasting Groups\(^1\).
  - WA special education practitioners studied AALDs for each achievement level
  - Open to all educators with students participating in WA-AIM
  - Considered knowledge, skills, and abilities of classroom students
  - Determine achievement level best describing students within a content area.

- Teacher classifications not part of assigned student scores.
  - Classifications part of standard setting process (*impact data*)
  - Part of a synthesis discussion procedure (similar to *articulation*)

\(^1\) Livingston & Zieky, 1982
Teacher Weighting Study

• Subset of educators from the achievement level study.
  • Submitted actual student performance data from WA-AIM submission (scores and access points)
  • Estimated performance on non-assessed access points for same students

• Goal – generate a common scale of performance for students using WA-AIM.
  • Unlike the general assessment students are not using the same items

• Process – apply numeric multipliers (weights) on scores
  • Statistical analyses conducted on survey data to generate weight
Profile Sorting

• In-person committee of WA educators considering types of performances aligned with each alternate achievement level.
  • The process designed to make determinations not on students’ disabilities, rather students’ demonstrated knowledge, skills, and abilities.

• Judgement-based process relying on expertise of state educators
  • Similar to other standard-setting designs
  • Process of impression/idea exchanges between practitioners

• Student profiles reviewed to determine knowledge, skills, and abilities.
  • Reviews scores associated by domain, strand, or PE (content-specific).
  • Profiles (also referred to as response vectors) summarize performance of a single student.

• Sorting occurs without attention to student's disabilities
  • Profiles display what the students can do, as evidenced through combination of scores and Access Points.
Contact Information

Michael Middleton
Director – Select Assessments
Office of Superintendent of Public Instruction
michael.middleton@k12.wa.us
360.725.6434

Connect with us:
www.facebook.com/waOSPI
www.twitter.com/waOSPI
www.youtube.com/waOSPI
www.flickr.com/waOSPI
- Approximately one out of five children in the U.S. speaks a different or additional home language other than English (language minority students) (United States Census Bureau, 2013)

- The number of English language learner students in Preschool-12 grades has grown exponentially in the past few decades (National Center for Education Statistics, 2014)
Considerations for Assessing Young ELLs

1) Learner-level characteristics

2) Construct: Distinction between social and academic language

3) Use of Technology
1. Learner-Level Characteristics

**Adults**  
*(Bialystok, 2001)*
- Relatively stable state

**Young ELLs**  
*(McKay, 2006)*
- State of constant cognitive, social, emotional, physical growth.
- Heterogeneity in the rate of development
Cognitive Growth

• School-age learners are developing abilities to think in new ways.

• They have shorter attention spans
Taking into Account Young ELL Students’ Cognitive Growth

• Cognitive demands of the assessment should be commensurate with their age-related abilities.
  - May not want to ask young learners (age 5-7) without meta-language to analyze or describe a language rule, such as the third person singular rule in English

• Assessment tasks should not extend beyond their experiences of the world.
  - e.g., Ask school-age learners living in San Diego, CA to describe what they generally do on a snow day or to distinguish between different types of snowfalls
Emotional/Social Growth

1) Start developing from a main interest in self towards greater social awareness

2) Need for love, recognition and belonging shifts from dependence on adults to peers

3) Heightened sensitivity to criticism and feelings of success and failure are dependent on the responses of adults and peers

Retrieved from: www.parenting.com
Taking into Account Young ELL Students’ Emotional/Social Growth

- Assessment tasks should be familiar and the environment should be “psychologically safe”, especially for younger learners.
- Take into account students’ likes and interests.
- Assessment should give students a sense of overall success and/or progression.
- Positive experiences and friendly feedback to help maintain their enthusiasm and motivation.
Physical Growth

Development of **fine-motor skills**: write, cut, draw, and hand-eye coordination.

Development of **gross-motor skills**: run, climb, balance.
Taking into Account Young ELL Students’ Physical Growth

• Assessment tasks need to consider learners’ ability to sit still, and hand-eye coordination.

• For younger learners (pre-k through early elementary grades):
  - Assessment tasks that involve physical activity to accompany language-related response may be more effective
Finally….Variable Rate of Development

• Rate of development may vary among Young ELL learners
• Each individual may also vary in their own development across dimensions
Finally….Variable Rate of Development

• Allow some flexibility in assessment (e.g., tasks catering for all levels)

• Interpret the assessment outcomes with caution

• Use multiple measures; include both summative and formative assessments (Brookhart, 2009; National Council on Measurement in Education, 1995)
2. Construct for Language and Content Area Assessments

What do we want to know about our ELL students’ language proficiency?

How can we tease apart ELL students’ language proficiency and content area knowledge?
Social vs. Academic Language

BICS and CALP

**BICS** (Basic Interpersonal Communication Skills)

“*conversational proficiency*” is the ability to use language successfully in everyday social situations.

**CALP** (Cognitive Academic Language Proficiency)

“*academic proficiency*” is the ability to use complex language for school-related tasks.

(Cummins, 1999)
LESS COGNITIVELY DEMANDING

SOCIAL LANGUAGE

CONCEPT-EMBEDDED

BICS

MORE COGNITIVELY DEMANDING

ACADEMIC LANGUAGE

CALP

CONTEXT-REDUCED
3. Technology

The use of technology in instruction can be beneficial to young children’s cognitive and social learning (Clements 1994; Haugland 1992).

Relatively little is known about how the use impacts students and the pros/cons of online assessment (Liang & Creasy, 2004), especially with young ELL students.
3. Considerations/Challenges for Online Assessment

**Pros:**
- Student-centered
- Faster access to results (Gaytan & McEwen, 2007)

**Cons:**
- Validity
- Developmental appropriateness
- Fairness and ethical issues
Contact Info:

Becky Huang
becky.huang@utsa.edu
An Integrated, Probabilistic Approach

Huan Wang
PhD. Sr. Research Scientist, CTB/McGraw-Hill
huan.wang@ctb.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?
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
     o Learning progression
     o 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)
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

Tom Bradley. ELP

Beginning

Progressing

Strong

Standard 1

Evidence

80%

10%

10%

Standard 2

Evidence

33%

33%

34%

Oct

Dec

Jan

Mar

May

Copyright © 2014 CTB/McGraw-Hill LLC
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)


THANK YOU!