Evaluating Alignment of Arizona ELP Standards to State Achievement Assessment

Kate Wright (Arizona Department of Education)
Hiro Fukuhara (Pearson)

Presentation at National Conference of Student Assessment in Austin, TX on June 30, 2017

Background

Arizona English Language Learner Assessment (AZELLA)

New Placement and Reassessment since Spring 2013

Proficiency Levels

- Four Proficiency Levels on Total Combined
- Three Proficiency Levels on Other Domains/Subdomains
- Overall Proficient if and only if Proficient on Total Combined, Reading and Writing (Three criteria)

ELL students who earn Overall Proficient on Reassessment will be reclassified as Reclassified Fluent English Proficient (RFEP) and go to mainstream classrooms in the next school year.
Aligning EL Proficiency with Academic Proficiency

How can we determine if the cuts on an ELL assessment are set at a right place? (Expectation: A proficient student no longer has a language barrier to follow regular classrooms.)

- During a standard setting meeting
  - Rounds of review
  - Impact data

- After the standard setting meeting
  - Comparison of performance between native English speakers and RFEP students on a state achievement assessment
  - Interview teachers how RFEP students do in a classroom

AZELLA and AIMS
Setting Standards on the AZELLA

1st meeting in May 2013

- Bookmark
  - Proficiency level cuts set on Total Combined
  - Adjustment on Reading and Writing Proficient cut after the cuts on Total Combined were finalized

- Impact data on AZELLA only

Setting Standards on the AZELLA

2nd meeting in June 2013

- Standard validation meeting

- Impact data on Arizona’s Instrument for Measuring Standards (AIMS) Reading, Mathematics, and Writing for Stages III-V.
  - Analyses analogous to Cook, Linquanti, Chinen, Jung (2012).

- Panelists had an opportunity to adjust the Proficiency level cuts from May 2013
  - They decided to keep the original cuts.
Validation of AZ ELP Standards

Research Question: How do former ELLs perform, compared to native English speaking peers?

Comparison of % Pass on AIMS Reading in 2014 between Non-PHLOTE and RFEP/FEP students

<table>
<thead>
<tr>
<th>Test Level</th>
<th>Non-PHLOTE</th>
<th>RFEP</th>
<th>% Difference</th>
<th>Non-PHLOTE vs RFEP</th>
<th>FEP 1</th>
<th>% Difference</th>
<th>Non-PHLOTE vs FEP 1</th>
<th>FEP 2</th>
<th>% Difference</th>
<th>Non-PHLOTE vs FEP 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>82.0%</td>
<td>85.1%</td>
<td>-3.1%</td>
<td>82.4%</td>
<td>-0.5%</td>
<td>91.7%</td>
<td>-9.7%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>80.7%</td>
<td>72.6%</td>
<td>8.1%</td>
<td>77.8%</td>
<td>3.0%</td>
<td>74.8%</td>
<td>6.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>84.3%</td>
<td>60.6%</td>
<td>23.7%</td>
<td>74.7%</td>
<td>9.6%</td>
<td>83.1%</td>
<td>1.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>84.3%</td>
<td>60.7%</td>
<td>23.6%</td>
<td>57.8%</td>
<td>26.5%</td>
<td>71.0%</td>
<td>13.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>89.4%</td>
<td>66.7%</td>
<td>22.7%</td>
<td>75.0%</td>
<td>14.4%</td>
<td>73.2%</td>
<td>16.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>73.2%</td>
<td>14.7%</td>
<td>58.4%</td>
<td>30.7%</td>
<td>42.5%</td>
<td>25.0%</td>
<td>48.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H5</td>
<td>81.8%</td>
<td>54.4%</td>
<td>27.4%</td>
<td>65.4%</td>
<td>16.3%</td>
<td>50.3%</td>
<td>31.6%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It appears that RFEP students performed much worse than native English speaker peers for the most of grades.

Is it a fair comparison to make?

• A group of ELL students tend to come from a lower SES family.
Regression Discontinuity Design

Regression Discontinuity Design (RDD; Jacob & Zhu, 2012; Lee & Munk, 2008; Robinson, 2011; Smith, 2014)
Quasi-experimental design to evaluate program effects (e.g., ELL program, regular classrooms)
  • Research question: Are former ELL students benefiting to be in a regular classroom as opposed to in an ELL program or vice versa? If so, the Proficient cut was set right (External validity evidence).

Regression Discontinuity Design

Powerful tool when investigation of such effects is not feasible under a randomized experiment design
  • No special data collection is needed

Can control for background variables (e.g., demographics)

Needs a threshold (e.g., AZELLA Overall Proficient cut) for a group assignment (e.g., ELL reclassification)
Model

Parametric RDD Approach
- Regression analysis with centering technique
  \[ Y_j = \alpha + f(M_j) + \delta R_j + \epsilon_j \]
  - \(Y_j\): Outcome of interest after a group assignment
  - \(\alpha\): Intercept
  - \(f(M_j)\): Regression function of a measure, \(M_j\), centered at a threshold for the group assignment
  - \(R_j\): Group indicator
  - \(\delta\): Effect of group assignment on the outcome of interest at the threshold
  - \(\epsilon_j\): Residual

- \(\delta\) is the parameter of interest in RDD (i.e., performance of students just below or above the threshold)

- A key is to find an appropriate \(f(M_j)\) to minimize a bias of estimating \(\delta\)

Possible Outcomes

Possible Outcomes from the Parametric RDD Study (e.g., Smith 2014)
1. ‘No effect’ (\(\delta=0\))
   \[ Y_j = \alpha + f(M_j) + \delta R_j + \epsilon_j \]
   - No disconnect between regression lines for RFEP and ELL at AZELLA cut score
   - Reclassification or retention of ELL students doesn’t impact either positively or negatively
   - AZELLA cut score is appropriate (desirable outcome)
Possible Outcomes

Possible Outcomes from the Parametric RDD Study (e.g., Smith 2014)

2. ‘Positive Effect’ ($\delta > 0$, RFEP = 1; ELL = 0)

- ELL students could get more benefits if they were in regular classrooms
- AZELLA cut score is higher than it should

3. ‘Negative Effect’ ($\delta < 0$, RFEP = 1; ELL = 0)

- RFEP students could get more benefits if they were retained in an ELL program
- AZELLA cut score is lower than it should
Data

AZELLA Reassessment from Spring 2013
- Students at Overall Proficient as RFEP
- Rest of students as ELL

AIMS from Spring 2014
- Reading: Grades 3 – 8, HS
- Mathematics: Grades 3 – 8, HS
- Writing: Grades 5 – 7, HS

<table>
<thead>
<tr>
<th>AIMS</th>
<th>Grade</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>High School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td></td>
<td>11551</td>
<td>5939</td>
<td>5757</td>
<td>4486</td>
<td>2874</td>
<td>1427</td>
<td>1139</td>
</tr>
<tr>
<td>Mathematics</td>
<td></td>
<td>11553</td>
<td>5937</td>
<td>5756</td>
<td>4486</td>
<td>2873</td>
<td>1426</td>
<td>1139</td>
</tr>
<tr>
<td>Writing</td>
<td></td>
<td>-</td>
<td>-</td>
<td>5752</td>
<td>4468</td>
<td>2863</td>
<td>-</td>
<td>1125</td>
</tr>
</tbody>
</table>

Method

Model Selection

Binding Score
- RDD involves with a single cut.
- AZELLA has three cuts to decide Overall Proficient.
- Robinson (2011) suggested constructing a ‘binding score’, \( M_j = \min(\text{AZELLA Total Combined scale score, AZELLA Reading scale score, AZELLA Writing scale score}) \), in which each of the scale scores is standardized and recentered around its respective Proficient cut.
- RFEP students have positive (or 0) binding scores
- ELL students have negative binding scores

Analytic Sample
- Students just below or above the cut
Results

Scatter Plot of AIMS Score against AZELLA Binding Score

\[ \text{Cor.} = 0.47 \]

Scatter Plot of AIMS Score against AZELLA Binding Score

\[ \text{Cor.} = 0.31 \]
Results

AZELLA Overall Proficient cut to reclassify ELL students as RFEP or retain them as ELL was appropriate for most of the grade and subject combinations.

There are only 2 out of 18 samples that drew different conclusions
- Grade 8 Reading result showed that the AZELLA cut score was statistically too high
- Grade 6 Mathematics results showed that the cut was statistically too low
- The results were not inconsistent for other subjects for those grades
- The magnitude of effect on ELL reclassification for both grades was trivial from the measurement perspective

Discussion

Two approaches for validity evidence of ELP standards drew different conclusions.

Arizona Department of Education keeps monitoring the alignment of ELP standards to the state achievement standards as their achievement assessment is aligned to Arizona College and Career Ready Standards (Arizona’s Measurement of Educational Readiness to Inform Teaching, AzMERIT).
- ‘Once performance standards have been set for a test it may be desirable or necessary to review or revalidate those standards for a variety of reasons’ (Mattar, Hambleton, Copella & Finger, 2012).
References


History

The Arizona Department of Education has been in agreements with the Office of Civil Rights regarding the AZELLA test since 2012.

Most recently, it was determined that the existing Proficiency level cut scores for some AZELLA tests might not be set high enough to ensure that EL students have sufficient English skills to meaningfully participate in mainstream classrooms where instruction is aligned to more rigorous college and career ready standards and assessments.

Increasing AZELLA Cut Scores

ADE has raised the Proficient cut scores for the AZELLA Stages III, IV, and V tests to a level that makes it more likely Fluent English Proficient – Year 2 (FEP2) students will perform like their English-only peers on AzMERIT ELA.

The old and new Total Combined Proficient cut scores for Stages III-V are shown below.

<table>
<thead>
<tr>
<th>Total Combined Proficient Cut Score</th>
<th>Stage III</th>
<th>Stage IV</th>
<th>Stage V</th>
</tr>
</thead>
</table>
Meeting Higher Standards

The change in the Proficient cut scores was effective retroactively to school year 2012-13 and applied to both Placement and Reassessment tests, so the English proficiency of some students was reevaluated compared to the performance of English-only students.

Students who were performing at or above the average AzMerit ELA score for English-only students were considered to be performing like their English-only peers. These students maintained their FEP status.

Students performing below their English-only peers were eligible for additional services depending on school data and the input of parents or guardians.

AZEELLA Technological Innovations

1. Online Testing
   – AZELLA is moving to an online format for Stages III-IV in 2017-2018.
   – This aligns with AzMerit tests in grades 3-12 which are also primarily online.

2. Multi-stage Adaptive Testing
   – AZELLA is planning to field-test multi-stage adaptive testing in 2018-2019
   – 1-2 design (more difficulty or easier module in stage 2)

3. Automated Scoring/ Open-ended Items
   – All Oral Reading and Speaking items currently scored by an automated scoring engine with 10% human backread for reliability check.
   – Writing prompt will be scored by the engine starting in Spring 2018
Realignment of ELP Standards

New Arizona College and Career Ready ELA and Mathematics Standards were adopted by the State Board of Education in December, 2016.

An ELP Standards alignment and revision committee has been created with the task of aligning the ELP Standards to the College and Career Ready Standards, with adoption in 2018.

Assessment Alignment

AZMerit test items will be aligned to the newly adopted Arizona College and Career Ready Standards

AZELLA test items will be aligned to the revised ELP Standards (2019-2020)
Accountability

K - 8

- Proficiency: 30%
- Growth: 50%
- Proficiency & Growth of English Learners (5% each)
- Acceleration Measures (i.e., inclusion of students with disabilities in general education, growth of subgroups, decreases in chronic absenteeism)