Improving Assessment and Instruction with Process Data and Learning analytics

Mingyu Feng
Senior Research Associate, WestEd

Nov 7, 2019    2019 MARC Conference
Overview

Two strands of work we have done using learning analytics based on continuous process data from technology-based learning systems

1. Assess Student Learning Outcomes
2. Support Teacher Practices and Improve Student Learning
SECTION 1

Using Process Data to Assess Learning Outcome

In ASSISTments
Motivation

Concerns about poor student performance on new state tests

• High-stakes standards-based tests are required every state
• Student performance are not satisfactory
  ➤ National report card 2017, only 34% percent of 8th students performed at or above the Proficient level on the mathematics assessment

Secondary teachers are asked to be data-driven

• Assessment takes valuable classroom time away
  ➤ Can we assess student performance without sufficing learning time?
• Not all kinds of data are equally accessible and informative
  ➤ State test reports, even interim assessment reports are not readily available soon enough to guide instruction
  ➤ The integration of DDI takes great effort of teachers
ASSISTments: Blending Assisting and Assessment

When solving problems in ASSISTments,

• Students get immediate feedback, hints, guided practice / scaffolding questions

ASSiSTments
A Free Public Service of Worcester Polytechnic Institute
A Grade Book Report

Where does this score come from?

<table>
<thead>
<tr>
<th>Student Name</th>
<th>Elapsed time (hh:mm)</th>
<th>Original Items</th>
<th>Scaffolding + Original Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Done</td>
<td>% Correct</td>
<td>Est. MCAS Scaled Score*</td>
</tr>
<tr>
<td>Tom*</td>
<td>4:12</td>
<td>90</td>
<td>39%</td>
</tr>
<tr>
<td>Dick*</td>
<td>4:01</td>
<td>98</td>
<td>66%</td>
</tr>
<tr>
<td>Harry*</td>
<td>4:07</td>
<td>78</td>
<td>40%</td>
</tr>
<tr>
<td>Mary*</td>
<td>4:17</td>
<td>114</td>
<td>20%</td>
</tr>
<tr>
<td>Jack*</td>
<td>3:53</td>
<td>104</td>
<td>39%</td>
</tr>
</tbody>
</table>

*Note: Scaled scores are estimates based on performance.

WestEd.org
Why should we be more complicated?

- **Static** – does not distinguish “Tom” and “Jack”
- **Average** – ignores development over time
- **Uninformative** – not informative for classroom instruction

<table>
<thead>
<tr>
<th>Student Name</th>
<th>Elapsed time (hh:mm)</th>
<th># Done</th>
<th>% Correct</th>
<th>Est. MCAS Scaled Score</th>
<th>Est. MCAS Performance Level</th>
<th># Done</th>
<th>% Correct</th>
<th># Hint Req.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom*</td>
<td>4:12</td>
<td>90</td>
<td>39%</td>
<td>214</td>
<td>Warning/Failing</td>
<td>228</td>
<td>44%</td>
<td>353</td>
</tr>
<tr>
<td>Dick*</td>
<td>4:01</td>
<td>98</td>
<td>66%</td>
<td>244</td>
<td>Proficient</td>
<td>158</td>
<td>59%</td>
<td>58</td>
</tr>
<tr>
<td>Harry*</td>
<td>4:07</td>
<td>78</td>
<td>40%</td>
<td>224</td>
<td>Needs improvement</td>
<td>154</td>
<td>38%</td>
<td>77</td>
</tr>
<tr>
<td>Mary*</td>
<td>4:17</td>
<td>114</td>
<td>20%</td>
<td>200</td>
<td>Warning/Failing</td>
<td>356</td>
<td>20%</td>
<td>705</td>
</tr>
<tr>
<td>Jack*</td>
<td>3:53</td>
<td>104</td>
<td>39%</td>
<td>214</td>
<td>Warning/Failing</td>
<td>267</td>
<td>43%</td>
<td>187</td>
</tr>
<tr>
<td>John*</td>
<td>4:24</td>
<td>92</td>
<td>55%</td>
<td>244</td>
<td>Proficient</td>
<td>40</td>
<td>52%</td>
<td>55</td>
</tr>
</tbody>
</table>

**Dynamic assessment**

**Longitudinal modeling**

**Cognitive diagnostic assessment**
Dynamic Assessment – The idea

- Dynamic testing began before computerized testing (Brown, Bryant, & Campione, 1983).

<table>
<thead>
<tr>
<th>Student Name</th>
<th>Elapsed time (hh:mm)</th>
<th># Done</th>
<th>% Correct</th>
<th>Est. MCAS Scaled Score*</th>
<th>Est.MCAS Performance Level</th>
<th># Done</th>
<th>% Correct</th>
<th># Hint Req.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom*</td>
<td>4:12</td>
<td>90</td>
<td>39%</td>
<td>214</td>
<td>Warning/Failing</td>
<td>228</td>
<td>44%</td>
<td>353</td>
</tr>
<tr>
<td>Dick*</td>
<td>4:01</td>
<td>98</td>
<td>66%</td>
<td>244</td>
<td>Proficient</td>
<td>158</td>
<td>59%</td>
<td>58</td>
</tr>
<tr>
<td>Harry*</td>
<td>4:07</td>
<td>78</td>
<td>40%</td>
<td>224</td>
<td>Needs improvement</td>
<td>154</td>
<td>38%</td>
<td>77</td>
</tr>
<tr>
<td>Mary*</td>
<td>4:17</td>
<td>114</td>
<td>20%</td>
<td>200</td>
<td>Warning/Failing</td>
<td>356</td>
<td>20%</td>
<td>705</td>
</tr>
<tr>
<td>Jack*</td>
<td>3:53</td>
<td>104</td>
<td>39%</td>
<td>214</td>
<td>Warning/Failing</td>
<td>267</td>
<td>43%</td>
<td>187</td>
</tr>
<tr>
<td>John*</td>
<td>4:24</td>
<td>92</td>
<td>55%</td>
<td>244</td>
<td>Proficient</td>
<td>40</td>
<td>52%</td>
<td>55</td>
</tr>
</tbody>
</table>
Dynamic vs. Static Assessment

Dynamic testing metrics

• # attempts
• # minutes to come up with an answer; # minutes to complete an ASSISTments
• # hint requests; # hint-before-attempt requests; #bottom-out hints
• % correct on scaffolds
• # problems solved

“Static” measure

• correct/wrong on original questions
Dynamic Assessment - Modeling

Three linear stepwise regression models

1-parameter IRT proficiency estimate
The standard test model

1-parameter IRT proficiency estimate + all online metrics
The mixed model

The assistance model

All online metrics

MCAS Score
Dynamic Assessment - Evaluation

**Bayesian Information Criterion (BIC)**
- Widely used model selection criterion
- Resolves overfitting problem by introducing a penalty term for the number of parameters
- Formula
- Prefer model with lower BIC

**Mean Absolute Deviation (MAD)**
- Cross-validated prediction error
- Prefer model with lower MAD

\[ BIC = -2 \log(L) + p \log(n) \]

\[ MAD = \frac{1}{n} \sum_{i=1}^{n} |MCAS_i - pred_i| \]

\[ Pct_{Err} = \frac{MAD}{(Max \ Raw \ Score)} \]

Dynamic Assessment - Results

1-parameter IRT proficiency estimate
1-parameter IRT proficiency estimate + all online metrics

All online metrics

The standard test model
The assistance model
The mixed model

<table>
<thead>
<tr>
<th>Model</th>
<th>MAD</th>
<th>BIC</th>
<th>Correlation with 2005 8th grade MCAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>The standard test model</td>
<td>6.40</td>
<td>-295</td>
<td>0.733</td>
</tr>
<tr>
<td>The assistance model</td>
<td>5.46</td>
<td>-402</td>
<td>0.821 <strong>p=0.001</strong></td>
</tr>
<tr>
<td>The mixed model</td>
<td>5.04</td>
<td>-450</td>
<td>0.841 <strong>p=0.001</strong></td>
</tr>
</tbody>
</table>
Dynamic Assessment - Robustness

- See if model can generalize
  - Train/test model on split of same year’s data
  - Test model on other year’s data

### Results of testing the mixed models on a different year’s data

<table>
<thead>
<tr>
<th>Year of training data</th>
<th>Year of Testing data</th>
<th>MAD on Testing data</th>
<th>Correlation with MCAS scores on testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004-2005</td>
<td>2005-2006</td>
<td>5.16</td>
<td>0.827</td>
</tr>
<tr>
<td>2005-2006</td>
<td>2004-2005</td>
<td>5.80</td>
<td>0.824</td>
</tr>
</tbody>
</table>
Section 1: Conclusion

ASSISTments process data enables us to assess more accurately.

The relative success of the assistance model over the standard test model highlights the power of the dynamic measures.
SECTION 2

Leveraging Learning Analytics to Support Teacher Practices and Improve Student Learning

In ASSISTments
ASSISTments for homework support

- Students get immediate feedback, hints, guided practice while doing homework
- Teachers can see how students are doing on homework
- Teachers can change homework reviews to adapt to students

ASSISTments™
A Free Public Service of Worcester Polytechnic Institute
ASSISTments Item Report

<table>
<thead>
<tr>
<th>Student/Problem --- [Unanonymize]</th>
<th>Average</th>
<th>PRAHESY</th>
<th>PRAHESW</th>
<th>PRAHESX</th>
<th>PRAHE</th>
<th>PRAHE2</th>
<th>PRAHE52</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem Average Graph</td>
<td>60%</td>
<td>90%</td>
<td>100%</td>
<td>100%</td>
<td>27%</td>
<td>61%</td>
<td>84%</td>
</tr>
<tr>
<td>Common Wrong Answers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct Answer(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1/5*3</td>
<td>1/7*8</td>
<td>1/6*2</td>
<td>1/3*10</td>
<td>1/5*3</td>
<td>1/5*3</td>
<td>1/16*2</td>
</tr>
<tr>
<td>Common Wrong Answers will present clues as to why students answer incorrectly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Percent correct per problem. This data identifies class weaknesses helping you drive your instruction.

This means 56% of the incorrect answers were 1/9*10. The percent will be in red if it is over 50%.
Research Overview
A Goal 3 Efficacy Trial

Sample
- Maine gives all students a laptop to take home, hence hardware is available
- Focused on 7th grade math students
- Recruited two cohorts:
  - Summer 2012 - June 2014
  - Summer 2013 - June 2015
- 46 schools recruited
- 43 schools in final sample
  - 87 teachers
  - 2800+ students
Research Overview
A Goal 3 Efficacy Trial

Random Assignment
- Schools paired by similar size and prior math scores
- “Coin toss” for each pair

Schools stay in condition for two years
- Year 1: Teacher preparation, practice and coaching
- Year 2: Teachers use with new cohort of students

Teachers expected to follow school homework policies
Content comes from their textbooks, plus pre-built content in ASSISTments
Findings

1. Impact on student learning
2. Benefit to students with low prior math achievement
3. Change in teacher practices
Finding 1: Impacts on Student Achievement

**effect size = 0.18; Unadjusted TerraNova Score 11 points higher**

![TerraNova National Norms vs. Measured Difference](chart.png)
Finding 2: Greater Effect for Low Prior Math

Via Median Split Analysis

Makes sense: higher achievers need less homework help
Finding 3: Change in Teacher Practice

- Shift from reviewing “all problems” to targeted problems

“Now it’s easier for me to go through and pick out the specific ones that I want them to do... I can use the reports to target those questions, those areas that it’s obvious that the entire class or most of the class is struggling with. So I can go back and spend a little more time with that.”

### How Did You Determine Which Problems to Review? (Survey)

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Control</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All the problems</td>
<td>24%</td>
<td>43%</td>
<td>0.000</td>
</tr>
<tr>
<td>Targeted problems</td>
<td>75%</td>
<td>36%</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### How Did You Determine Which Problems to Review? (Log)

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Control</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All the problems</td>
<td>21%</td>
<td>21%</td>
<td>0.968</td>
</tr>
<tr>
<td>Targeted Problems</td>
<td>93%</td>
<td>68%</td>
<td>0.003</td>
</tr>
<tr>
<td>Student Driven</td>
<td>57%</td>
<td>76%</td>
<td>0.072</td>
</tr>
</tbody>
</table>
Finding 3: Change in Teacher Practice

- Homework review: same total time, fewer problems
- Emphasis on problems that were difficult for students
- Addressing student errors

“I'm not having to go over every single problem. I can look at the big picture and say, ‘All right, this is how you did.’ Kids see it, I see it . . . you know, ‘Only 30 percent of the class got this right. We’re definitely looking at this question.’”

“We do more talking and discussing about the homework. . . . not just finding and fixing, but they’re able to discuss maybe a process or a step that they’re missing. . . . I think it’s really changed how we talk about homework, as opposed to right and wrong answers. But what we got wrong and why.”
Section 2: Conclusion

Homework CAN be improved

• Intervention needs to be easy for the public to digest - they care about homework
• Relatively easy for schools to adopt - modest intervention in an area of stable school
• *How* to do homework, not what or how much

Teachers’ practice CAN be changed

• When the intervention can be readily integrated into general practices
• When the right kind of data is available
• There is support on strategies to respond to the data
More On-going Studies

Replication study in North Carolina - IES funded, led by WestEd
  • Face-to-face training + local coach visits

National effectiveness study - IES funded, led by AIR
  • Face-to-face training + Google hang-out for support

Scale up study - DoE EIR, led by WPI
  • All virtual training + Instructional recommendation based on common errors
Thanks!

Questions?

mfeng@wested.org
Summary

The data we focused on is

- Individualized
- Generated by online learning systems
- Unobtrusive collected at scale
- Large amount (and noisy)
- Behavior or performance data

Best practices

- Identify useful data
- Become familiar with teachers’ and students’ experience of using the online platforms
- Understand the ways in which these data were collected and stored by the system
- Build explainable, actionable analytics that need low-effort for adoption and integration
- Check face validity and external validity of measures
ASSISTments Used for Textbook Work

The students benefit by knowing when they are doing a problem wrong allowing them to fix their approach before moving on to the next problem instead of continuing to do the work wrong.

Vicki Davis, 7th Grade Math Teacher, Blue Hills, ME
Design Workshops - where data became actionable

**Measures were put in front of faculty to better understand**
- how well they captured students’ learning strategies and behaviors
- whether they could be used to help faculty more effectively intervene with students

**Researchers + Faculty**
- Researchers learn from faculty on what they found meaningful and whether certain patterns that were identified had face validity
- An efficient touch-point for faculty to engage in data-intensive research activities

**Structured activities**
- Jointly interpreting data products
- Co-developing data products and follow up actions such as change ideas that faculty could implement using a data