Using Curriculum-Based Measures to Predict Math Performance on Statewide Assessments

Keri Menesses, M.A.
Louisiana State University
LAS*PIC Intern
Overview

• Background Information
• Previous Research
• Current Project
• Results and Implications
Background Information

• No Child Left Behind Act (NCLB, 2001)
  • Accountability standards
  • Standardized assessments of basic skills
    • Grades 3-8
    • Reading/Language Arts and Math
  • Consequences tied to test results
    • Students: summer school, retention
    • Teachers: salary cuts, job loss
    • Schools: loss of funding, accreditation, or administrative control
Background Information

• Statewide tests provide too little information, too late
  • Near end of school year
  • Untimely feedback
  • One-shot assessment
  • Do not specify areas of need
Background Information

• Curriculum-based measurement (CBM) addresses the issues of statewide testing
  • Can be administered frequently
    • Easy and efficient to administer and score
    • Provides feedback on progress in timely manner
  • Reliable and valid way of identifying specific academic concerns
  • Used for screening, establishing local norms, monitoring student progress, and evaluating intervention effectiveness
Previous Research

• Relationship between CBM and statewide tests
  • Most research investigates reading
  • Correlations range from $r = .44$ to $.79$ across eight states
  • The few studies that have investigated math have found similar correlations, ranging from $r = .11$ to $.83$
Previous Research

- Computational – foundational strategies and procedural knowledge
- Conceptual – reasoning and application of knowledge
  - Helwig, Anderson, & Tindall, 2002
  - Shapiro, Keller, Lutz, Santoro, & Hintze, 2006
  - Keller-Margulis, Shapiro, & Hintze, 2008
Helwig et al., 2002

• Created a conceptually-based math CBM
  • 11 problems
  • 8th graders
  • Correlated with computer-adaptive test (which paralleled Oregon state test) $r = .83$
  • The conceptual probe predicted with 81% accuracy which students would and would not score high enough to meet state standard
Shapiro et al., 2006

• Used Monitoring Basic Skills Progress - Math Computation and Math Concepts and Applications probes (Fuchs, Hamlett, & Fuchs, 1998)
  • Three benchmark periods
  • 3rd, 4th, and 5th graders
• Concepts and applications probes had somewhat higher correlations ($r$ ranged from .46 to .64) than computational probes ($r$ ranged from .41 to .53)
• ROC curve analysis
  • Both probes had similar results:
    • Overall correct classification 65%
Keller-Margulis et al., 2008

- 1st – 5th graders
- Math computation
  - $r$ ranged from .23 to .69
- Math concepts and applications
  - $r$ ranged from .25 to .66
- ROC curve analysis
  - Overall correct classification: 70% and 69%
  - Individual benchmark scores resulted in better diagnostic accuracy than slope data
Rationale

- Focus on testing not likely to cease
- Although relationship between math probes and test performance appears to be promising, further replication across states is required
- Recently published conceptual/application probes will be investigated
Primary Research Questions

• Will scores on computational and conceptual probes be significantly correlated with test scores? Which probe will be a stronger predictor of test performance?

• Hypotheses: Both probes will be moderately and significantly correlated with test scores; however, conceptual probes will have stronger correlations and better predictive ability due to similarity of items
Secondary Research Questions & Hypotheses

- Which CBM probes will be significantly correlated with final math grades? Which probes will better predict final grades?
  - There will be a significant relationship with both types of probes; however, concepts/applications probes will be more highly correlated and will have better predictive ability than computational probes.
Secondary Research Questions & Hypotheses

• Will teacher prediction of student test performance be significantly correlated with actual test performance?
  • There will be a significant correlation; however, teacher prediction will not add significantly to the variance associated with test scores above that explained by CBM scores
Participants

- 146 third, fourth, and fifth grade students from three schools across two districts
  - Students were required to give assent, return parental consent, and complete all math probes and the statewide assessment to be included in data analyses
  - Student demographics including age, ethnicity, grade, sex, and socioeconomic status were collected
# Participants

<table>
<thead>
<tr>
<th></th>
<th>School A</th>
<th>School B</th>
<th>School C</th>
</tr>
</thead>
<tbody>
<tr>
<td>District</td>
<td>Small, rural</td>
<td>Small, rural</td>
<td>Moderate, urban/suburban</td>
</tr>
<tr>
<td>Grade Levels</td>
<td>2-3</td>
<td>4-5</td>
<td>K-5</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>4%</td>
<td>16%</td>
<td>94%</td>
</tr>
<tr>
<td>Caucasian</td>
<td>89%</td>
<td>79%</td>
<td>3%</td>
</tr>
<tr>
<td>Asian</td>
<td>7%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0%</td>
<td>4%</td>
<td>3%</td>
</tr>
<tr>
<td>Alaskan American</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Socioeconomic Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free/Reduced Lunch</td>
<td>39%</td>
<td>54%</td>
<td>100%</td>
</tr>
<tr>
<td>Paid Lunch</td>
<td>61%</td>
<td>46%</td>
<td>0%</td>
</tr>
<tr>
<td>Educational Classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular Ed</td>
<td>98%</td>
<td>97%</td>
<td>97%</td>
</tr>
<tr>
<td>Special Ed</td>
<td>2%</td>
<td>3%</td>
<td>3%</td>
</tr>
</tbody>
</table>
Materials

• AIMSweb computation (M-CBM) and concepts/applications (M-CAP) probes
  • M-CBM: 72 problems - addition, subtraction, multiplication, and division
    • Hintze, Christ, & Keller (2002) median of 3 multiple skills probes is more dependable than one score
  • M-CAP: 29 or 30 problems - measurement, money, time, geometry, algebra, number relations, rounding, graphs, word problems, fractions, number patterns
Materials

- CBMs administered in winter (early Feb) and spring (late April) in group format according to manual
  - M-CBM: 2 minutes for third grade, 4 minutes for fourth and fifth grade
  - M-CAP: 8 minutes for all students
  - 30% were observed by research assistant
  - Procedural integrity = 100%
- Probes were be scored according to manual
  - 30% will be scored by research assistant
  - Interscorer reliability = 96%
Materials

- Statewide test: LEAP (4th grade) and iLEAP (3rd and 5th grade) were administered in mid-April
  - Five achievement levels - Advanced, Mastery, Basic, Approaching Basic, and Unsatisfactory
- Math portion of tests comprise six strands:
  (a) numbers and number relations; (b) algebra; (c) measurement; (d) geometry; (e) data analysis, probability, and discrete math; and (f) patterns, relations, and functions.
Materials

• Teacher reports
  • Teacher prediction of student performance on statewide test
    • Collected during winter administration of CBM probes
  • Teacher report of final math grades
    • Collected at end of school year
Preliminary Data Analyses

- Distributional properties
- Outliers
- Parametric assumptions
- Missing data
Results

• Pearson product-moment correlations

   Test Score

   Winter M-CBM  .32*
   Winter M-CAP  .55*
   Spring M-CBM  .30*
   Spring M-CAP  .53*

   *p<.001

• As predicted, the M-CAP probes were more highly correlated with performance on the LEAP and iLEAP
Results

- Hierarchical regression analyses:
  - The winter M-CAP helped explain more variance than the winter M-CBM alone
  - The spring M-CBM and the spring M-CAP both significantly account for variance in test scores
Results

- Hierarchical regression analyses:
  - Neither the winter nor the spring M-CBM probe significantly explained more variance in test scores than the other
  - Both the winter and spring M-CAP probes significantly accounted for variance in test scores
Results

• Teacher report was moderately and significantly correlated with the statewide assessment scores ($r = .67, p < .001$)

• Teacher reports of future student performance did in fact add significantly to the variance associated with test scores when entered after the probes

• Likewise, the probes explained additional unique variance in test scores when entered after teacher prediction
Results

- Pearson product-moment correlations

<table>
<thead>
<tr>
<th></th>
<th>Final Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter M-CBM</td>
<td>.25*</td>
</tr>
<tr>
<td>Winter M-CAP</td>
<td>.37**</td>
</tr>
<tr>
<td>Spring M-CBM</td>
<td>.25*</td>
</tr>
<tr>
<td>Winter M-CAP</td>
<td>.40**</td>
</tr>
</tbody>
</table>

*p < .01, **p < .001

- As predicted, the M-CAP probes were more highly correlated with final grades
Diagnostic Accuracy

• **Sensitivity** - percentage of students who were not successful on the LEAP/iLEAP and scored below the cut score on the math probe

• **Specificity** - the percentage of students who passed the LEAP/iLEAP and scored at or above the cut score on the math probe

• **Positive predictive power** - probability that the students who scored below the cut score on the CBM measure will score below Basic on the LEAP/iLEAP

• **Negative predictive power** - probability that students who scored at or above the cut score on the CBM probe will score in the Basic range or above on the LEAP/iLEAP
## Diagnostic Accuracy

<table>
<thead>
<tr>
<th>Predictor (CBM)</th>
<th>Criterion (Test)</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>True positive (sensitivity) ( A )</td>
<td>False positive (type I) ( B )</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>False negative (type II) ( C )</td>
<td>True negative (specificity) ( D )</td>
</tr>
</tbody>
</table>

- **Sensitivity** = \( \frac{A}{A+C} \)
- **Specificity** = \( \frac{D}{D+B} \)
- **PPP** = \( \frac{A}{A+B} \)
- **NPP** = \( \frac{D}{D+C} \)
- **Overall Correct Classification** = \( \frac{A+D}{A+B+C+D} \)
Diagnostic Accuracy

• Receiver operator characteristic (ROC) curve analyses
  • Graphically displays the tradeoff between sensitivity and specificity
  • Used to identify explicit cut scores on the math probes that are associated with passing or failing the LEAP/iLEAP
ROC curve graph
Pairwise Comparison of ROC curves: Winter
Pariwise Comparison of ROC curves: Spring
Comparison of 3\textsuperscript{rd} grade ROC curves
Comparison of 4th grade ROC curves
Comparison of 5th grade ROC curves

![ROC curves graph]
Diagnostic Accuracy

- Overall correct classification rates ranged from 61% to 73%
- As predicted, the M-CAP probes had higher overall correct classifications than the computational probes
### Overall Diagnostic Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Winter M-CBM</th>
<th>Winter M-CAP</th>
<th>Spring M-CBM</th>
<th>Spring M-CAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cut Score</td>
<td>40</td>
<td>10</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>Hit Rate</td>
<td>62%</td>
<td>72%</td>
<td>61%</td>
<td>73%</td>
</tr>
</tbody>
</table>
# Diagnostic Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Winter</th>
<th>Winter</th>
<th>Spring</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M-CBM</td>
<td>M-CAP</td>
<td>M-CBM</td>
<td>M-CAP</td>
</tr>
<tr>
<td>Cut Score</td>
<td>16</td>
<td>6</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Hit Rate</td>
<td>84%</td>
<td>80%</td>
<td>79%</td>
<td>76%</td>
</tr>
</tbody>
</table>
## Diagnostic Accuracy

<table>
<thead>
<tr>
<th>Grade</th>
<th>Winter M-CBM</th>
<th>Winter M-CAP</th>
<th>Spring M-CBM</th>
<th>Spring M-CAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cut Score</td>
<td>45</td>
<td>11</td>
<td>46</td>
<td>10</td>
</tr>
<tr>
<td>Hit Rate</td>
<td>67%</td>
<td>76%</td>
<td>68%</td>
<td>76%</td>
</tr>
</tbody>
</table>
## Diagnostic Accuracy

**5th grade**

<table>
<thead>
<tr>
<th></th>
<th>Winter</th>
<th>Winter</th>
<th>Spring</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-CBM</td>
<td>38</td>
<td>10</td>
<td>48</td>
<td>6</td>
</tr>
<tr>
<td>M-CAP</td>
<td>76%</td>
<td>76%</td>
<td>69%</td>
<td>74%</td>
</tr>
</tbody>
</table>

**Cut Score**

**Hit Rate**
Discussion

• Similar to the results of previous research, math CBM probes had moderate to strong relationships with outcomes on high-stakes assessments.
• Correlations between the math probes and statewide math test scores were consistent across winter and spring assessment periods.
• Correlations with test scores were consistently higher for the M-CAP probes ($r = .53$ to $.55$) than for the M-CBM probes ($r = .30$ to $.32$).
• Same pattern existed for correlations between the math probes and final math grades (M-CAP = .37-.40 and M-CBM = .25).
• This held true across all three grade levels.
Discussion

• Regression analyses:
• Adding a conceptual/applications measure offers a better explanatory model than does a computational measure alone
• Adding teacher prediction of future student performance explains more variance in state test scores than either probe alone
• Overall correct classification rates: 61% - 72%, which suggests that CBM metrics are appropriate measures for universally screening students
• As hypothesized, M-CAP probes had consistently higher rates of overall correct classification
Implications

• Math CBM probes can serve as efficient and inexpensive screening tools which can potentially identify a large group of students who are at-risk

• The progress monitoring capability of CBMs allows school personnel to have consistent and on-going data on the status of their students in regard to state standards, rather than waiting for the return of statewide testing results
Implications

• Knowing how many and which students may be at-risk can guide school administrators in implementing an intensive, short-term remediation program (i.e. tier two interventions such as small group instruction or peer tutoring) which focuses on teaching the skills necessary to be successful on statewide tests and in the general curriculum.
Limitations

• Only students who returned parental permission and who assented to participate were eligible for participation
  • It is fairly likely more responsible and higher-performing students participated in this study, which does not result in a fully representative sample of students
• Only students who had complete data sets were included in final data analyses: several students were absent or withdrew
• This study was conducted in only three schools across two districts in Louisiana
Conclusions

• The outcomes of this research study link one of the political pressures affecting educators today (accountability testing) with an evidence-based practice: universal benchmarking using curriculum-based measures

• Utilizing screening measures may promote proactive strategies rather than reactive approaches to education
Conclusions

• Due to the accountability laws currently in place and the ramifications associated with poor test results, it is unlikely that schools will cease to focus on test results.

• It is imperative to utilize tools that have the ability to forecast test performance as well as specify deficits that require remedial instruction, with the ultimate goal of increasing the number of students who perform successfully.