Developing Benchmark Goals and Cut-points for Risk: Odds of Achieving Subsequent Reading Goals

Pacific Coast Research Conference
Coronado, CA
February 4, 2011

Roland H. Good III
Dynamic Measurement Group, Inc.
University of Oregon
Kelly Powell-Smith
Ruth A. Kaminski
Dynamic Measurement Group, Inc.

Overview

- Panel:
The decision utility of educational assessment is enhanced when there are interpretive rules and recommendations for support associated with levels of performance. Benchmark goals provide a level of progress where students are judged to be making adequate progress. A cut point for risk indicates a level where a teacher can be confident the student will need additional intervention and support in order to make adequate progress. This panel will present alternative approaches to establishing benchmark goals and cut points for risk in the context of infants and toddlers, preschool students, and elementary students in kindergarten through sixth grade.

- Section:
One approach will examine the odds or likelihood of attaining important subsequent outcomes. The odds can be modeled with logistic regression and interpreted with receiver operator characteristic curves.

Benchmark Goal Study Research Questions

- The Benchmark Goal Study was designed to address three research questions:
  1. What levels of performance on DIBELS Next assessments predict a student is likely to score at or above the 40%ile on selected outcome measures?
  2. What levels of performance on DIBELS Next assessments predict a student is unlikely to score at or above the 40%ile on selected outcome measures?
  3. What are the correlations between DIBELS Next assessments and the Group Reading Assessment and Diagnostic Evaluation (GRADE), a criterion measure of reading proficiency that includes comprehension?

Participants

- Students recruited for the study were from 13 schools in five school districts representing five US regions.
- Participating school districts had a median of 10 years experience using DIBELS.
- Kindergarten through 6th grade students participated in DIBELS Next assessments \((n = 3,816 \text{ total}; 433 \text{ to } 569 \text{ per grade})\). The percentage at benchmark ranged from 65% - 79% across grades and times of year.
- Subsamples of students participated in testing with an external criterion measure (Group Reading Assessment and Diagnostic Evaluation; GRADE) \((n = 1257 \text{ total}; 103 \text{ to } 219 \text{ per grade})\). The GRADE subsample was 50% female on average across grades.
Participant Demographics

Figure 1. Racial/Ethnic Background

Figure 2: Parent-Reported Level of Education

Figure 3: Parent-Reported Household Income

Measures: DIBELS Next

- The measures included all *DIBELS Next* assessments. *DIBELS Next* assessments include:
  
  - Letter Naming Fluency
  - First Sound Fluency
  - Phoneme Segmentation Fluency
  - Nonsense Word Fluency Correct Letter Sounds and Whole Words Read
  - Oral Reading Fluency Words Correct, Accuracy, and Retell.
  - Daze Adjusted Score (DIBELS-maze)
  - DIBELS Composite Score
Measures: Group Reading Assessment and Diagnostic Evaluation (GRADE)

• Un-timed and group administered. Appropriate for students in preschool through grade 12.
• Five components and 16 subtests. Subtests combine to form the following composites:
  • Phonemic Awareness, Early Literacy Skills, Comprehension, Vocabulary, and Total Test.
  • We used the Total Test Raw Score for analyses.
• The GRADE has excellent reliability and validity for its intended purposes.
  • Reliability ranges from .77 to .98.
  • Correlation coefficients range from .69 to .86 with other group- and individually-administered achievement tests.

Procedures: Data Collection

• All Data were collected during the 2009-2010 school year.
• DIBELS Next assessments were administered at regular benchmark intervals by trained school personnel using standardized procedures.
• GRADE testing occurred in the spring at the end of the year and was conducted across two to three sessions. Total testing time ranged from 60 to 90 minutes. The GRADE was administered by trained school personnel and onsite coordinators.

DIBELS Composite Score

• For each grade and time of year, the DIBELS Next measures that correlate highly with later outcomes are combined into a DIBELS Composite Score.
• Each measures is weighted so that all contribute approximately equally to the DIBELS Composite Score.
  – Weighted scores have approximately equal standard deviations.
• The DIBELS Composite Score represents a rich and broad sample of behavior.
• The DIBELS Composite Score conveys that all of the aspects of reading proficiency are critical – a student whose DIBELS Composite Score is At or Above Benchmark is reading accurately, at an adequate rate, and attending to the meaning of the passage.
• The DIBELS Composite is highly correlated with a broad range of reading outcomes.

For Example: Third Grade DIBELS Composite Score. Benchmark Goal: 220
DIBELS Composite Score is the Best Indicator of Reading Proficiency

- DIBELS ORF is very good, DIBELS Composite Score is substantially better.
  - For example, beginning of year 3rd grade DORF Words Correct correlates with end of the year GRADE Total Score .66, which is very good.
  - Beginning of year 3rd grade DIBELS Composite Score correlates .73, explaining 10% more variance than DORF alone.
- DIBELS Composite Score beats the single best DIBELS Next measure at almost every grade and time of year.
- DIBELS Composite Score provides a more complete sample of reading behavior than any single measure.

Predictive Validity of the DIBELS Composite Score Compared to the Best Single DIBELS Measure by Grade and Time of Year

<table>
<thead>
<tr>
<th>Kindergarten</th>
<th>Best single measure</th>
<th>DIBELS Composite Score</th>
<th>Middle of year</th>
<th>Best single measure</th>
<th>DIBELS Composite Score</th>
<th>End of year</th>
<th>Best single measure</th>
<th>DIBELS Composite Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADE Total score</td>
<td>0.52</td>
<td>0.50</td>
<td>0.47</td>
<td>0.48</td>
<td>0.40</td>
<td>0.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>End of year DIBELS Composite Score</td>
<td>0.43</td>
<td>0.52</td>
<td>0.65</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First grade</td>
<td>GRADE Total score</td>
<td>0.43</td>
<td>0.55</td>
<td>0.64</td>
<td>0.70</td>
<td>0.75</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>End of year DIBELS Composite Score</td>
<td>0.71</td>
<td>0.73</td>
<td>0.83</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second grade</td>
<td>GRADE Total score</td>
<td>0.69</td>
<td>0.75</td>
<td>0.76</td>
<td>0.80</td>
<td>0.73</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>End of year DIBELS Composite Score</td>
<td>0.81</td>
<td>0.81</td>
<td>0.87</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third grade</td>
<td>GRADE Total score</td>
<td>0.66</td>
<td>0.73</td>
<td>0.67</td>
<td>0.78</td>
<td>0.66</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>End of year DIBELS Composite Score</td>
<td>0.86</td>
<td>0.88</td>
<td>0.86</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fourth grade</td>
<td>GRADE Total score</td>
<td>0.76</td>
<td>0.80</td>
<td>0.76</td>
<td>0.80</td>
<td>0.75</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>End of year DIBELS Composite Score</td>
<td>0.86</td>
<td>0.89</td>
<td>0.87</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fifth grade</td>
<td>GRADE Total score</td>
<td>0.69</td>
<td>0.76</td>
<td>0.64</td>
<td>0.76</td>
<td>0.66</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>End of year DIBELS Composite Score</td>
<td>0.84</td>
<td>0.86</td>
<td>0.86</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sixth grade</td>
<td>GRADE Total score</td>
<td>0.64</td>
<td>0.71</td>
<td>0.59</td>
<td>0.68</td>
<td>0.61</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>End of year DIBELS Composite Score</td>
<td>0.86</td>
<td>0.90</td>
<td>0.87</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What is the Purpose of Benchmark Goals and Screening for Risk in Education?

Different standards, procedures, and requirements are necessary if our purpose is:

1. To quickly identify students that are likely to need additional support to prevent later academic difficulty.
   - To specify important and meaningful future goals—a level of skills at a point in time where we can change the odds to being in favor of an individual’s meeting subsequent goals.

We are troubled by the purpose of identifying true Tier 3 students. We think the future is not set. Tier 3 is not a characteristic of the student. There are no true Tier 3 students. Tier 3 is a level of support necessary for the student to make adequate progress. No fate but what we make.

Our purpose is to prevent reading difficulty and enhance reading outcomes by providing targeted, differentiated instruction early.
Goal: Adequate Reading Skills

- Adequate reading skills should generalize across different state, national, and published reading tests.
- Adequate reading skills are not a normative decision, but are a socio-political judgment.
- The 40th percentile or above on a high quality, nationally norm-referenced test can serve as an approximation for adequate reading performance.
- Students at or above the 40th percentile on a high quality, nationally norm-referenced test are on track to be rated Basic or above on NAEP.
- We used the Group Reading Assessment and Diagnostic Evaluation (GRADE) in our initial research to provide an initial approximation of adequate reading skills.

### Evidence Base, Score Level, Likely Need for Support

<table>
<thead>
<tr>
<th>Odds of achieving subsequent early literacy goals</th>
<th>Score level</th>
<th>Likely need for support to achieve subsequent early literacy goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>80% to 90%</td>
<td>At or Above Benchmark scores at or above the benchmark goal</td>
<td>Likely to Need Core Support</td>
</tr>
<tr>
<td>40% to 60%</td>
<td>Below Benchmark scores below the benchmark goal and at or above the cut point for risk</td>
<td>Likely to Need Strategic Support</td>
</tr>
<tr>
<td>10% to 20%</td>
<td>Well Below Benchmark scores below the cut point for risk</td>
<td>Likely to Need Intensive Support</td>
</tr>
</tbody>
</table>

The fundamental rationale for benchmark goals and screening decisions is based on the odds of achieving subsequent early literacy goals.
Primary Design Specifications for DIBELS Goals and Cut Points for Risk

• **Primary Specification**: At or Above Benchmark Decision on initial (screening) DIBELS assessment should provide favorable odds (80% -- 90%) of achieving subsequent reading outcomes. Benchmark Goal should provide a level where we are reasonably confident the student is making adequate progress.

• **Below Benchmark Decision** on initial DIBELS assessment should provide 50 – 50 odds (40% -- 60%) of achieving subsequent reading outcomes. Below the Benchmark Goal but above the Cut Point should provide a zone of uncertainty where we don’t know if the student is making adequate progress or not.

• **Well Below Benchmark Decision** on initial DIBELS assessment should provide low odds (10% -- 20%) of achieving subsequent reading outcomes — unless intensive intervention is implemented. Below the Cut Point should provide a zone where we are reasonably confident the student will not make adequate progress -- unless we provide additional support.

Secondary Specifications for Benchmark Goals and Cut Points

• Marginal percents for the predictor close to marginal percents for the outcome.
  – The sample for the Benchmark Goal Study was a relatively high performing sample.
  – We tried have them appear equally high performing on DIBELS Next and the GRADE.

• Logistic Regression Analysis
  – Logistic regression predicted odds of about 60% or better at the exact goal score.
  – Logistic regression predicted odds of about 40% or below at the exact cut point for risk score.

Other Considerations
DIBELS Goals and Cut Points

• Other considerations
  – Receiver Operator Characteristic Curve (ROC) analysis with large area under curve
  – Other metrics for decision utility
    • sensitivity,
    • specificity,
    • percent correct classification,
    • kappa
  – Coherent pattern of goals across measures and grades.

Setting Benchmark Goals and Cut Points for Risk

1. Examine scatterplot illustrating the relation between the screening assessment (earlier assessment or predictor) and the outcome assessment (later assessment).
   – DIBELS is a step-by-step model, so the outcome of one step is the predictor of the next step.
2. Examine the table of counts for each zone of the scatterplot.
3. Primary: Consider odds of students with each screening decision achieving goal.
4. Secondary: Consider marginal percents
5. Secondary: Consider logistic regression analysis
6. Other: Consider ROC curve and decision utility metrics
7. Other: Consider the overall pattern of goals and cut points.
Example Analysis Detail

Third Grade DIBELS Composite Score for Beginning (DCS3b) to Middle of Year (DCS3m)

DCS3b Screening Decision:
- Likely to need intensive support
- Likely to need strategic support
- Likely to need core support

DCS3m Outcome:
- At or Above Benchmark
- Below Benchmark
- Well Below Benchmark

Likely to need intensive support
Likely to need strategic support
Likely to need core support

Marginal total
350
57
83
490

Primary consideration: Odds of achieving goal
- Core support beginning of year screening decision:
  324 of 349 students achieve the middle of year goal, or 93% odds.
- Strategic support: 22 of 47 students achieve the middle of year goal, or 47% odds.
- Intensive support: 4 of 94 students achieve the goal, or 4% odds.

Secondary consideration: Marginal Percents
Also Considered Marginal Percents

DCS3b Screening Decision:

DCS3m Outcome: Strategic support decision
Core support decision

DCS3b Screening Decision:
Marginal percent
Intensive support decision
Strategic support decision
Core support decision
Marginal total
Marginal percent

<table>
<thead>
<tr>
<th>DCS3m Outcome</th>
<th>Intensive support decision</th>
<th>Strategic support decision</th>
<th>Core support decision</th>
<th>Marginal total</th>
<th>Marginal percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>At or Above Benchmark</td>
<td>4</td>
<td>22</td>
<td>324</td>
<td>350</td>
<td>71%</td>
</tr>
<tr>
<td>Below Benchmark</td>
<td>20</td>
<td>16</td>
<td>21</td>
<td>57</td>
<td>12%</td>
</tr>
<tr>
<td>Well Below Benchmark</td>
<td>70</td>
<td>9</td>
<td>4</td>
<td>83</td>
<td>17%</td>
</tr>
</tbody>
</table>

- Percent At or Above Benchmark at beginning of year is very close to the percent At or Above Benchmark in the middle of the year.
- Desirable for the screening decision to identify about the same percent of students that are expected on the outcome.

Logistic Regression Estimates Odds of Adequate Outcomes for each Score

- Blue diamonds are moving proportion with adequate outcome.
- Red line is logistic regression estimated odds of adequate outcomes.

DIBELS is a Step-by-Step Model:
Beginning to Middle; Middle to End;

- Mastering each step puts the odds in favor of mastering the next step.
  - At or Above Benchmark: Odds are generally 80% to 90% of achieving subsequent benchmark goals and important reading outcomes. **Student is likely to make adequate progress with effective core instruction.**
  - Below Benchmark: Odds are generally 40% to 60% of achieving subsequent benchmark goals and important reading outcomes. **Student is likely to need strategic support to make adequate progress.**
  - Well Below Benchmark: Odds are generally 10% to 20% of achieving subsequent benchmark goals and important reading outcomes. **Student is likely to need intensive support to make adequate progress.**

- Contiguous Continuity. Each step is a continuous process with a strong linkage. Each step is contiguous with the next step.
End of Year Benchmark Goals

DIBELS® Next: Summary of Benchmark Goals and Cut Points for Risk

Third Grade DIBELS Composite Score for Beginning of Year (DCS3b) and Middle of Year (DCS3m)
- .91 correlation to DIBELS Composite Score at Middle of Year

Third Grade DIBELS Composite Score for Middle of Year (DCS3m) and End of Year (DCS3e)
- .90 correlation to DIBELS Composite Score at End of Year
Third Grade DIBELS Composite Score for End of Year (DCS3e) and GRADE Total Raw Score (gtotr3e)

- .75 correlation to GRADE Total Raw Score at End of Year

Receiver Operator Characteristic Curve

- Larger area under the curve indicates favorable trade off of sensitivity and specificity.
- Decision points in the upper left bend of the curve indicate a favorable balance of sensitivity and specificity.

Other Decision Utility Metrics End of Third Grade

We are troubled by the terminology. We think a “True Positive” is actually a student for whom we were not effective in changing the future.

<table>
<thead>
<tr>
<th>Role</th>
<th>At or Above Benchmark outcome</th>
<th>Well Below Benchmark outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Core support decision</td>
<td>Intensive support decision</td>
</tr>
<tr>
<td>True Negative</td>
<td>123</td>
<td>134</td>
</tr>
<tr>
<td>False Negative</td>
<td>14</td>
<td>26</td>
</tr>
<tr>
<td>True Positive</td>
<td>37</td>
<td>25</td>
</tr>
<tr>
<td>False Positive</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>.73</td>
<td>.49</td>
</tr>
<tr>
<td>Specificity</td>
<td>.90</td>
<td>.99</td>
</tr>
<tr>
<td>Negative Predictive Power</td>
<td>.90</td>
<td>.84</td>
</tr>
<tr>
<td>Positive Predictive Power</td>
<td>.74</td>
<td>.93</td>
</tr>
<tr>
<td>Accurate Classification</td>
<td>.86</td>
<td>.85</td>
</tr>
<tr>
<td>Kappa</td>
<td>.63</td>
<td>.56</td>
</tr>
</tbody>
</table>

Early Intervention and Prevention are Active Ingredients Between Screening and Outcomes

- The effectiveness of the school-wide system of instruction can change the odds.
  - Differences in the effectiveness of Tier 1 instruction and Tier 2 & 3 intervention change the underlying relation between screener and outcome.
  - Less effective school-wide system Tier 1 instruction can decrease the odds of achieving subsequent early literacy goals for students who are at or above benchmark.
  - Increasing the effectiveness of Tier 2 & 3 intervention can increase the odds of achieving subsequent early literacy goals for students who are at risk.
Building Futures

• Key Point: The student’s outcome is unknown and not fixed at the time of the screening. Instead, the outcome is the result of the targeted, differentiated instruction and intervention we provide as a direct result of the screening information.

• **Our instructional goal is to ruin screening predictions**

• For Example: If a child screens as at high risk on a measure of early literacy skills in Kindergarten, we know they are likely to need additional instructional support to be successful. Their later outcome, their reading skills in first grade for example, are a direct result of the targeted, differentiated instruction and early intervention that we provide.