Best Practices in Oral Reading Fluency Administration

Daniel Anderson, Julie Alonzo & Gerald Tindal
Behavioral Research & Teaching
College of Education - UO
Disclaimer

• We are from Behavioral Research and Teaching (BRT), which is where easyCBM was born.
• The research reported here compares easyCBM recommendations for Oral Reading Fluency administration, with those of other vendors.
Our Study

• Explore the effect of different administration procedures on the reliability of Oral Reading Fluency (ORF) passages.

Why?

• Different test vendors provide different recommendations, which have practical and psychometric repercussions. Our aim is to provide research evidence to inform decisions.
BACKGROUND: ORF
This is Tom's first year of playing on a team at school. He is on a basketball team. Tom's dad played on a team in high school when he was younger. Tom wants to be just like his dad. Every day after school, Tom goes home. He practices throwing the ball in the hoop. He also practices dribbling with both hands. He is getting better every day. At first, he was lucky if he made two baskets, now he can make almost every shot he takes. Sometimes Tom's dad comes outside. He helps him with his skills. Tom works hard on what his dad shows him. He can now steal the basketball away from his dad. On some days, Tom and his dad play one-on-one.

Tom looks forward to his daily practices with his coach and teammates. All of his friends from school are on his team. Tom is still young, and both boys and girls play on the same team. Tom makes sure that he always passes the ball to everyone on his team. This way, they all get practice every day. He wants to make sure that all his friends get a chance to shoot the ball. Tom loves being on the same team as all of his friends. When they have recess at school, they all play basketball. His team is going to be playing against a very good team soon. He hopes that he and his friends can beat the other team. But Tom knows that playing fairly is the real goal.
<table>
<thead>
<tr>
<th>What’s the point?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Brain internalizes ‘rules’ about grapheme (written words) / phoneme (sound units) relationships.</td>
</tr>
<tr>
<td>• Repeated exposure to words = move to sight word vocabulary bank</td>
</tr>
<tr>
<td>• +/-150 CWPM needed to read with comprehension</td>
</tr>
</tbody>
</table>
Regular ORF administration provides teachers with a powerful set of data from which decisions can be based.
## Median Score versus Single Probe

<table>
<thead>
<tr>
<th>Median Score</th>
<th>Single Probe</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Approach includes the administration of three passages in succession.</td>
<td>• Approach includes administering a single passage, and taking the score the student receives as “the” score.</td>
</tr>
<tr>
<td>- High and low scores dropped (i.e., median maintained)</td>
<td>• Recommended by easyCBM</td>
</tr>
<tr>
<td>• Recommended by numerous vendors (e.g., DIBELS, AIMSweb, etc.)</td>
<td></td>
</tr>
</tbody>
</table>
### Overview

- **Passage comparability**
- Standard error of measurement

### Study overview and results

### Discussion

- Practical Considerations: resources devoted to testing
- Intra-construct reliability versus inter-construct reliability
- Consequences of the decisions made with the data
Passage comparability

• ORF probes are routinely used to evaluate growth

• The *validity* of fluency-based growth estimates *depend upon* adequate passage comparability
Example
### Passage comparability

<table>
<thead>
<tr>
<th><strong>Median Score</strong></th>
<th><strong>Single Passage</strong></th>
</tr>
</thead>
</table>
| Generally “safer”  
  - Unusually **high** or **low** scores (easy or difficult passages) will often be dropped  
| Importance of passage comparability becomes more pronounced |
| Passage comparability still critical, but perhaps not as much so as with a single passage approach  
| Decisions are based off a single passage, so if that passage is not comparable to others, the validity of educational decisions becomes threatened |
Passage Comparability: Development consideration

• Passages can be more or less comparable depending on the procedures followed during development (Poncy, Skinner, & Axtell, 2005).

• Some (e.g., Francis et al., 2008) have recommended equating ORF passages to increase comparability
  – This would require test administrators to use a lookup table or enter the data into a computer to obtain an ORF scale score.
## Overview

- Passage comparability
- **Standard error of measurement**

## Study overview and results

## Discussion

- Practical Considerations: resources devoted to testing
- Intra-construct reliability versus inter-construct reliability
- Consequences of the decisions made with the data
Standard Errors

• Standard error of measurement – flip side of reliability
• Lower standard errors (obviously) lead to higher precision and higher reliability of measurement
• Measurement errors compounded when multiple measures are used to measure growth.
Example: Compounded Errors

Notice the substantial difference in the steepness of the two slopes highlighted in orange.

Error around the test administration

The slope from one point to another may thus vary substantially based on error.
Standard Error of Measurement

<table>
<thead>
<tr>
<th>Median Score</th>
<th>Single Passage</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Will nearly always produce lower SEM</td>
<td>• More efficient administration</td>
</tr>
</tbody>
</table>

Studies

- Poncy, Skinner, & Axtell, 2005: 5 to 7 wcpm
- Christ & Silberglitt, 2007: Median \( \cong 10 \) wcpm, varied across grades 1-5, ranging from 4 to 15 wcpm
- Poncy, Skinner, & Axtell, 2005: 12 to 18 wcpm, depending on the construction of the passage
## Overview

- Passage comparability
- Standard error of measurement

## Study overview and results

### Discussion

- Practical Considerations: resources devoted to testing
- Intra-construct reliability versus inter-construct reliability
- Consequences of the decisions made with the data
Study overview and results

Evaluating the reliability of ORF under different measurement conditions

• Measures and Study Sample
• Analytic Methodology
  – Generalizability Theory
• Results
  – What do they mean?
Measures

- **easyCBM passage reading fluency** measures, grades 1-5.
- Piloted and analyzed with Analysis of variance (ANOVA).
- **Development details**: Alonzo & Tindal (2007)
- **Validity studies**: Jamgochian et al., 2010; Sáez et al., 2010; and Lai et al., 2010.
Study Sample

• Small \( n \), but generally acceptable statistical power nonetheless (exception at grade 5).
• Convenience sample from Pacific NW
  – Data analyzed were collected as part of a larger study
• Students administered a battery of easyCBM assessments (specific number varied by grade).
• Gathered on two occasions, one week apart.
Study Design

<table>
<thead>
<tr>
<th>Grade</th>
<th>Total n</th>
<th>Condition</th>
<th>Test forms: Day 1</th>
<th>Test forms: Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
<td>1</td>
<td>11-13</td>
<td>13-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>11-13</td>
<td>11-13</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>1</td>
<td>13-12-11</td>
<td>13-11-12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>11-12-13</td>
<td>12-13-11</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>1</td>
<td>16-15-14</td>
<td>16-14-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>14-15-16</td>
<td>15-16-14</td>
</tr>
<tr>
<td>4</td>
<td>39</td>
<td>1</td>
<td>13-12-11</td>
<td>13-11-12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>11-12-13</td>
<td>12-13-11</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>1</td>
<td>8-9-10-12</td>
<td>9-10-8-12</td>
</tr>
</tbody>
</table>

Note. For each grade, roughly half the sample was assigned to each condition. Data were combined across conditions for all analyses.
Analysis: Generalizability Theory

• Method for estimating the variance associated with different **facets** of the measurement process.

• G- and D-Study components
  – **G-Study**: Estimate variance components
  – **D-Study**: Estimate how the reliability would change under different **levels** of each facet.
Potential Facets in ORF

- Assessor
- Test form
- Testing Occasion
- Order of test forms
- Testing location
- etc.

Note. The object of measurement (in this case Student) is not generally referred to as a facet in Generalizability Theory.

Fully crossed, two-facet design used in this study. One analysis presented for each of grades 1-5.
Because the design was fully crossed, the G-Study portion estimated the variance associated with persons and each facet, as well as all interactions.
Note that we will use **absolute** rather than **relative** error variances.

### Study Variance Components

#### Absolute Error Variance

\[
\sigma^2_{\text{Absolute}} = \frac{\sigma^2_F}{n'_F} + \frac{\sigma^2_O}{n'_O} + \frac{\sigma^2_{F\cdot O}}{n'_F n'_O} + \frac{\sigma^2_{P\cdot F\cdot O\cdot e}}{n'_F n'_O}
\]

#### Relative Error Variance

\[
\sigma^2_{\text{Relative}} = \frac{\sigma^2_{P\cdot F}}{n'_F} + \frac{\sigma^2_{P\cdot O}}{n'_O} + \frac{\sigma^2_{P\cdot F\cdot O\cdot e}}{n'_F n'_O}
\]

Error terms not included in relative variance components.
**D-Study**

- How might the error variance change with **different levels** of each facet?
- Estimates obtained in a similar fashion to the Spearman-Brown Prophecy formula. Substitute in the level for the particular facet you are interested in.

\[
\sigma_{\text{Absolute}}^2 = \frac{\sigma_F^2}{n'_F} + \frac{\sigma_O^2}{n'_O} + \frac{\sigma_{FO}^2}{n'_F n'_O} + \frac{\sigma_{PF}^2}{n'_F} + \frac{\sigma_{PO}^2}{n'_O} + \frac{\sigma_{PF,FO,e}^2}{n'_F n'_O}
\]
## Results: G-Study

### Variance Components for G-Theory Analyses

<table>
<thead>
<tr>
<th>Grade</th>
<th>Persons</th>
<th>Forms</th>
<th>Occasion</th>
<th>Persons*Forms</th>
<th>Persons*Occasion</th>
<th>Forms*Occasion</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2143.91 (.95)</td>
<td>8.43 (.00)</td>
<td>20.32 (.01)</td>
<td>35.61 (.02)</td>
<td>9.76 (.00)</td>
<td>0.00 (.00)</td>
<td>32.39 (.01)</td>
</tr>
<tr>
<td>2</td>
<td>1306.29 (.88)</td>
<td>5.87 (.00)</td>
<td>16.44 (.01)</td>
<td>26.17 (.02)</td>
<td>29.24 (.02)</td>
<td>4.98 (.00)</td>
<td>94.12 (.06)</td>
</tr>
<tr>
<td>3</td>
<td>1237.18 (.82)</td>
<td>21.83 (.02)</td>
<td>36.58 (.02)</td>
<td>56.67 (.04)</td>
<td>83.81 (.06)</td>
<td>3.52 (.00)</td>
<td>61.07 (.04)</td>
</tr>
<tr>
<td>4</td>
<td>1363.10 (.88)</td>
<td>0.00 (.00)</td>
<td>65.52 (.04)</td>
<td>31.15 (.02)</td>
<td>15.25 (.01)</td>
<td>7.90 (.01)</td>
<td>71.91 (.05)</td>
</tr>
<tr>
<td>5</td>
<td>621.75 (.79)</td>
<td>26.74 (.03)</td>
<td>18.96 (.02)</td>
<td>0.00 (.00)</td>
<td>9.46 (.01)</td>
<td>0.00 (.00)</td>
<td>108.55 (.14)</td>
</tr>
</tbody>
</table>

*Note.* Proportion displayed in parentheses. Residual term represents a person by form by occasion interaction.

Overall, between 79-95% of the variance in the measurement process was associated with persons (i.e., the object of measurement). Also note how the variance associated with persons generally decreases as grade level increases. Very little variance associated with test forms. Slightly more variance associated with occasions, but still quite modest. Generally more variance associated with person by form interactions than with forms individually. With the exception of grade 3, there is no variance associated with a person by occasion interaction.
## Results: D-Study

### Absolute Standard Errors

Predicted absolute standard errors by administration practice

<table>
<thead>
<tr>
<th>Reliability index</th>
<th>Grade</th>
<th>$n$ Occasions</th>
<th>$n$ Forms</th>
<th>D studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>10.32</td>
<td>8.26</td>
<td>7.45</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>13.85</td>
<td>11.67</td>
<td>10.85</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>12.80</td>
<td>9.80</td>
<td>8.57</td>
</tr>
</tbody>
</table>

Notice that the 3-4 wcpi by $SE, \sigma(\Delta_{pi})$ instead of 1.

Approximately a 2-3 point reduction was observed by increasing the occasion.

Another reduction was observed by the second testing occasion.
### Results: D-Study

#### Absolute Dependability Coefficients

<table>
<thead>
<tr>
<th>Phi, Φ</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.95</td>
<td>.88</td>
<td>.82</td>
<td>.88</td>
<td>.79</td>
</tr>
<tr>
<td>2</td>
<td>.97</td>
<td>.92</td>
<td>.87</td>
<td>.91</td>
<td>.87</td>
</tr>
<tr>
<td>3</td>
<td>.98</td>
<td>.94</td>
<td>.88</td>
<td>.92</td>
<td>.91</td>
</tr>
<tr>
<td>4</td>
<td>.98</td>
<td>.95</td>
<td>.89</td>
<td>.93</td>
<td>.92</td>
</tr>
<tr>
<td>5</td>
<td>.98</td>
<td>.96</td>
<td>.93</td>
<td>.95</td>
<td>.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D studies</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>2</th>
</tr>
</thead>
</table>

Dependability coefficients increase modestly by using three forms rather than one.

A similar increase was observed during the second testing occasion.

Absolute dependability coefficients for a single test form taken during a single occasion, were all modest to high.
Dependability Coefficients

Absolute dependability coefficients for a sample of two grades. Figure displays how the dependability coefficients were predicted to change based on various conditions of measurement. Each line represents a different number of testing occasions.
What do the results mean?

• Overall, the reliability for a single form on a single occasion was quite good.
  – Standard errors quite low
• Increasing to 3 forms universally increased reliability and decreased SEM

**Study limitation:** Increases in reliability estimated by moving to 3 forms likely overestimates the effect of a median score approach. Analysis here assumes the information from all 3 forms would be used. Median score approach discards data from 2 of 3 forms.
### Discussion

- Passage comparability
- Standard error of measurement

### Study overview and results

**Discussion**

- **Practical Considerations: resources devoted to testing**
- Intra-construct reliability versus inter-construct reliability
- Consequences of the decisions made with the data
ORF Resource Allocation

Hypothetical Example

- Imagine we’re in a large school district, with approximately **10,000 students** in grades 1-5.
- The district has formally adopted a response to intervention plan, including **seasonal benchmark screenings** for all students.
- You are the test coordinator for the district, and must make some decisions.
## Resource Allocation

<table>
<thead>
<tr>
<th>Median Score Approach</th>
<th>Single Passage Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 3 minutes of testing per student (not including transition times)</td>
<td>• 1 minute of testing per student (not including transition times)</td>
</tr>
<tr>
<td>• = 30,000 minutes of testing (~500 hours) <strong>district-wide</strong> for a single administration</td>
<td>• = 10,000 minutes of testing (~167 hours) <strong>district-wide</strong> for a single administration</td>
</tr>
<tr>
<td>• * 3 testing occasions (fall, winter, and spring) = 90,000 minutes (~1,500 hours) <strong>district-wide</strong></td>
<td>• * 3 testing occasions (fall, winter, and spring) = 30,000 minutes (~500 hours) <strong>district-wide</strong></td>
</tr>
</tbody>
</table>
Resource Allocation

• Median score approaches will nearly always have higher reliability, and lower standard errors.

However, median score approaches also require a substantial increase in the amount of testing time when aggregated across a district. The overall testing costs are thus also increased.

**Question:** Are the increases in technical quality worth the financial costs and increased testing time?
Discussion

- Passage comparability
- Standard error of measurement

Study overview and results

Discussion

- Practical Considerations: resources devoted to testing
- **Intra-construct reliability versus inter-construct reliability**
- Consequences of the decisions made with the data
Inter-Construct Validity

- Fluency is a single facet of reading.
- Other facets include
  - Phonemic awareness
  - Phonics
  - Vocabulary
  - Comprehension

When a median score approach is used, **additional time** is dedicated to the assessment of fluency (relative to a single passage approach). *Is the additional time at the expense of assessing other reading constructs?*
### Discussion

- Passage comparability
- Standard error of measurement

### Study overview and results

**Discussion**

- Practical Considerations: resources devoted to testing
- Intra-construct reliability versus inter-construct reliability
- **Consequences of the decisions made with the data**
**Decision Making**

**What type of decision will be made?**

- **High-Stakes**
  - e.g., referral to special programs

- **Low-Stakes**
  - progress-monitoring

Which administration practice is most appropriate for each context?
Decision Making

• In the end the decision for a median-score approach versus a single-probe approach must balance:
  – Need for technical adequacy (e.g., decision to be made)
  – Practical repercussions of assessment approach (financial cost, time devoted to fluency assessment, etc.)
<table>
<thead>
<tr>
<th>Median Score</th>
<th>Single Probe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased reliability</td>
<td>Increased efficiency</td>
</tr>
<tr>
<td>Lower standard errors</td>
<td>Reliability and standard errors still generally within acceptable range</td>
</tr>
<tr>
<td>Better for high-stakes decisions</td>
<td>Better for benchmarking?</td>
</tr>
</tbody>
</table>
References


Funding Source

Funds for the dataset used in this presentation came from a federal grant awarded to the UO from the U.S. Department of Education, Institute for Education Sciences:


The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.
Thanks!

• Daniel Anderson, Behavioral Research & Teaching
  – daniela@uoregon.edu