Best Practices in Oral Reading Fluency Administration

Daniel Anderson, Julie Alonzo & Gerald Tindal

Behavioral Research & Teaching

College of Education - UO

BRT

behavioral research & teaching

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.

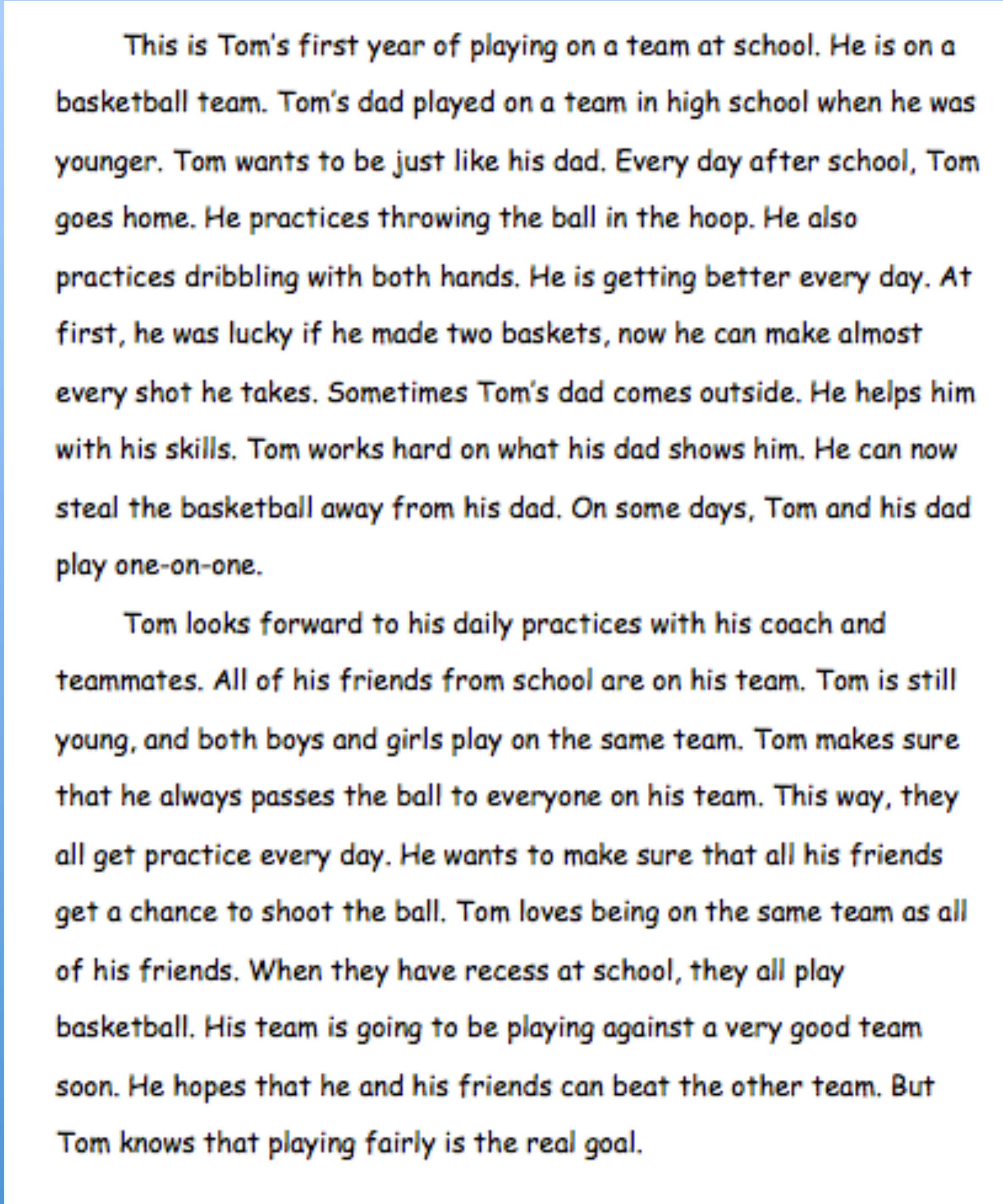
Our Study

- 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



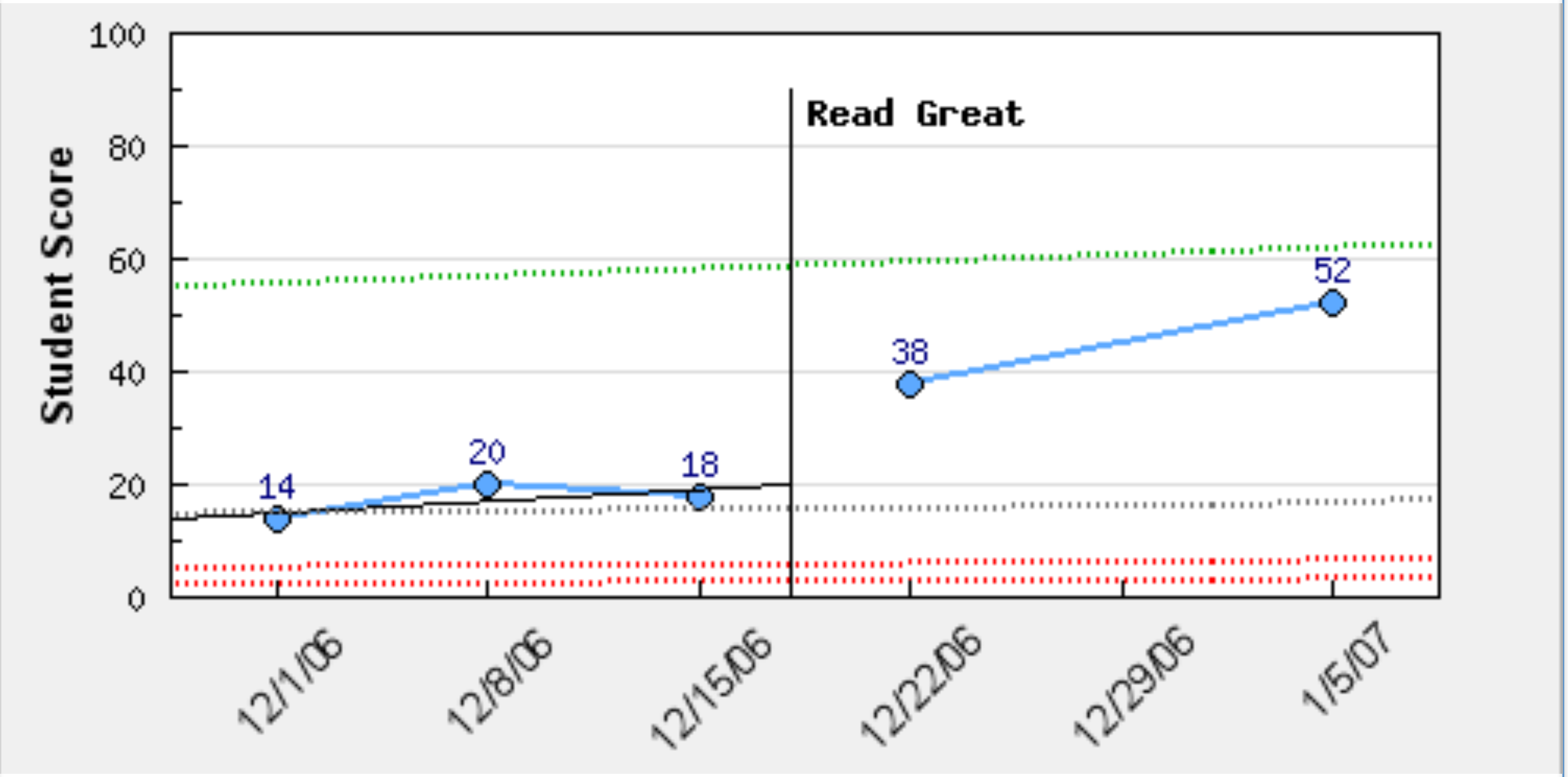
This is Tom's first year of playing on a team at school. He is on a basketball team. Tom's dad played on a team in high school when he was younger. Tom wants to be just like his dad. Every day after school, Tom goes home. He practices throwing the ball in the hoop. He also practices dribbling with both 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.

What’s the point?

- Brain internalizes ‘rules’ about grapheme (written words) / phoneme (sound units) relationships.
- Repeated exposure to words = move to sight word vocabulary bank
- +/-150 CWPM needed to read with comprehension

Repeated Measures



Bottom line

Regular ORF administration provides teachers with a powerful set of data from which decisions can be based.

LOST
CONFUSED
UNSURE
UNCLEAR
PERPLEXED

Median Score versus Single Probe

Median Score

- Approach includes the administration of three passages in succession.
  - High and low scores dropped (i.e., median maintained)
- Recommended by numerous vendors (e.g., **DIBELS**, **AIMSweb**, etc.)

Single Probe

- Approach includes administering a single passage, and taking the score the student receives as “the” score.
- Recommended by **easyCBM**

Overview

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

Study overview and results

Discussion

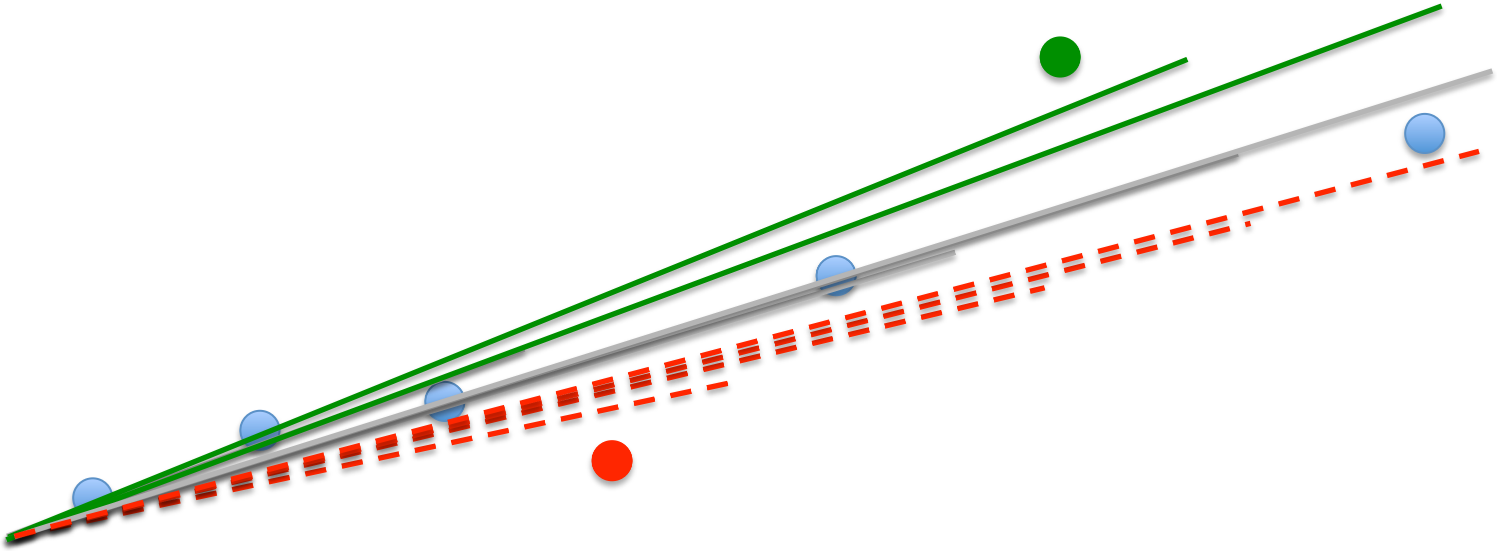
- Practical Considerations: resources devoted to testing
- Intra-construct reliability versus inter-construct reliability
- Consequences of the decisions made with the data

BRT
behavioral research & teaching

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

Median Score

- Generally "safer"
  - Unusually **high** or **low** scores (easy or difficult passages) will often be dropped
- Passage comparability still critical, but perhaps not as much so as with a single passage approach

Single Passage

- Importance of passage comparability becomes more pronounced
- 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

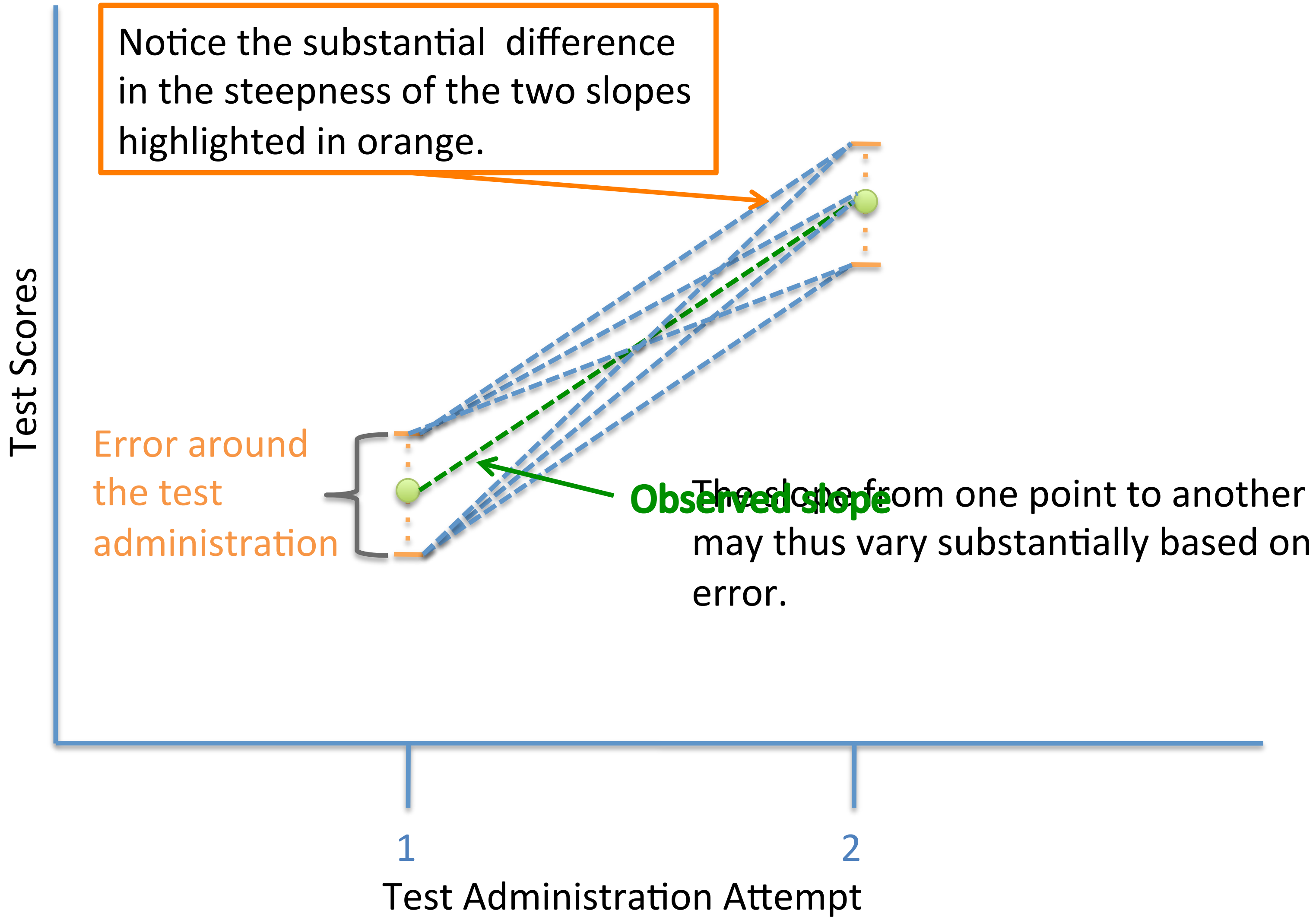
- Practical Considerations: resources devoted to testing
- Intra-construct reliability versus inter-construct reliability
- Consequences of the decisions made with the data

BRT
behavioral research & teaching

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.

The slope from one point to another may thus vary substantially based on error.

Error around the test administration
Observed slope
Test Scores
1
2
Test Administration Attempt

Standard Error of Measurement

Median Score

- Will nearly always produce **lower SEM**

Single Passage

- More efficient administration

Studies

- Poncy, Skinner, & Axtell, 2005: 5 to 7 wcpm
- Christ & Silberglitt, 2007:

  Median ≅ 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?

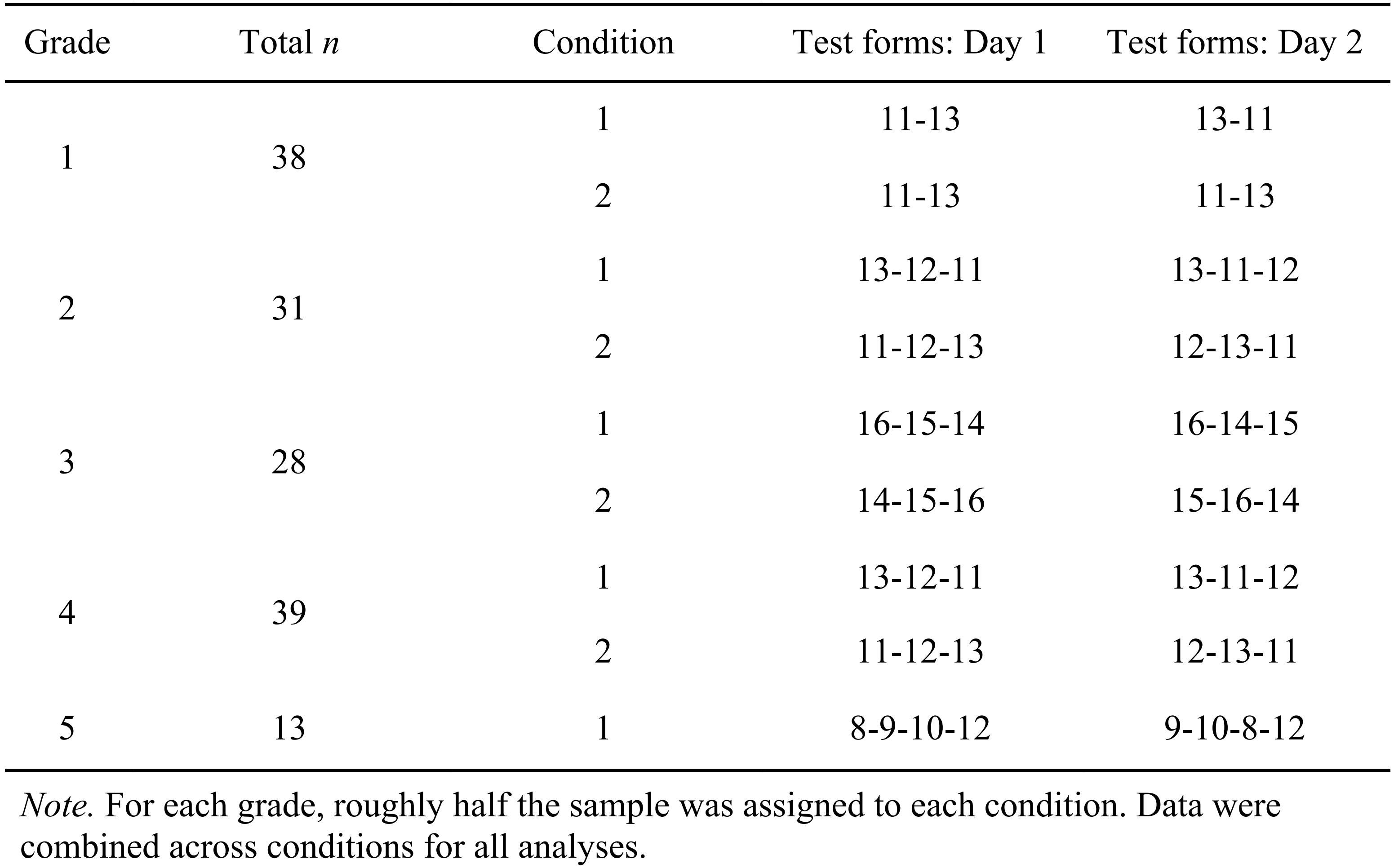
Measures

- **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



Grade	Total *n*	Condition	Test forms: Day 1	Test forms: Day 2
1	38	1	11-13	13-11
		2	11-13	11-13
2	31	1	13-12-11	13-11-12
		2	11-12-13	12-13-11
3	28	1	16-15-14	16-14-15
		2	14-15-16	15-16-14
4	39	1	13-12-11	13-11-12
		2	11-12-13	12-13-11
5	13	1	8-9-10-12	9-10-8-12

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.

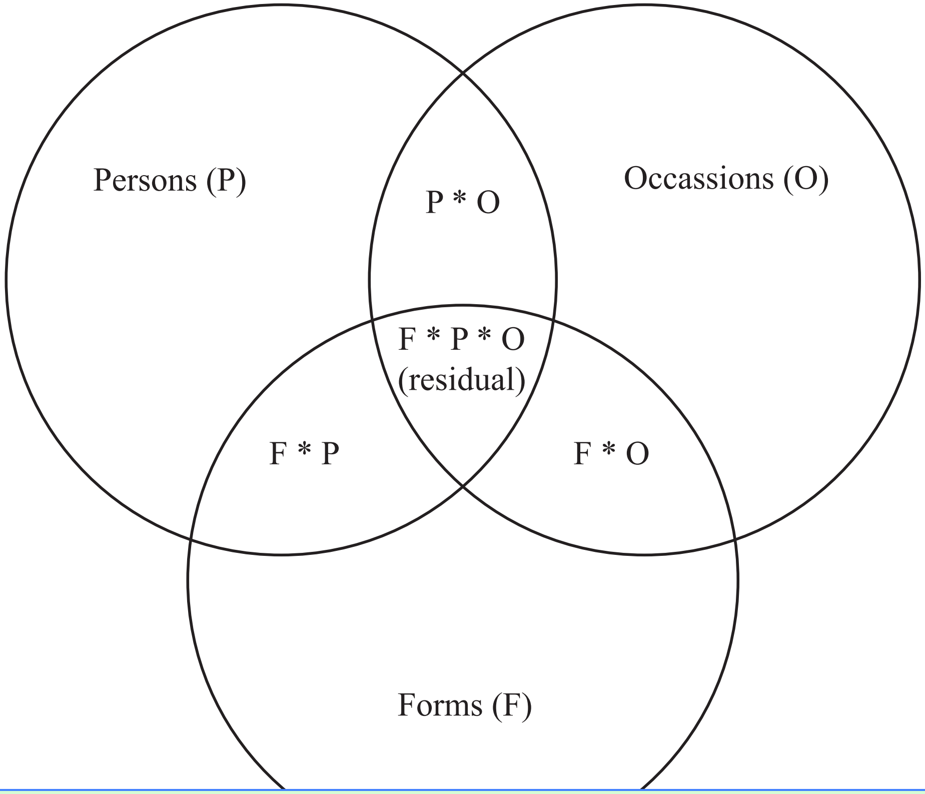
Potential Facets in ORF

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

Fully crossed, two-facet design used in this study. One analysis presented for each of grades 1-5.

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

G-Study: Variance Components

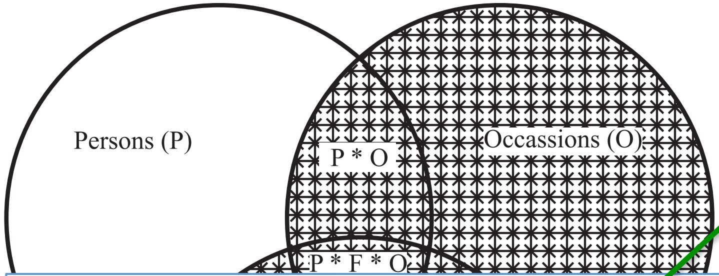


Because the design was fully crossed, the G-Study portion estimated the variance associated with persons and each facet, as well as all interactions.

Study Variance Components

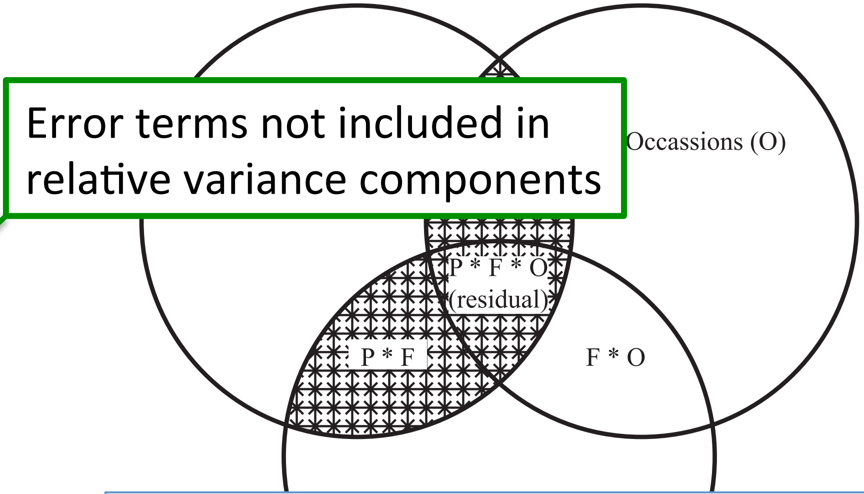
Note that we will use **absolute** rather than **relative** error variances

Absolute Error Variance



$$\sigma^2_{Absolute} = \boxed{\frac{\sigma^2_F}{n'_F} + \frac{\sigma^2_O}{n'_O} + \frac{\sigma^2_{FO}}{n'_F n'_O}} + \frac{\sigma^2_{PF}}{n'_F} + \frac{\sigma^2_{PO}}{n'_O} + \frac{\sigma^2_{PFO,e}}{n'_F n'_O}$$

Relative Error Variance



Error terms not included in relative variance components

$$\sigma^2_{Relative} = \frac{\sigma^2_{PF}}{n'_F} + \frac{\sigma^2_{PO}}{n'_O} + \frac{\sigma^2_{PFO,e}}{n'_F n'_O}$$

D-Study

- How might the error variance change with **different levels** of each facet?
- Estimates obtained in a similar fashion to the Spearman-Brown P

Substitute in the level for the particular facet you are interested in.

$$\sigma^2_{Absolute} = \frac{\sigma^2_F}{n'_F} + \frac{\sigma^2_O}{n'_O} + \frac{\sigma^2_{FO}}{n'_F n'_O} + \frac{\sigma^2_{PF}}{n'_F} + \frac{\sigma^2_{PO}}{n'_O} + \frac{\sigma^2_{PFO,e}}{n'_F n'_O}$$

Results: G-Study

Variance Components for G-Theory Analyses

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

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

Overall, [illegible] measurement process was associated [illegible] measurement)

Also note how the variance associated with persons generally decreases as grade level increases

Very little variance [illegible] the person by [illegible]

With the exception of grade 3.

Slightly more [illegible] with occasion [illegible] modest

Generally more variance associated with person by form interactions than with forms individually

behavioral research & teaching

Results: D-Study
Absolute Standard Errors

Predicted absolute standard errors by administration practice

			D studies									
Reliability index	Grade	*n* Occasions	1	1	1	1	1	2	2	2	2	2
		n Forms	1	2	3	4	5	1	2	3	4	5
	1	-	10.32	8.26	7.45	7.01	6.74	8.68	6.72	5.93	5.49	5.20
	2	-	13.30	10.55	9.46	8.86	8.48	10.22	7.98	7.07	6.58	6.26
SE, $\sigma(\Delta_p)$	3	-	16.23	13.85	12.96	12.50	12.21	13.08	10.75	9.86	9.38	9.08
	4	-	13.85	11.67	10.85	10.42	10.15	10.56	8.71	8.00	7.63	7.39
	5	-	12.80	9.80	8.57	7.89	7.45	9.76	7.40	6.42	5.87	5.52

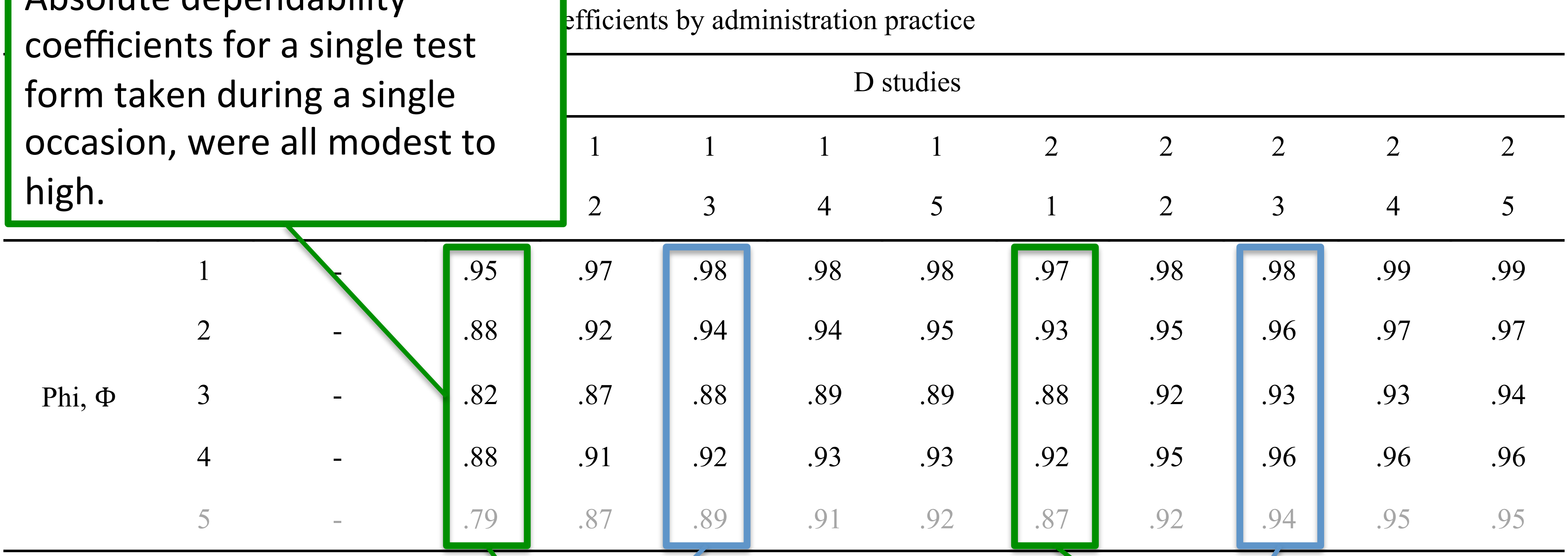
Notice that th
3-4 wcpm by
instead of 1.

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

r reduction was observed
he second testing occasion.

Results: D-Study
Absolute Dependability Coefficients

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



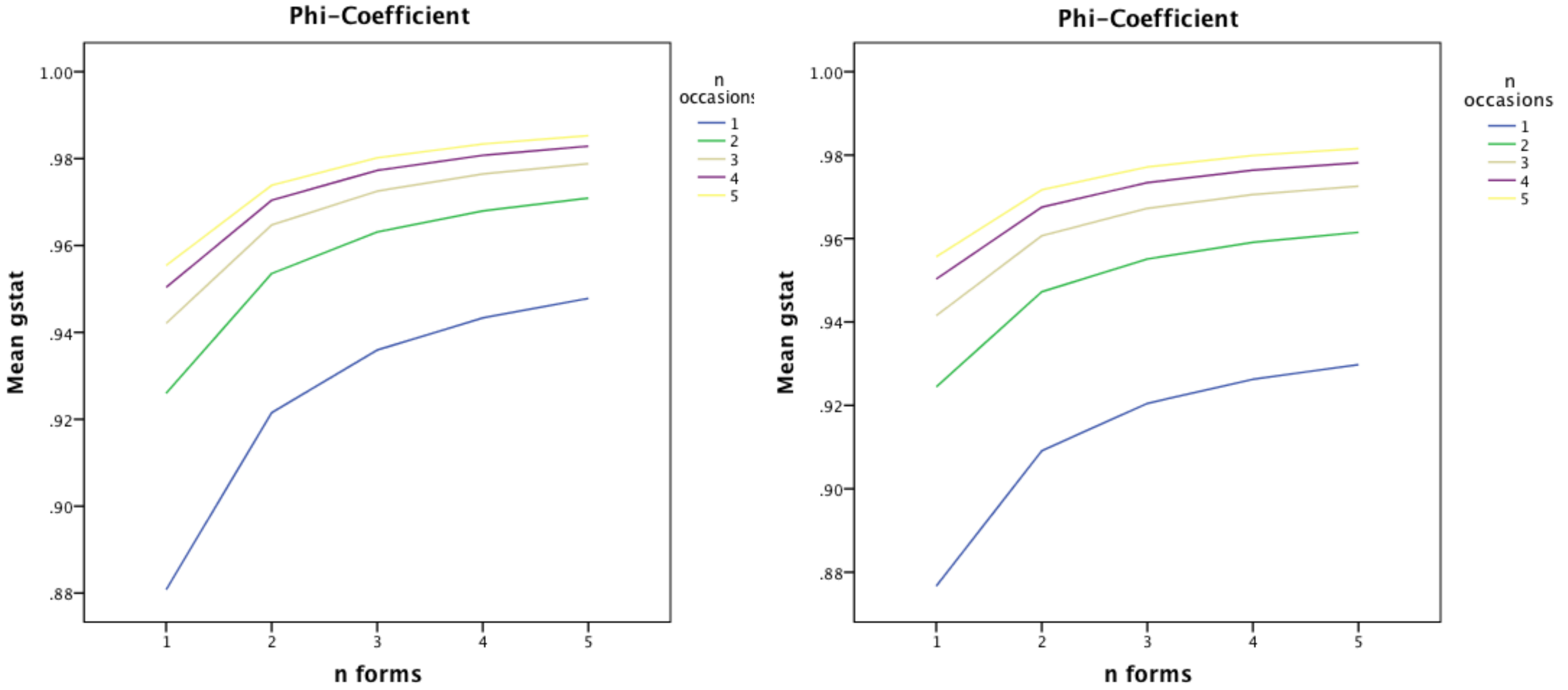
efficients by administration practice

				D studies								
				1	1	1	1	2	2	2	2	2
				2	3	4	5	1	2	3	4	5
	1	-	.95	.97	.98	.98	.98	.97	.98	.98	.99	.99
	2	-	.88	.92	.94	.94	.95	.93	.95	.96	.97	.97
Phi, Φ	3	-	.82	.87	.88	.89	.89	.88	.92	.93	.93	.94
	4	-	.88	.91	.92	.93	.93	.92	.95	.96	.96	.96
	5	-	.79	.87	.89	.91	.92	.87	.92	.94	.95	.95

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

A similar increase was observed during the second testing occasion.

Dependability Coefficients



Grade 2 Grade 4

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 occassions.

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

Median Score Approach

- **3 minutes** of testing per student (not including transition times)
- = **30,000 minutes** of testing (~500 hours) **district-wide** for a single administration
- * 3 testing occasions (fall, winter, and spring) = **90,000 minutes** (~1,500 hours) **district-wide**

Single Passage Approach

- **1 minute** of testing per student (not including transition times)
- = **10,000 minutes** of testing (~167 hours) **district-wide** for a single administration
- * 3 testing occasions (fall, winter, and spring) = **30,000 minutes** (~500 hours) **district-wide**

Resource Allocation

- Median score approaches will nearly always have **higher reliability**, and **lower standard**

  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?

BRT
behavioral research & teaching

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

BRT
behavioral research & teaching

Inter-Construct Validity

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

Essential components of effective reading instruction (National Reading Panel)

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**

BRT
behavioral research & teaching

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.)

Conclusions

Median Score

- Increased reliability
- Lower standard errors
- Better for high-stakes decisions

Single Probe

- Increased efficiency
- Reliability and standard errors still generally within acceptable range
- Better for benchmarking?

References

Alonzo, J., & Tindal, G. (2007). *Examining the technical adequacy of word reading and passage reading fluency measures in a progress monitoring assessment system* (Technical Report # 40). Eugene, OR: Behavioral Research and Teaching.

Christ, T. J., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review, 36*, 130-146.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*, 315-342. doi: 10.1016/j.jsp.2007.06.003

Jamgochian, E.M., Park, B.J., Nese, J.F.T., Lai, C.F., Sáez, L., Anderson, D., Alonzo, J., & Tindal, G. (2010) *Technical adequacy of the easyCBM grade 2 reading measures, 2009-2010 version.* (Technical Report #1004). Eugene, OR: Behavioral Research and Teaching.

Lai, C.F., Nese, J.F.T., Jamgochian, E.M., Kamata, A., Anderson, D., Park, B.J., Alonzo, J., & Tindal, G. (2010). *Technical adequacy of the easyCBM primary-level reading measures (Grades K-1), 2009-2010 version.* (Technical Report #1003). Eugene, OR: Behavioral Research and Teaching.

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurment of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment, 23*, 326-338. doi: 10.1177/073428290502300403

Sáez, L , Park, B.J., Nese, J.F.T, Jamgochian, E.M., Lai, C.F., Anderson, D., Alonzo, J., & Tindal, G. (2010) *Technical adequacy of the easyCBM reading measures (Grades 3-8), 2009-2010*

BRT
behavioral research & teaching

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:

> Reliability and Validity Evidence for Progress Measures in Reading. U.S. Department of Education, Institute for Education Sciences. R324A100014. June 2010 - June 2014.

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