Technical Advisory Committee Meeting for DARA Project

Cara Cahalan-Laitusis
Educational Testing Service

ETS

ies NATIONAL CENTER FOR SPECIAL EDUCATION RESEARCH
Institute of Education Sciences

NARAP Projects Goals

1. Develop a definition of reading proficiency
2. Research the assessment of reading proficiency
3. Develop research-based principles and guidelines making large-scale reading assessments more accessible for students who have disabilities that affect reading
4. Develop and field trial a prototype reading assessment

National Accessible Reading Assessment Projects

- Designing Accessible Reading Assessments (DARA)
- Partnership for Accessible Reading Assessment (PARA)
- Technology Assisted Reading Assessment (TARA)

ETS®
ies NATIONAL CENTER FOR SPECIAL EDUCATION RESEARCH
Institute of Education Sciences

Partnership for Accessible Reading Assessments (PARA)

- Collaboration between the National Center for Educational Outcomes, CRESST, and Westat
- Focus on all disabilities that impact reading, particularly:
  - learning disabilities,
  - speech/language impairments,
  - mental retardation, and
  - deafness/hard of hearing
- Investigate varied obstacles to accessible reading assessments and identify possible solutions

Designing Accessible Reading Assessments (DARA)

- Educational Testing Service (ETS)
- Focuses on students with learning disabilities
- Focuses on component approach to assessing reading skills. Primary focus are:
  - Word Recognition
  - Reading Fluency
  - Vocabulary Knowledge
  - Comprehension

Technology Assisted Reading Assessment (TARA)

- ETS, NCEO and Center for Applied Special Technology (CAST)
- Focus on students with visual impairments
- Focus on:
  - Examining the performance of operational ELA tests for students with visual impairments
  - Development of prototype Technology Assisted Reading Assessment
  - Inclusion of VI students in NARAP field test

Collaborative Dissemination

- 2006 Presentations
  - AERA
  - NCME
  - CCSSO LSAC
  - ATP
  - CEC
  - LDA

- 2007 Submissions
  - Awaiting response from AERA/NCME/CCSSO/CEC
  - ATP (accepted)
  - IRA (accepted)
  - ASCD (accepted presenting March 18th in Anaheim, CA)

Progress for Goal 1: Definition

- Reading First Definition was adopted by NARAP
- Two reports were written which are available on the NARAP website www.narap.info
  - – Focus Group Results
  - – Issues and Principles Paper

Primary Questions for Year 2

- Can comprehension be assessed in audio format if word recognition and fluency are assessed separately?
  - Are listening comprehension and reading comprehension similar constructs (highly correlated) in proficient readers?
  - Do students with reading-based learning disabilities receive differential performance gains from read aloud?
  - Do tests and test items taken with and without read aloud perform the same psychometrically (same factor structure, no evidence of differential item performance)?

Year 2 Research

- Differential Boost from Read Aloud on Reading Test
- Psychometric Studies of ELA test
  - Differential Item Function
  - Differential Distractor Analysis
  - Factor Analysis

Year 3 Research

- Continue psychometric research on GMRT
- Continue analysis of differential boost data
- Think Aloud studies with LD and non-LD students to examine how students approach
  - items shown to have DIF
  - new item types designed to assess fluency and word recognition in a large scale assessment
  - Families of items with slight variations (e.g., operational item and universally designed items)

Focus of this meeting

- Review of Research Results from Year 2
  - Psychometric Research
  - Differential Boost
- Feedback on Research Plans for Year 3
  - Psychometric Research
  - Additional analysis of differential boost
  - Cognitive Labs
- Feedback on Field Test Plans for Goal 4

Using Factor Analysis to Investigate the Impact of Accommodations on the Scores of Students with Disabilities on English-Language Arts Assessments

Linda Cook
DARA Technical Advisory Committee
October 5-6, 2006

Factor Analysis of STAR ELA Assessment

- Report sent as background information for meeting discusses item level analyses
- Today’s discussion will focus on factor analyses carried out using item parcels as input

Factor Analyses of STAR ELA Assessment

- Item level factor analyses
  - Exploratory and Confirmatory Analyses
    - Grade 4 and Grade 8
    - Students without disabilities
    - Students with disabilities (no accommodations)
    - Students with disabilities (504/IEP accommodations)

Factor Analyses of STAR ELA Assessments

- Item level analyses
- Common factor exploratory analysis
- Confirmatory analyses
  - Base-line model
  - Multi-group analyses
- Grade 4 and grade 8 tests measuring single dimensions for all three groups
- Not able to test hypotheses of equal factor loadings or equal intercorrelations

Factor Analyses of Item Parcel Data for the STAR ELA Assessments

- **Purpose of the Study**
  - **First purpose**
    - **To determine whether the ELA assessments measured the same constructs for**
    - **Students without disabilities**
    - **Students with disabilities who took the test without accommodations**
    - **Students with disabilities who took the test with accommodations (504/IEP)**
    - **Students with disabilities who took the test with accommodations (Read aloud)**
  - **Second purpose**
    - **To demonstrate that the matching criterion for DIF study (total test score) was unidimensional**

Number of Items for Grade 4 English-Language Arts Assessment

Test	Content	No. of Items
Reading	Word Analysis, Fluency, and Systematic Vocabulary Development	18
	Reading Comprehension	15
	Literary Response and Analysis	9
	Total—Reading	42
Writing	Writing Strategies	15
	Writing Applications (Genres and Their Characteristics)	1*
	Written and Oral English Language Conventions	18
	Total—Writing	34

*Essay item (all others are multiple-choice). The essay item was not used in the study

Number of Items for Grade 8 English-Language Arts Assessment

Test	Content	No. of Items
Reading	Word Analysis, Fluency, and Systematic Vocabulary Development	9
	Reading Comprehension	18
	Literary Response and Analysis	15
	Total—Reading	42
Writing	Writing Strategies	17
	Written and Oral English Language Conventions	16
	Total—Writing	33

STAR ELA Grade 4 and Grade 8 Summary Statistics

Grades 4 and 8 Total Group Sizes, Sample Sizes and Summary Statistics for English-Language Arts Assessment

	Grade 4 Total Groups			Grade 4 Samples (N = 500)		Grade 8 Total Groups			Grade 8 Samples (N = 500)	
Group	N	Mean	SD	Mean	SD	N	Mean	SD	Mean	SD
(1) Students without disabilities	298,622	48	14	47	14	357,374	46	12	46	12
(2) LD, without accommodations	9,045	29	12	29	12	18,512	29	10	29	10
(3) LD, 504/IEP accommodations	4,724	27	10	27	10	4,325	27	9	27	9
(4) LD, read-aloud accommodation	1,367	29	11	29	11	874	27	9	27	9

Factor Analyses of STAR ELA Assessments

- Underlying structure of the tests
  - 5 factors based on strands
  - Two factors based on reading and writing items
  - Single ELA factor
- Began exploratory analyses using item parcels constructed within strands

Summary of Item Parcel Information for Grade 4 Factor Analysis of English-Language Arts Assessment

				Average P+			
Parcel No.	Content	Strand	No. Items	Students Without Disabilities	LD, Without Accom.	LD, With 504/IEP Accom.	LD, With Read Aloud Accom.
1	Reading	1	5	.708	.402	.408	.412
2	Reading	1	4	.700	.460	.422	.475
3	Reading	1	5	.704	.443	.413	.474
4	Reading	1	4	.710	.401	.406	.489
5	Reading	2	5	.618	.399	.357	.414
6	Reading	2	5	.609	.346	.322	.344
7	Reading	2	5	.610	.348	.344	.363
8	Reading	3	5	.549	.364	.338	.342
9	Reading	3	4	.551	.357	.328	.349
10	Writing	4	5	.645	.396	.368	.372
11	Writing	4	5	.623	.398	.389	.402
12	Writing	4	4	.652	.354	.324	.358
13	Writing	4	4	.645	.404	.395	.402
14	Writing	5	5	.587	.372	.344	.360
15	Writing	5	5	.596	.375	.355	.349
16	Writing	5	5	.589	.338	.315	.343

Summary of Item Parcel Information for Grade 8 Factor Analysis of English-Language Arts Assessment

				Average P+			
Parcel No.	Content	Strand	No. Items	Students Without Disabilities	LD, Without Accom.	LD, With 504/IEP Accom.	LD, With Read Aloud Accom.
1	Reading	1	5	.655	.423	.407	.363
2	Reading	1	4	.653	.454	.431	.410
3	Reading	2	5	.608	.396	.360	.342
4	Reading	2	4	.614	.414	.370	.401
5	Reading	2	5	.610	.362	.353	.351
6	Reading	2	4	.597	.404	.359	.325
7	Reading	3	5	.592	.392	.355	.358
8	Reading	3	5	.608	.374	.352	.376
9	Reading	3	5	.606	.376	.341	.344
10	Writing	4	4	.562	.323	.290	.297
11	Writing	4	4	.588	.362	.314	.321
12	Writing	4	4	.598	.390	.391	.326
13	Writing	4	4	.578	.400	.383	.379
14	Writing	5	4	.651	.395	.368	.357
15	Writing	5	4	.660	.385	.346	.374
16	Writing	5	4	.663	.361	.322	.315
17	Writing	5	5	.662	.440	.441	.400

Factor Analyses of STAR ELA Assessments

- Exploratory analyses (separately in each group)
  - how many factors
- Confirmatory (separate analyses of groups ;multi-group analysis)
  - Establish base-line model
  - Determine number of factors needed to describe data across four groups

Factor Analyses of STAR ELA Assessments

Summary of Results of Grade 4 Multi-Group Confirmatory Factor Analyses of English-language Arts Assessment

Analysis	df	Normal Theory Chi-square	RMSEA	CFI	GFI
2-factor model (baseline)	412	532.137	.012	.987	.969
2-factor model, all loadings constrained as equal (model 1)	460	698.284	.016	.974	.958
2-factor model, all loadings constrained as equal and interfactor correlations constrained as 1 (model 2)	464	786.892	.019	.965	.952

Factor Analyses of STAR ELA Assessments

Summary of Results of Grade 8 Multi-Group Confirmatory Factor Analyses of English-language Arts Assessment

Analysis	df	Normal Theory Chi-square	RMSEA	CFI	GFI
2-factor model (baseline)	472	511.404	.006	.994	.971
2-factor model, all loadings constrained as equal (model 1)	523	741.587	.014	.965	.957
2-factor model, all loadings constrained as equal and interfactor correlations constrained as 1 (model 2)	527	843.157	.017	.949	.951

Summary

- Factor analyses indicate one factor needed to account for data
- Provides some evidence of validity of test for students with disabilities
- Additional validity evidence needs to be collected

Questions?

Comments?

Using Differential Item Functioning to Investigate the Impact of Accommodations on the Scores of Students with Disabilities on English-Language Arts Assessments

Mary J. Pitoniak
Educational Testing Service

Purpose and Overview of the Study

- The purpose of this study was to examine differential item functioning on the English-Language Arts assessments at grades 4 and 8 described by Linda
- DIF analyses are statistical procedures that are used to identify items that function differently for different subgroups of examinees
- DIF "exists when examinees of equal ability differ, on average, according to their group membership in their responses to a particular item" *(Standards)*

ETS
ies NATIONAL CENTER FOR SPECIAL EDUCATION RESEARCH
Institute of Education Sciences

Purpose and Overview of the Study (continued)

- Issues investigated:
  - How many items are flagged as showing DIF?
  - Are the results interpretable in terms of a priori or a posteriori evaluation of item content?
  - Of particular interest:
    When the read-aloud accommodation is used, do the items function differentially for students?

Purpose and Overview of the Study (continued)

- Features of study:
  - Used Mantel-Haenszel DIF method, with purification step as recommended by literature
  - Large enough sample sizes (which is not always the case)
  - A priori codings of characteristics made, along with prediction of effect of read-aloud accommodation on difficulty of item
  - A posteriori interpretations of flagged items

Method

- Mantel-Haenszel Categorization—3 Levels
  - A ➔ Negligible DIF
  - B ➔ Slight to Moderate DIF
  - C ➔ Moderate to Large DIF

    (At ETS, operational items categorized as *C* are carefully reviewed to determine whether there is a plausible reason why any aspect of that question may be unfairly related to group membership, and may or may not be retained on the test.)

Method (continued)

- Directions of DIF Flags

  - ➔ Favors reference group

  + ➔ Favors focal group

- The table on the following page shows the reference and focal groups for each comparison

Comparisons Made in the Study

Comparison Number	Reference Group	Focal Group
1.3	Without disabilities	LD no accommodations
1.4	“	LD IEP/504 accommodations
1.5	“	LD read-aloud accommodation (& IEP/504 accommodations)
3.1	LD no accommodations	LD IEP/504 accommodations
3.2	“	LD read-aloud accommodation (& IEP/504 accommodations)

Results

- Level *C (moderate to large DIF)*
  - 1 item flagged at each grade
  - Grade 4: Reading,
    Grade 8: Writing
  - Both flagged as favoring the <u>reference</u> group of students without disabilities, with the focal group being students with disabilities who received the read-aloud accommodation (comparison 1.5)

Results (continued)

- Level *B (slight to moderate DIF)*
  - 9 items flagged at 4th grade,
    8 items flagged at 8th grade
  - Majority of flagged items were Reading items
  - Many items favored the focal group of students with disabilities who received the read-aloud accommodation over the reference group of students without disabilities (comparison 1.5)

Grade 4

Ref. Group	Non-LD			LD No Acc.		Total Number of Flags
Focal Group	LD no acc (1.3)	LD IEP/504 (1.4)	LD read-aloud (1.5)	LD IEP/504 (3.1)	LD read-aloud (3.2)	
Item						
3 (R)	*B-*	*B-*				2
10 (R)			*B+*		*B+*	2
13 (R)			*B+*			1
25 (R)			*B+*			1
32 (R)			*B-*			1
33 (R)			*B+*			1
34 (R)			*B+*			1
45 (W)			*B-*			1
64 (W)		*B-*	*C-*		*B-*	3

Grade 8

Ref. Group	Non-LD			LD No Acc.		Total Number of Flags
Focal Group	LD no acc (1.3)	LD IEP/504 (1.4)	LD read-aloud (1.5)	LD IEP/504 (3.1)	LD read-aloud (3.2)	
Item						
1 (R)			*C-*			1
2 (R)			*B-*			1
15 (R)			*B+*			1
20 (R)			*B-*			1
28 (R)			*B+*		*B+*	2
29 (R)			*B+*			1
42 (R)			*B+*		*B+*	2
71 (W)			*B+*			1

A Priori Theories About Read-Aloud Accommodation

- How accurate were the predictions about whether a read-aloud accommodation would make an item easier or more difficult?

Grade 4

	Impact of Read-Aloud Accommodation on Item Difficulty	
Item	Prediction	Result
10 (R)	Difficult	Easier
13 (R)	Easier	Easier
25 (R)	Easier	Easier
32 (R)	Easier	Difficult
33 (R)	Difficult	Easier
34 (R)	Difficult	Easier
45 (W)	Difficult	Difficult
64 (W)	Difficult	Difficult

Red shading indicates an inaccurate prediction

Grade 8

	Impact of Read-Aloud Accommodation on Item Difficulty	
Item	Prediction	Result
1 (R)	Easier	Difficult
2 (R)	Easier	Difficult
15 (R)	Easier	Easier
20 (R)	Difficult	Difficult
28 (R)	Difficult	Easier
29 (R)	Difficult	Easier
42 (R)	Easier	Easier
71 (W)	Difficult	Easier

Red shading indicates an inaccurate prediction

A Posteriori Interpretation About Read-Aloud Accommodation Results

- The reasons why some of the items were easier with read-aloud accommodation were not obvious to test developers.

Differential Distractor Functioning Analyses

- Differential Distractor Functioning (DDF) Analyses were also carried out (see paper by Middleton & Cahalan-Laitusis, 2006)
- These analyses yielded information about which distractors functioned differently for the reference and focal groups

What the Results of the DIF Study Say About the 3 Questions Posed

1. How many items were flagged for DIF?
   - Only 1 at each grade at Level C (moderate to large DIF), which is in line with other comparisons for this test (gender, race/ethnicity)
   - 8 or 9 items for each grade at Level B (slight to moderate DIF), which is slightly above normal rates for other comparisons for this test

What the Results of the DIF Study Say About the 3 Questions Posed (continued)

2. Are the results interpretable in terms of a priori or a posteriori evaluation of item content?
   - No, for the most part

3. Of particular interest: When the read-aloud modification is used, do the items function differentially for students?
   - Some items were easier when read-aloud, though at the B level; which somewhat supports this state’s decision to view read-aloud as a modification

Next Steps

- ELL and ELL/LD groups to be compared
- Additional DIF method to be utilized

Questions?
Comments?

Results from Differential Boost Study

Cara Cahalan-Laitusis
Educational Testing Service

Differential Boost from Read Aloud (Non-disabled vs. RLD)

1. Is there a Differential Boost from read aloud?
2. How well do test scores (standard, audio, and fluency) predict variance in teacher ratings of reading comprehension?
3. Are teachers' able to predict which students will benefit from read aloud?

Prior Research

- No Differential Boost
  - Kosciokek & Ysseldyke (2000)- Small sample size (n=31)
  - Meloy, Deville, and Frisbie (2002) – Between subjects design (n=260, 76% non-disabled, randomly assigned to audio or standard)
  - McKevitt & Elliott (2003)-Small sample size (n=39)
- Differential Boost
  - Crawford and Tindal (2004)-(n=338, 78% non-disabled)
  - Fletcher, et. al (2006)-Between subjects design (randomly assigned to audio or standard). Sample included 91 Dyslexic (poor decoder) and 91 average decoders

Data Collected

- GMRT Forms S and T
  - Extra Time
  - Extra Time with Read Aloud via CD
- 2 Fluency Measures
  - WJ Reading Fluency
  - Test of Silent Word Reading Fluency
- 2 Decoding Measures (4th grade only)
  - WJ Letter Word ID
  - WJ Word Recognition
- Demographic and Survey Data

Sample

- 1170 4th Graders
  - 522 Students with RLD
  - 648 Students without a disability

- 855 8th Graders
  - 394 Students with RLD
  - 461 Students without a disability

Design

Group	Session 1		Session 2	
	Form	Accommodation	Form	Accommodation
1	S	Standard	T	Audio
2	S	Audio	T	Standard
3	T	Standard	S	Audio
4	T	Audio	S	Standard

Means for Grade 4

	Non-LD			RLD		
Test/Condition	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
WJ Letter Word ID	604	**504**	21	469	**473**	29
WJ Word Attack	604	**504**	15	469	**484**	20
TOSWRF	604	**102**	10	469	**89**	12
WJ Fluency	604	**501**	24	469	**474**	21
Audio	604	**502**	32	469	**477**	30
Standard	604	**497**	37	469	**457**	31
Boost	604	**5**	24	469	**19**	27

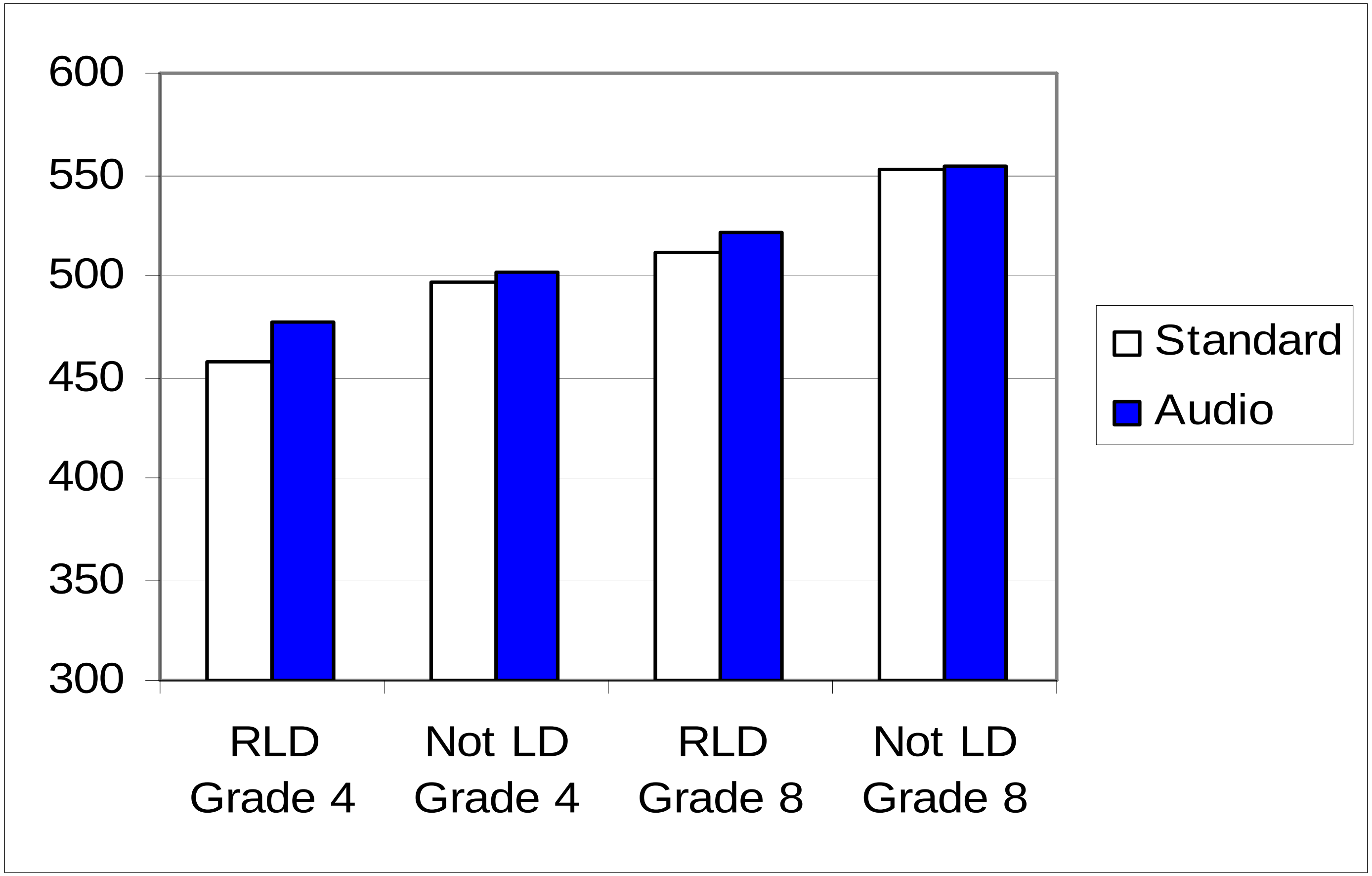
ETS®
ies NATIONAL CENTER FOR SPECIAL EDUCATION RESEARCH
Institute of Education Sciences

Means for Grade 8

	Non-LD			RLD		
Test/Condition	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
TOSWRF	463	**103**	13	373	**90**	12
WJ Fluency	463	**560**	42	373	**514**	34
Audio	463	**555**	31	373	**521**	27
Standard	463	**553**	3	373	**511**	28
Boost	463	**2**	21	373	**10**	23

1. Is there a Differential Boost from read aloud?

Scores by RLD and Grade



Repeated Measures ANOVA

- Dependent Variables:
  - GMRT “Standard”
  - GMRT Audio
- Independent Variables:
  - Disability Status (RLD vs. NLD)
  - Form/Order (STSA, STAS, TSSA, TSAS)
- Covariate: WJ-III Reading Fluency

RM ANOVA for Grade 4

Repeated Measures Analysis of Variance for Grade 4

Source	df	F	p
Within subjects			
Boost	1	265.81***	.000
Boost x Reading LD	1	96.46***	.000
Boost x Form/Order	3	0.62	.602
Boost x Reading LD x Form/Order	3	1.35	.258
Error(Boost)	1,173	(342.85)	

Note. Value enclosed in parentheses represent mean square errors.
*p < .05. **p < .01, *** p <.001.

RM ANOVA for Grade 4 with Fluency Covariate

Table 8. Repeated Measures Analysis of Variance for Grade 4 with Fluency

Source	df	F	p
Within subjects			
Boost	1	71.43***	.000
Fluency (Covariate)	1	58.87***	.000
Boost x Reading LD	1	22.50***	.000
Boost x Form/Order	3	0.91	.438
Boost x Reading LD x Form/Order	3	1.50	.213
Error(Boost)	1,171	(323.03)	

Note. Value enclosed in parentheses represent mean square errors, *p < .05. **p < .01, *** p <.001

Differential Boost Findings

- Differential Boost at both 4th and 8th grades (i.e., students with LD had significantly greater score gains from read aloud than non-LD students)
- When WJ reading fluency ability is controlled for a Differential Boost is still found at both grades

2. How well do test scores (standard, audio, and fluency) predict variance in teacher ratings of reading comprehension?

Summary of Regression Analysis for Variables Predicting Reading Comprehension for 4th grade students with Reading Learning Disabilities

Variable	*B*	*SE B*	β
Step 1			
Standard reading comprehension	0.01	0.00	.46**
Step 2			
Standard reading comprehension	0.01	0.00	.24**
Reading fluency	0.01	0.00	.38**
Step 3			
Standard reading comprehension	0.00	0.00	.17**
Reading fluency	0.01	0.00	.38**
Audio reading comprehension	0.00	0.00	.13**

Note. R^2 = .21 for Step 1; ΔR^2 = .10 for Step 2; ΔR^2 = .01 for Step 3 (ps < .01). ***p < .05. <.001, **p < <.01.

Regression Findings

- Audio score does not significantly predict variance in Teacher Ratings of Reading Comprehension (beyond standard and fluency) for Grade 8 RLD
- Audio score adds to prediction of reading comprehension (beyond standard and fluency scores) for three groups (NLD grade 4, NLD grade 8, and RLD grade 4), but incremental change is small

3. Are teachers’ able to predict which students will benefit from read aloud?

Accuracy of Teacher Prediction

For this study each student took a reading comprehension test that was read aloud by a CD player and another reading comprehension test that they read to themselves. Which test do you predict the student did better on?

Ⓐ Test read aloud by CD player

Ⓑ Test the student read to themselves

Ⓒ No difference

Findings from Teacher Predictions

- On average teachers were able to predict score gain from audio at grade 4 but not grade 8
- At the individual level teachers accurately predicted if a student would benefit from the audio version about 35% of the time and were completely wrong about 5% of the time

Teachers Predictions

	Grade 4					
	RLD (n=519)			NLD (n=639)		
Teacher Prediction	*M*	*N*	*SD*	*M*	*N*	*SD*
Audio	21.6	411	29.7	8.5	292	23.3
Standard	5.4	43	27.6	0.2	128	24.5
No Difference	21.4	65	23.3	3.2	219	23.5

Teacher Predictions

	Grade 8					
	RLD (n=363)			NLD (n=433)		
Audio	11.4	254	23.1	2.6	162	20.8
Standard	4.7	49	19.7	1.4	120	20.6
No Difference	6.8	60	23.3	1.7	151	21.3

	Actual Performance					
	Audio	Same	Standard	Audio	Same	Standard
	Grade 4					
Teacher Prediction	RLD (N=519)			NLD (N=639)		
Audio	30%	47%	**3%**	9%	34%	**2%**
No Difference	5%	8%	0%	5%	26%	4%
Standard	**2%**	6%	1%	**2%**	16%	2%

Questions/Comments

Plans for Factor Analysis of GMRT Data

Linda Cook
DARA Technical Advisory Committee
October 5-6, 2006

Outline of Presentation

- Why we want to factor analyze the data from the differential boost study
- Advantages of using data from the differential boost study
- Analyses we plan to carry out
- How we plan to analyze the data
- Questions that we have about our data analysis plan

Why Factor Analyze Data From the Differential Boost Study

- Understand implications of using total test score as criterion in DIF studies
- Aid in interpretation of results of differential boost study
- Increase understanding of impact of disability and accommodation on reading test scores

Why Factor Analyze Data From the Differential Boost Study

- Understand implications of using total test score as criterion in DIF studies

Why Factor Analyze Data From the Differential Boost Study

- Aid in interpretation of results of differential boost study

Why Factor Analyze Data From the Differential Boost Study

- Increase understanding of impact of disability and accommodation on reading test scores

Advantages of Using Differential Boost Data

- Characteristics of the Samples
- Specification of the Accommodations

Advantages of Using Differential Boost Data

- Characteristics of Samples

Advantages of Using Differential Boost Data

- Specification of Accommodations

Possible Factor Analyses

- Comparisons of factor structures for four groups
  - Reading based learning disability, no accommodation
  - Reading based learning disability, audio accommodation
  - No disability, no accommodation
  - No disability, audio accommodation

Possible Factor Analyses

Summary of Possible Comparisons of Factor Structures

Comparison	Group 1	Group 2
1	RLD Standard	RLD Audio
2	RLD Standard	NLD Audio
3	RLD Standard	NLD Standard
4	RLD Audio	NLD Audio
5	RLD Audio	NLD Standard
6	NLD Standard	NLD Audio

Analyses We Plan to Carry Out

- What are the implications of using total test score as a criterion in DIF studies of accommodated and non-accommodated scores?
  - Compare factor structure for students without disabilities taking test without accommodation with factor structure for students with a disability taking the test with an accommodation

Analyses We Plan to Carry Out

- Aid in interpretation of results of differential boost study
  - Compare factor structures for students without disabilities who took test with and without accommodation
  - Compare factor structures for students with disabilities who took test with and without accommodation

Analyses We Plan to Carry Out

- Increase understanding of impact of disability and accommodation on reading test scores
  - Compare factor structures of test given to examinees with and without disabilities under standard conditions
  - Compare factor structure of test given to examinees with disabilities who take test with accommodations and examinees without disabilities who take test without accommodations

How We Plan to Analyze the Data

- Item parcels
- Develop base-line model
- Test factor structure
- Test measurement model

Questions About Our Analysis Plans

- Should we start with item level or parcel level data?
- Do the comparisons that we described earlier make sense?
- Are there other comparisons that we should be examining?

Questions About Our Analysis Plans

- Should we start with item level or parcel level data?

Questions About Our Analysis Plans

- Are the comparisons we have specified for each of the questions the most appropriate?

Questions About Our Analysis Plans

- Are there other questions or comparisons that we should be considering?

Additional Questions or Comments?

Plans for DIF and DDF Analyses of GMRT Data

Mary J. Pitoniak
Educational Testing Service

Purpose of Doing DIF and DDF Analyses on Data From the Differential Boost Study

- Aid in interpretation of results of differential boost study
- Increase understanding of impact of disability and accommodation on reading test scores

Possible DIF and DDF Analyses

- Comparisons of scores for 4 groups
  - Reading based learning disability (RLD), no accommodation (Standard)
  - Reading based learning disability (RLD), audio accommodation (Audio)
  - No disability (NLD), no accommodation (Standard)
  - No disability (NLD), audio accommodation (Audio)

Possible Comparisons

	Focal Group			
Reference Group	RLD Standard	RLD Audio	NLD Standard	NLD Audio
RLD Standard	--	1	3	6
RLD Audio		--	5	4
NLD Standard			--	2
NLD Audio				--

Comparisons 1 and 2:
Same Group, Different Mode

	Focal Group			
Reference Group	RLD Standard	RLD Audio	NLD Standard	NLD Audio
RLD Standard	--	1	3	6
RLD Audio		--	5	4
NLD Standard			--	2
NLD Audio				--

Comparisons 3 and 4:
Same Mode, Different Groups

	Focal Group			
Reference Group	RLD Standard	RLD Audio	NLD Standard	NLD Audio
RLD Standard	--	1	3	6
RLD Audio		--	5	4
NLD Standard			--	2
NLD Audio				--

Comparisons 5 and 6: Different Mode, Different Group

	Focal Group			
Reference Group	RLD Standard	RLD Audio	NLD Standard	NLD Audio
RLD Standard	--	1	3	6
RLD Audio		--	5	4
NLD Standard			--	2
NLD Audio				--

What Each Comparison Will Show

- Comparisons 1 and 2
(Same Group, Different Mode)
  - For each group, does the accommodation change the functioning of the item?
  - However: Is the matching criterion the same (i.e., does the total score mean the same thing) if the mode of administration is different?

What Each Comparison Will Show (continued)

- Comparisons 3 and 4 (Same Mode, Different Groups)
  - Do the items function differently when the mode of administration is the same for both groups?

What Each Comparison Will Show (continued)

- Comparisons 4 and 5 (Different Mode, Different Groups)
  - – We do not think that these comparisons will yield useful data (though we welcome other viewpoints)

Procedures for Analyzing Data

- Differential Item Functioning: Mantel-Haenszel
- Differential Distractor Analysis: Standardized Distractor Analysis

Questions About Our Analysis Plans

- Do the comparisons that we have described make sense?
- Are the interpretations that we think we can make from these comparisons the appropriate ones?
- Do you have any other suggestions?

Questions?
Comments?

Next Steps for Differential Boost Data

Cara Cahalan-Laitusis
Educational Testing Service

- Examine impact of Decoding Measures
- Examine which factors best predict boost from Read Aloud or reduced score from Read Aloud
- Tryout Two (or Three) Staged Tailored Testing Model

Decoding and Fluency Measure

Means for Grade 4

	Non-LD			RLD		
Test/Condition	N	Mean	SD	N	Mean	SD
WJ Letter Word ID	604	504	21	469	473	29
WJ Word Attack	604	504	15	469	484	20
TOSWRF	604	102	10	469	89	12
WJ Fluency	604	501	24	469	474	21
Audio	604	502	32	469	477	30
Standard	604	497	37	469	457	31
Boost	604	5	24	469	19	27

Means for Grade 8

	Non-LD			RLD		
Test/Condition	N	Mean	SD	N	Mean	SD
TOSWRF	463	103	13	373	90	12
WJ Fluency	463	560	42	373	514	34
Audio	463	555	31	373	521	27
Standard	463	553	3	373	511	28
Boost	463	2	21	373	10	23

Correlations

	Standard	Audio	Boost	Standard	Audio	Boost
	Grade 4 RLD			Grade 4 no RLD		
Standard	1.00			1.00		
Audio	.56			.78		
Boost	-.52	.41		-.51	.14	
TOSWRF	.43	.23	-.23	.46	.42	-.15
WJ Fluency	.58	.38	-.25	.60	.55	-.20
WJ Letter Word ID	.53	.30	-.28	.59	.52	-.22
WJ Word Attack	.50	.27	-.28	.51	.42	-.23
	Grade 8 RLD			Grade 8 no RLD		
Standard	1.00			1.00		
Audio	.65			.79		
Boost	-.43	.40		-.43	.22	
TOSWRF	.36	.22	-.17	.36	.32	-.09 *
WJ Fluency	.47	.33	-.17	.47	.49	-.03 ns

All correlations are significant at .001 unless noted *p<.05 ns=not significant

Analyses

- RM ANOVA with decoding covariates
- Logistic regression to examine predictors of boost
- Simulate tailored test ideas

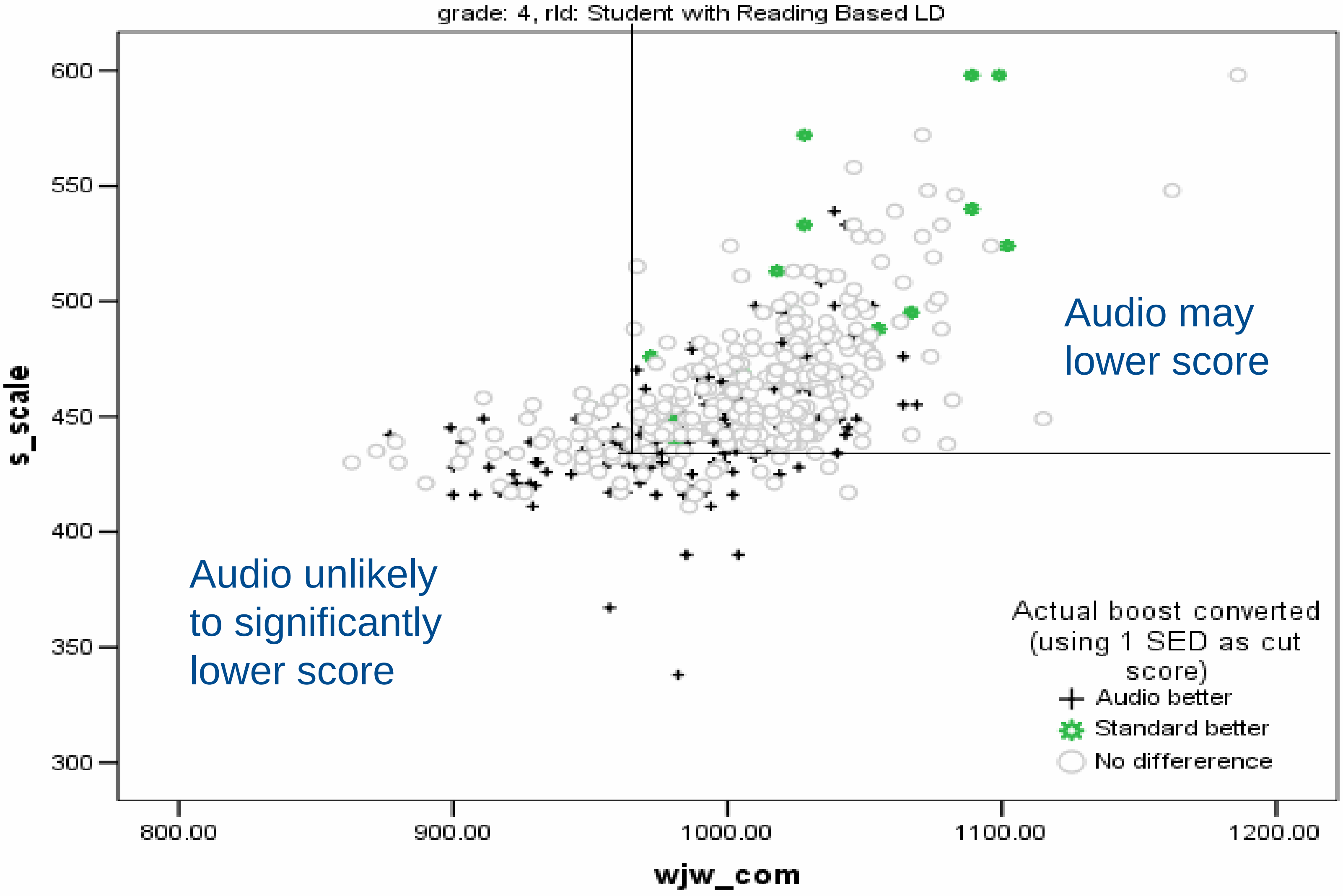
Other Questions/Comments

- Any other analyses of the decoding and fluency measures you would like to see?
- Are these correlations and means consistent with prior research?

IEP Decision Making

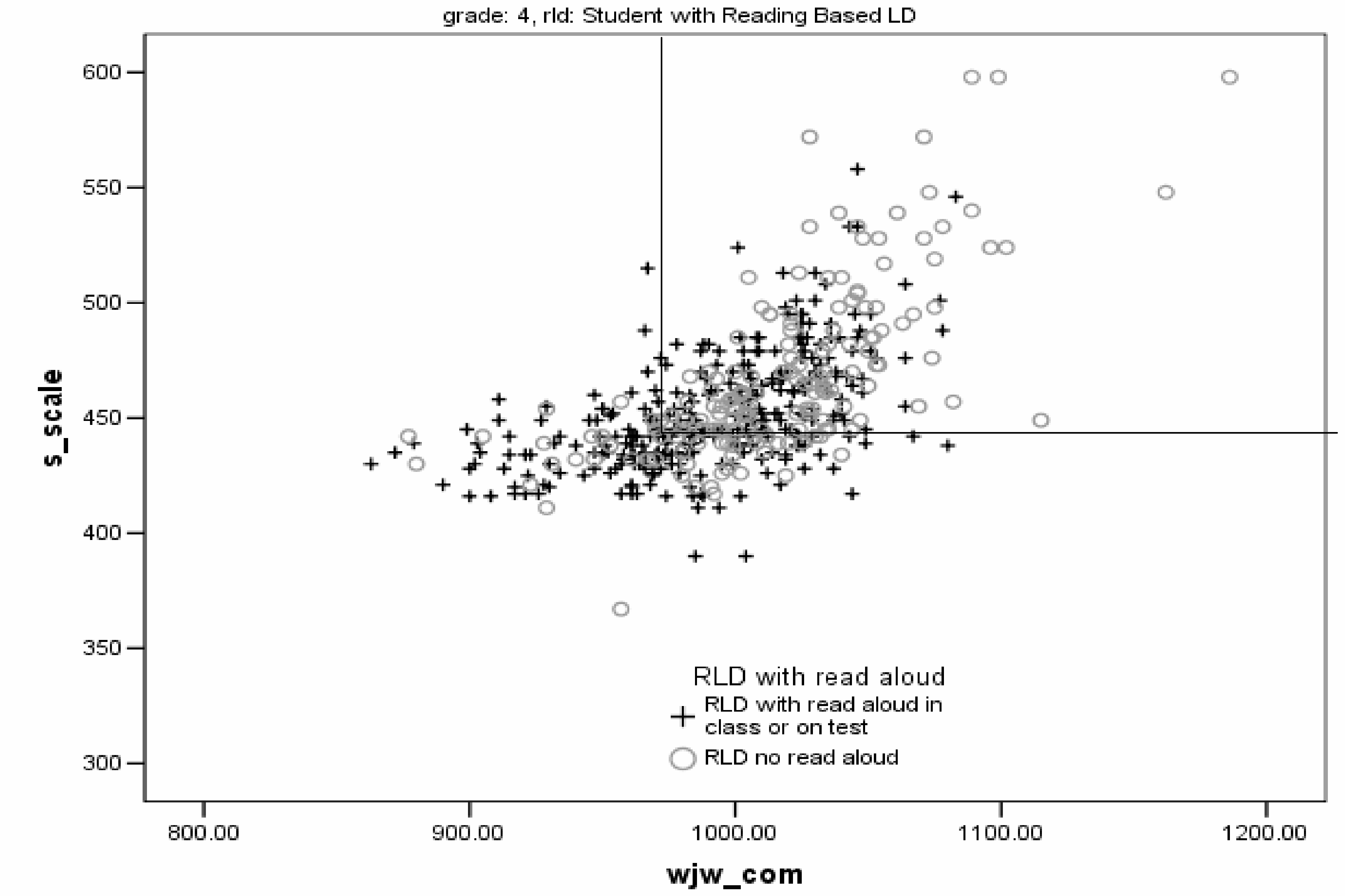
IEP Decision Making

- What factors contribute to boost?
  - – Low standard score
  - – WJ Reading Measures
  - – Teacher Predictions
  - – Student Preference



Audio may lower score

Audio unlikely to significantly lower score


grade: 4, rld: Student with Reading Based LD
s_scale
600
550
500
450
400
350
300
wjw_com
800.00
900.00
1000.00
1100.00
1200.00
RLD with read aloud
+ RLD with read aloud in class or on test
○ RLD no read aloud

Analyses planned

- Logistic regression analyses to predict boost for RLD students using
  - WJ scores
  - Standard score
  - Use of read aloud in class or on tests
  - Teacher predictions
  - NJ ASK from prior year

Questions

- Is there are better approach than logistic regression?
- If not, what cut score value should we use to determine boost?
  - Standard Error of Difference
  - 1/2 standard deviation
  - 1/10 of a standard deviation
  - Any boost

Questions

- What cut scores should be used for other measures?
  - Average score
- What other variables from our data set should we include in the analyses?

Two Staged Tailored Testing

Percentage of Students at Chance Level on STAR

		Percent Below Chance	
Grade	Group	Test Takers	Items
4	No Disability	2.2	1.3
	LD No accommodation	21.2	21.3
	LD Allowable accommodation	23.3	30.7
	LD Read Aloud	17.4	21.3
8	No Disability	1.5	0
	LD No accommodation	16.6	24
	LD Allowable accommodation	19.7	32
	LD Read Aloud	20	32

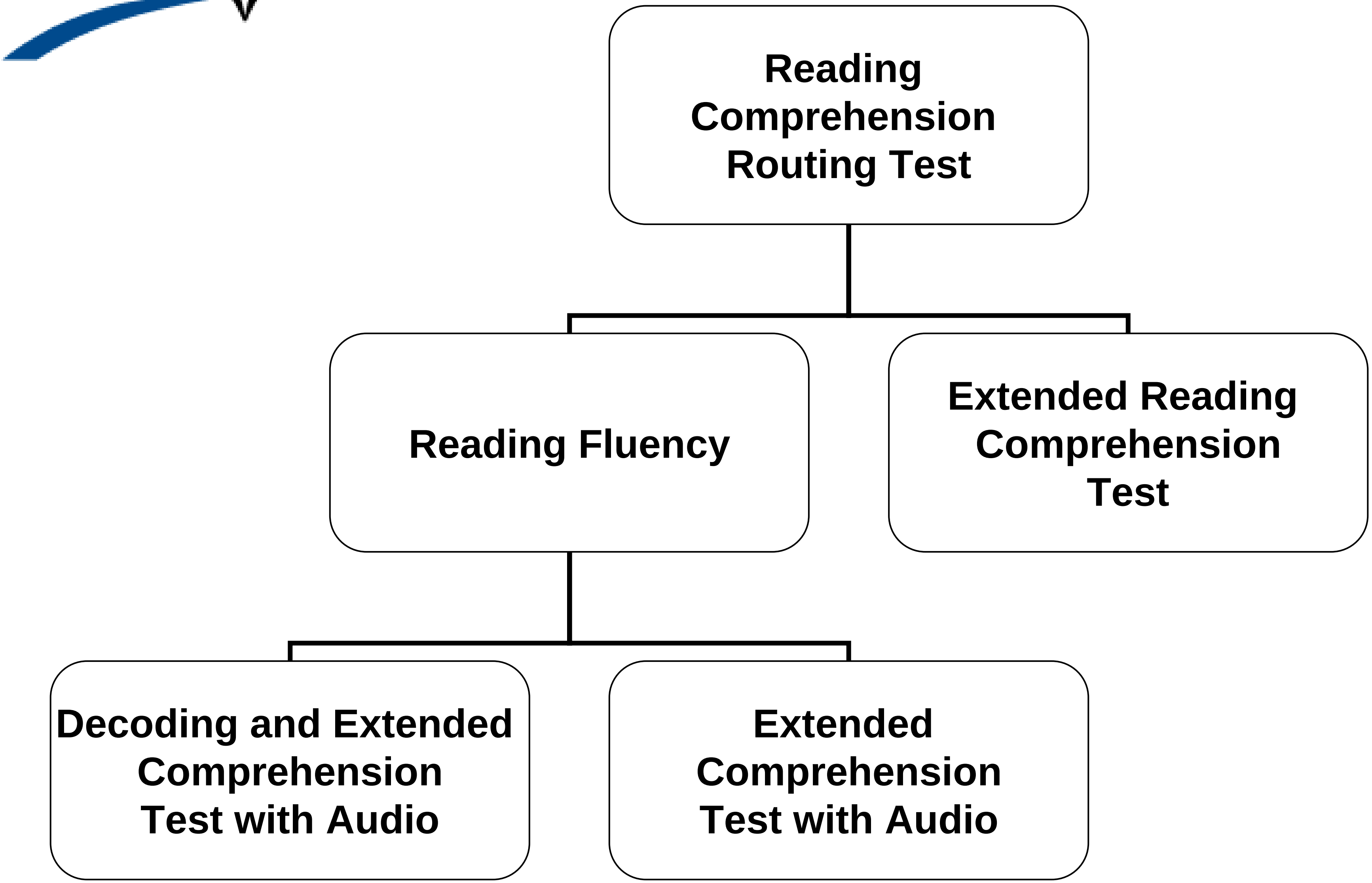
Percentage of Students at Chance Level on GMRT Standard

Grade 4	Test Takers (12 items or less)			Items (less than 30%)			
	Both Forms			Form S		Form T	
Group	Standard	Audio		Standard	Audio	Standard	Audio
Student with RLD	20.5%	4.4%		31.3%	2.1%	14.6%	8.3%
Student with no RLD	2.6%	1.5%		0.0%	0.0%	4.2%	2.1%

Grade 8	Test Takers (12 items or less)			Items (less than 30%)			
	Both Forms			Form S		Form T	
Group	Standard	Audio		Standard	Audio	Standard	Audio
Student with RLD	12.2%	4.8%		16.7%	12.5%	14.6%	10.4%
Student with no RLD	1.1%	0.2%		0.0%	0.0%	2.1%	2.1%

DARA Tailored Testing Model

- Two (or three) stages of testing
- Students subtests on stage 2 are determined by performance on routing test administered in stage 1
- Ideally computer administered but can be paper administered
- Some parts could be individually administered (e.g., decoding) if only a few students are routed into a decoding measure and this format reduces the number of students receiving individualize testing accommodations (e.g., read aloud by human)



Reading Comprehension Routing Test
Reading Fluency
Extended Reading Comprehension Test
Decoding and Extended Comprehension Test with Audio
Extended Comprehension Test with Audio

Advantages of Model

- Score is more reliable estimate since items are targeted to students ability level
- Students may feel less frustrated if they can do some of the items on the routing test
- Teacher receives more information on low performing students strengths and weaknesses
- Fundamental Skills and Comprehension are not confounded for students with poor fundamental skills (some LD) or poor comprehension (some LD and ELL)

Disadvantages of Model

- Requires computer administration or teacher scoring of items after stage 1
- Students who are routed to fluency test may be embarrassed
- Routing decision is made before test is scaled or standard setting is completed

Design Decisions from Lord (1977)

1. length of routing test,
2. length of second-stage test (s),
3. number of second-stage tests,
4. difficulty level of the routing test,
5. method of scoring routing test,
6. cutting scores on routing test for each second stage test,
7. method of scoring second-stage test,
8. method of combining scores from first and second stages

Questions for the DB Data?

- How many items (and of what difficulty) are needed for an accurate routing test? (Lord 1 and 4)
- Can we equate the audio extended and standard extended using the routing test? (Lord 8)
- What portion of students would be routed to fluency measure and what portion would be routed to decoding?
- Are the 2 alternate routes highly correlated with the standard administration?

Questions for the DB Data?

- What is the impact audio, fluency, and decoding scores on total test score.
  - If student is not a fluent reader should the total test score be non-proficient?
- Is the routing test accurate for all students?
  - Do some students do better on hard items?
  - Do some students having trouble with the first few items on the test?

Any Additional Questions or Comments?

Cognitive Labs

Teresa King
Educational Testing Service

ETS
ies NATIONAL CENTER FOR SPECIAL EDUCATION RESEARCH
Institute of Education Sciences

Overview

- Background
- Cognitive Labs
- Purpose of study
- Sample and test description
- Issues and proposed plans
  - Think aloud protocol
  - Data collection
  - Coding and analysis
- Questions for TAC

Background

- Cognitive labs using the think aloud method on reading comprehension questions
- Build off the findings of last year's large scale differential boost study
  - Gates MacGinitie Reading Comprehension Test
  - Use items found in preliminary findings of the DIF analysis of the GMRT data

Cognitive Labs

- Means of measuring mental processes through the use of a think aloud protocol
  - Unique and optimal way to capture otherwise unattainable information

- Involves 2 steps
  - Academic task
  - Verbal reporting task

Cognitive Labs Advantages

- Beneficial to learn about components of mental processes of reading (Afflerbach & Johnston, 1984, Alavi, 2005; Pressley & Afflerbach, 1995)
- Beneficial in the development of assessments (Caspar, Lessler, & Willis, 1999; Desimone & LeFloch, 2004; Willis, 2005)
- Open flexible procedure can be catered to the specific situation and activity (Davison, Vogel & Coffman, 1997)
- May use a small sample size
- Procedure has been successfully conducted with children as young as 3rd grade (Laing & Kamhi, 2002; Paulsen & Levine, 1999; Trambasso & Magliano, 1996)

ETS

ies NATIONAL CENTER FOR SPECIAL EDUCATION RESEARCH
Institute of Education Sciences

Cognitive Labs Disadvantages

- Thinking aloud is an unnatural step which may affect or interfere one's normal mental processes
- Students with disabilities may have difficulty with the procedure (Johnstone, Miller, & Thompson, in press)
- Responses have the potential to be incomplete or incorrect
  - Lack of desire/motivation
  - Embarrassment
  - Inability to understand the task

Purpose of Study

This study is being conducted to serve the following purposes:

1. How do students with and without reading-based learning disabilities differ as they approach a reading comprehension assessment?
2. Is this *type of information gathering* and *data quality* worthwhile to conduct in future large scale studies considering:
   - Age of students
   - Students with disabilities

Research Questions

1. In what way do students with reading-based disabilities respond differently to reading comprehension questions compared to students without disabilities?
2. What errors occur while reading the passage/reading the items?

Sample

- 50 students
- Students without LD and student with a documented reading-based LD as stated in an IEP
- 4th and 8th grade
- NJ public schools that participated in last year's differential boost study

Instruments

- Comprehension section of the Gates MacGinitie Reading test
  - Reading passages and multiple choice questions
- Think aloud protocol
- Measure of oral reading fluency
- Student survey

Think aloud protocol issues

Directions and practice

- Novel task for students
- Potential bias of responses

Concurrent versus retrospective verbal reports

- Concurrent reports
  - + More likely to remember mental processes
  - − Working memory may be overloaded when too many tasks are performed at once (Afflerbach & Johnston, 1984)
- Retrospective reports
  - + Frees up working memory to perform experimental task (Afflerbach & Johnston, 1984)
  - − Retrieval of thoughts allow more room for error (Leighton, 2004)

Data Collection Issues

- Length of session
  - Number of items able to be tested in one session is dependent upon the amount of verbal feedback provided
- Interaction during testing
  - Audio recording and note taking
  - Marked or unmarked passages
  - Probing

Proposed Study Design

- Individual administration
- 1 hour testing session
- Detailed instructions and practice
- Test full GMRT passage sets with DIF for non-LD and RLD students

Proposed Data Collection

- Concurrent
  - Test administrator will use probes during passage reading and test items when needed to elicit thoughts
- Retrospective
  - Follow up questions from the administrator if needed on an individual basis
  - Student survey

Data coding and analysis issues

- Accuracy
  - Important to carefully extract qualitative information because minor changes can affect the accuracy
- Coding
  - The amount of coding required is dependent upon the level and type of data needed
  - Categorization of student responses can be organized many ways

Coding Schema

- Errors on reading comprehension test can occur because of
  - Inability or inaccurate comprehension of passage
  - Inability or inaccurate understanding of a question and/or the answer options

Source of information

(Wolfe and Goldman, 2005)

- Self-explanations
- Surface text connections
- Irrelevant associations
- Predictions

Kendeau & van den Broek (2005)

- Understanding
- Uncertainty-confusion
- Explanations
  - Text-based
  - Knowledge-based
- Paraphrases

Level of Interpretation

(Wolfe and Goldman, 2005)

- Paraphrase
- Evaluation
- Comprehension Problems
- Comprehension Successes
- Elaborations

Type of Interpretation

(Laing & Kamhi, 2002)

- Literal
  - Paraphrases
  - Repetitions
- Inferential
  - Associative
  - Predictive
  - Explanatory
- Accuracy

Coding Test Items

(Tourangeau, 1984)

- Comprehension
- Retrieving relevant information
  - Within text
  - Prior knowledge
- Make judgment based on recall of information
  - Motivation of respondent
  - Social desirability of any response
- Mapping answer on the reporting system

Coding for Multiple Choice Reading Assessment

Text-Focused Processes	Test-Focused Processes		
Recall meaning	Guess	Refer to another question	Answer before reading options
Construct meaning	Use common sense	Think about an item’s clarity	Match options to text
Monitor meaning	Skip	Revise answer	Reflect on own performance
	Process of elimination		

Questions for the TAC

Any comments or suggestions on the following would be greatly appreciated:

- Test Administration
  - Directions
  - What is your opinion about asking students to think aloud their thoughts while they are reading?
  - Student reading passage and items silently vs. aloud

Questions for the TAC

- Measure of oral reading fluency
  - – Student reading section of passage aloud as a measure of reading fluency
  - – Woodcock Johnson reading fluency
- Sample
  - – Suggestions/ideas to make study more accommodating for students with reading-based learning disabilities?

Questions for the TAC

- Data collection
  - Recording
  - Should the test administrator read aloud to the student?
  - Are there any extraneous or confounding factors that we must control for that we are not?
- Coding schema
  - Best approach
  - Tie into reading comprehension theories?

DARA Proposal for Field Test (Goal 4)

Cara Cahalan-Laitusis
Educational Testing Service

NARAP Projects Goals

1. Develop a definition of reading proficiency
2. Research the assessment of reading proficiency
3. Develop research-based principles and guidelines making large-scale reading assessments more accessible for students who have disabilities that affect reading
4. Develop and field trial a prototype reading assessment

Goal 4 from RFP

“In collaboration with the other projects funded under this priority, and based on the definition formulated under Goal One and the research conducted under Goal Two, develop instruments and/or methods for assessing reading proficiency that are:

Goal 4 continued

- Suitable for large-scale administration for school accountability purposes.
- Accessible to students who have disabilities that affect reading,
- Maintain validity and comparability of scores,
- Can provide a valid measure of proficiency against academic standards.
- Can provide individual interpretive, descriptive, and diagnostic reports for the full range of students with disabilities that affect reading.

Goal 4 continued

"In collaboration with the other projects funded under this priority, a project must conduct a large-scale field test on the instruments and methods to determine the degree to which they provide for accessibility, validity, and comparability. The projects must provide sufficient sample size and diversity, as well as sound data collection and analysis procedures to ensure conclusive field test findings."

Primary Issues

- Development of an accessible and valid reading assessment.
- How to demonstrate that NARAP assessment more accessible while maintaining validity and comparability of scores.

Presentation

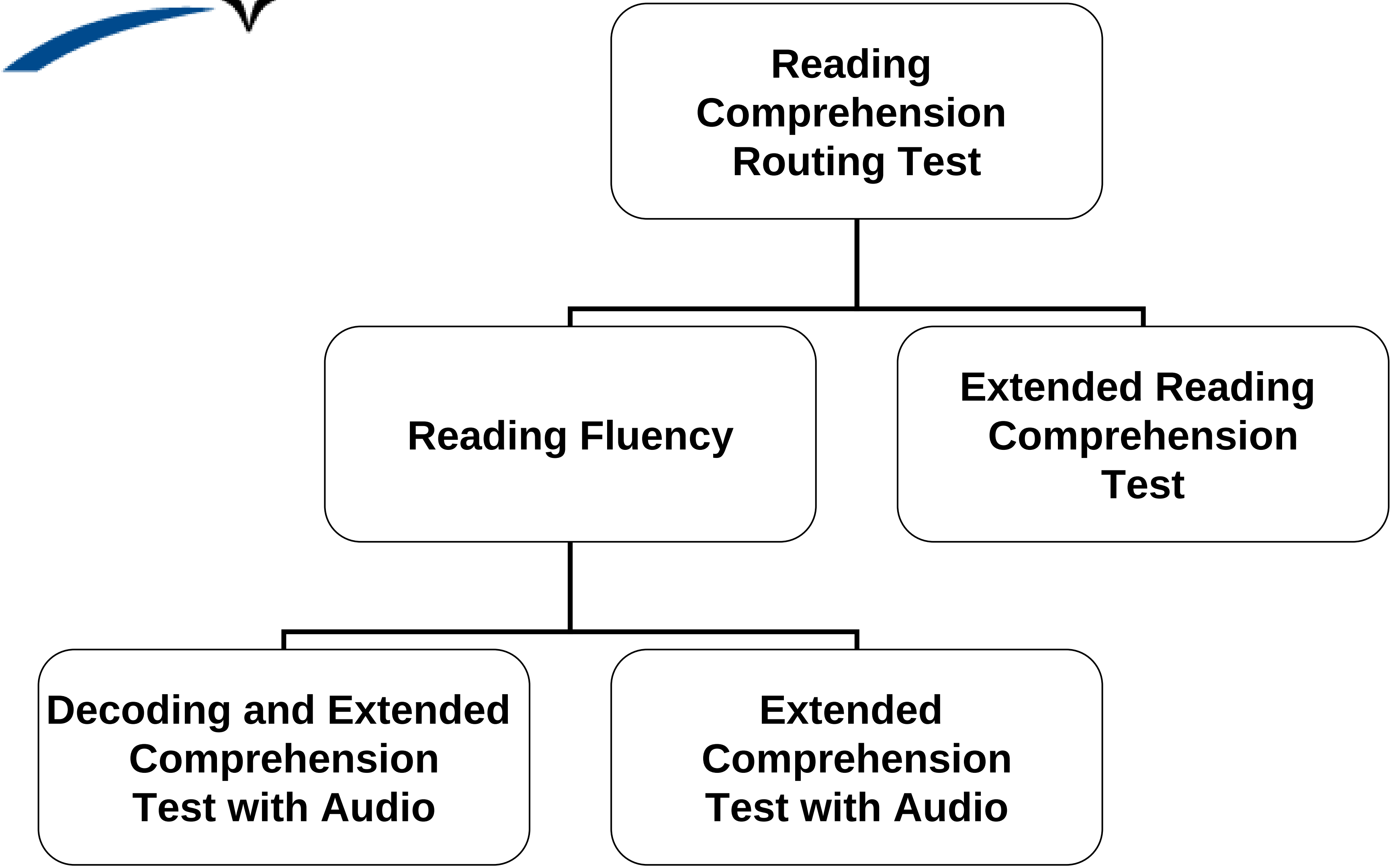
- Review requirements for Goal 4
- Sample sizes and disability subgroups included in each NARAP proposal
- Overview of the DARA projects initial vision for the assessment
- Ideas for ensuring conclusive field test findings that this assessment is valid and accessible

Original Sample Proposed

	PARA		DARA		TARA	
Disability Type	4	8	4	8	4	8
None	200	200	1000			
LD	200	200	1000			
SLI	200	200				
MR	200	200				
Deaf	100	100				
Hard of Hearing	100	100				
Low Vision						100
Blind (Braille)						50
Blind (Audio Users)						50

Data to Collect

- NARAP Assessment
- State Reading Assessment
- Survey data (teacher, student, and demographic)



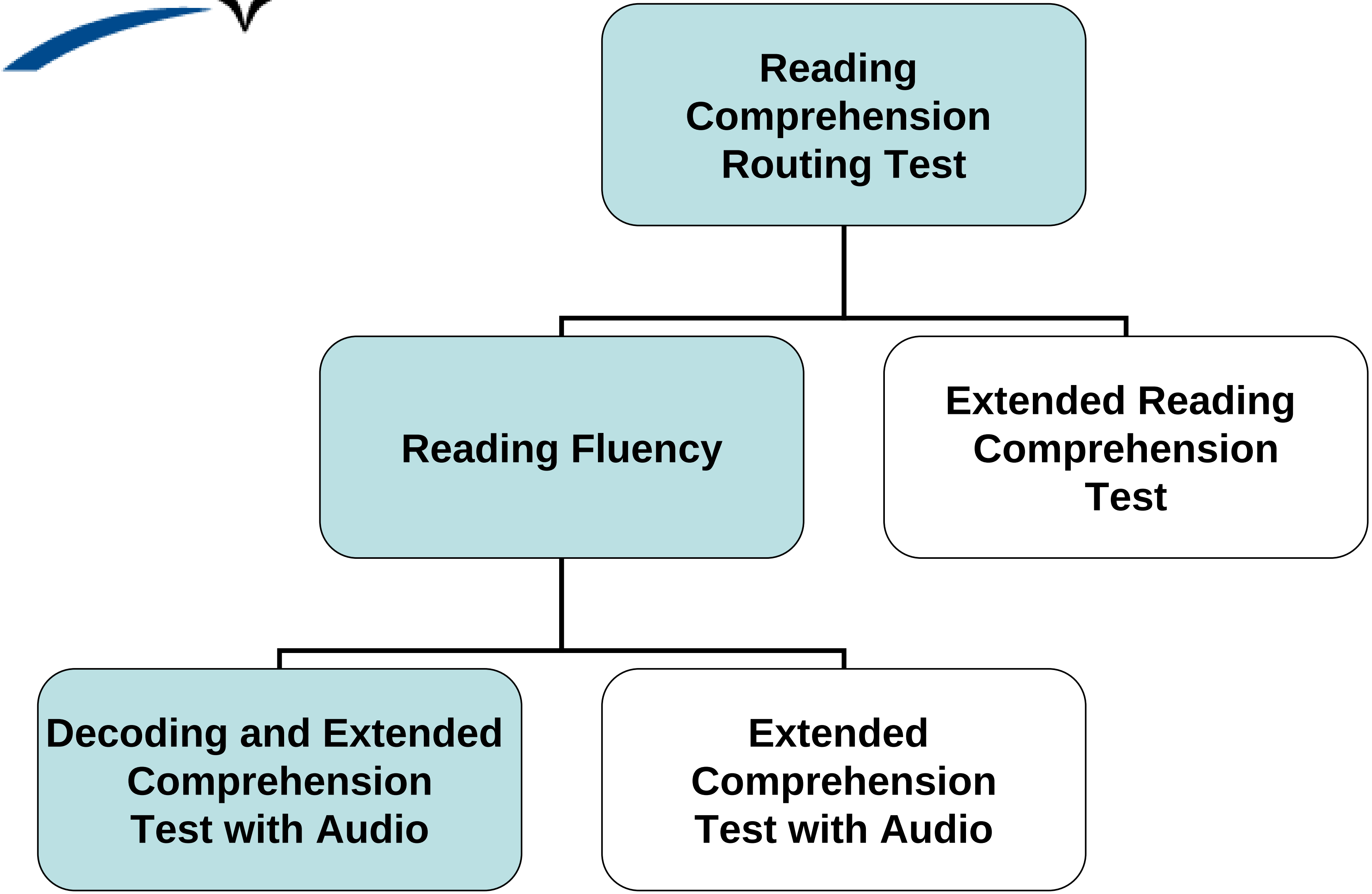
Reading Comprehension Routing Test

Reading Fluency

Extended Reading Comprehension Test

Decoding and Extended Comprehension Test with Audio

Extended Comprehension Test with Audio



Reading Comprehension Routing Test

Reading Fluency

Extended Reading Comprehension Test

Decoding and Extended Comprehension Test with Audio

Extended Comprehension Test with Audio

Possible Routes

Test 1=RC Routing + RC Extended

Test 2=RC Routing + RC Audio + Fluency

Test 3 = RC Routing + Fluency + Decoding + Audio

Also may consider another route for fluent readers w/ poor comprehension:

Test 4=RC Routing + Fluency + RC Extended

Possible Comprehension Scores

RC Routing + RC Extended

RC Routing + RC Extended w/ Audio

RC Extended w/ Audio

RC Extended

Lots of Questions for Year 3

- Is routing test accurate?
- Can scores be compared?
- How should we weight different measures?
- What portion of students would be routed to anything other than the extended RC test?

But our largest questions is regardless of our design . . .

Does the NARAP assessment result in a more accessible assessment while maintaining validity and comparability of scores from current state assessments?

Current Ideas

- Use differential boost framework to compare the changes in performance between students with and without disabilities on the NARAP assessment. Performance will be measures by rank ordering of scaled scores and z-scores.
- Examine the psychometric properties of both state and NARAP assessments to determine if the NARAP assessment results in reduced DIF and DDF and increased internal consistency compared to the state assessment.

Current Ideas

- Multiple regression analyses to determine which assessment is the best predictor of teacher’s ratings of reading comprehension
- Standard setting study using both the NARAP and state assessment items to determine changes in number of proficient students by disability subgroup.

- Reaction to these ideas
- Any other ideas
- Suggestions for pure measure of reading comprehension
  - Teacher ratings
  - Individually administered assessment
  - ???