CREATE
Center for Research on the Educational Achievement and Teaching of English Language Learners

Assessing Vocabulary Needed to Meet the New Standards

Diane August and Lauren Artzi
Center for Applied Linguistics

David Francis
University of Houston

Acknowledgements

Dorry Kenyon, Annie Duguay, Lindsey Massoud, Erin Haynes, Laura Wright

Chris Barr, Coleen Carlson, Ken Nieser

This project effort was supported by Grant Number 5P01HD039530-09 from the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Child Health and Human Development or the National Institutes of Health.

Overview of Presentation

- A variety of Instruments developed as part of the overall project
  - Test of Academic Vocabulary in English (TAVE)
    - Assessment of Multiword Knowledge
    - Word Associations Test
    - Test of Homonym Knowledge
    - Test of connectives
- Background
- Assessment development
- Analyses and results

Background

- The TAVE was developed, piloted, and revised prior to its administration with a larger sample of students.
- The TAVE was administered to 1,450 English learners (ELs) and English proficient students in a large urban district in the Southwest
  - Administered to students in grades 3-8
  - 100 ELs and 100 native-English speakers at each grade level.
- Two norm-referenced measures were also administered.

Norm-Referenced Measures

- The vocabulary subtest of the Gates-MacGinitie Reading Test (GMRT) was administered at each grade level to provide comparative data.
- Test of Silent Word Reading Fluency (TOWSRF) was administered to assess student's ability to recognize printed words accurately and efficiently.

Grade-level Form of the TAVE

- There is different form at each grade level.
- A grade-level test consists of four mini-tests, each composed of three units
- Each unit contains
  - Four items
  - A word bank with nine words: four target words and five distractors
- Participants are instructed to select a word from the word bank that matches a definition and completes a sentence.
- The test takes one hour to complete.

TAVE Unit (12 per grade-level form)

A. bold	B. chance	C. defeated
D. generous	E. important	F. jammed
G. skilled	H. solid	I. swift

1. _____: has great meaning or value
The picture is _____ to me because my dad drew it. ☐

2. _____: something that is stuck
The printer won't work because the paper is _____ in the printer. ☐

3. _____: something that is not hollow
The _____ iron bar is very heavy. ☐

4. _____: large in size
Tommy loves ice cream and cake. Tommy asks for _____ servings. ☐

CREATE

7

Rationale for Developing the TAVE

- Most vocabulary measures assess how well students compare with each other, but do not provide information about whether students know the meanings of words they are likely to encounter when reading grade level text.
- The TAVE assesses how well students are likely to understand words that appear frequently in grade-level text.
- Many standardized vocabulary measures do not have sufficient items at a level to be sensitive to interventions, but the TAVE does.

Challenges in Developing the TAVE

- Locating a corpus of words that provides information on grade-level frequencies
- Deciding which words from the corpus to retain in the domain of interest
- Assigning word meanings to word forms
- Identifying and measuring attributes that determine word difficulty

Locating a Corpus of High Frequency Words and Selecting Target Words

- Words were drawn from the most comprehensive and recent list of frequencies of words in written text in K-12 schools in the United States, *The Educator's Word Frequency Guide (EWFG)* (Zeno, Ivens, Millard, & Duvvuri, 1995).
- Words with frequency values between 10 and 999 were designated "target words" because Hiebert (2005) found words with these values accounted for 92% of words that appeared in NAEP and state reading tests.

CREATE

Assigning Word Meanings to Word Forms

- Because the *Educator's Word Frequency Guide* consists of word forms, grade-level meanings from the *Living Word Vocabulary (LWV)* (Dale & O'Rourke, 1981) were matched to the word forms.
- *The LWV* is a corpus of approximately 44,000 word meanings that indicate the grade at which 66% to 80% of students acquire word meanings.

Coding for Attributes that Might Determine Difficulty: Cognates

Target Word	LWV Grade Level	LWV Percent	LWV Meaning	Imageability	Cognate Status Word Form	Cognate Status Word Meaning	POS	Morphology
SECTION	4	0.67	A PART	2	C	CMC	noun	singular
SERVED	4	0.7	WAITED ON	2	C	CMC	verb	past
SHORT	4	0.79	NOT ENOUGH OF	3	N	N	adjective	bare
START	4	0.92	TO SET SOMETHING MOVING	2	N	N	verb	bare
WALK	4	0.88	GO BY FOOT	1	N	N	verb	bare
FORM	6	0.71	KIND OR VARIETY	4	C	CMN	noun	singular
CENTER	8	0.69	COMMUNITY HOUSE	3	C	CMC	noun	singular
BENT	10	0.76	FORCED	3	N	N	verb	irregular past
CAST	12	0.73	A LIGHT COVERING OF COLOR	4	N	N	noun	singular
CENTER	12	0.86	POLITICALLY NEUTRAL	4	C	CMC	noun	singular

CREATE 12

Establishing Inter-rater Reliability: Cognates

- Two bilingual coders rated cognate status.
- 500 target words were randomly selected to be coded by both coders to establish inter-rater reliability.
- The remaining 4,828 words were randomly assigned to the two coders and coded independently.
- At four time points, inter-rater reliability was assessed for the remaining words (Cohen's Kappa > .75).

Coding for Attributes that Might Determine Difficulty: Imageability

Target Word	LWV Grade Level	LWV Percent	LWV Meaning	Imageability	Cognate Status Word Form	Cognate Status Word Meaning	POS	Morphology
SECTION	4	0.67	A PART	2	C	CMC	noun	singular
SERVED	4	0.7	WAITED ON	2	C	CMC	verb	past
SHORT	4	0.79	NOT ENOUGH OF	3	N		adjective	bare
START	4	0.92	TO SET SOMETHING MOVING	2	N		verb	bare
WALK	4	0.88	GO BY FOOT	1	N		verb	bare
FORM	6	0.71	KIND OR VARIETY	4	C	CMN	noun	singular
CENTER	8	0.69	COMMUNITY HOUSE	3	C	CMC	noun	singular
BENT	10	0.76	FORCED	3	N		verb	irregular past
CAST	12	0.73	A LIGHT COVERING OF COLOR	4	N		noun	singular
CENTER	12	0.86	POLITICALLY NEUTRAL	4	C	CMC	noun	singular
DRAW	12	0.63	TO GET, RECEIVE	2	N		verb	bare

CREATE

14

Coding for Attributes that Might Determine Difficulty: Imageability

Try to image a word. How much effort is required?

Consider the following when rating:

- Number of frames* with more or less the same picture
- Amount of context within a frame (how much of the frame needs to be filled up to image a word)
- The imageability of each of the elements in a frame, excluding the target word
- The number of elements in each frame
- Whether the target word is an element in the frame or is defined by the relationship among the elements in the frame

*Frame refers to the borders of the whole image; picture refers to the whole image in each frame; element refers to the objects in a picture

Example Imageability Ratings

Target Word	LWV Definition	Rating
row	paddle a boat	1
aboard	on a ship	1
dinosaur	animal no longer living	1
direct	to order, command	2
directed	ordered, commanded	2
dining	eating	2
direct	to control or manage	3
directed	controlled or managed	3
abandoned	gave up	3
narrow	lacking a broad view	4
spirit	special quality	4

Establishing Inter-rater Reliability: Imageability

- Four coders rated imageability
- 350 target words were randomly selected to be coded by all four coders. The remaining 4978 words were randomly assigned to the four coders and coded independently.
- At four time points, inter-rater reliability was assessed for the remaining words and was high: Kendall's concordance > .75

Coding for Part of Speech and Morphology

Target Word	LWV Grade Level	LWV Percent	LWV Meaning	Imageability	Cognate Status Word Form	Cognate Status Word Meaning	POS	Morphology
SECTION	4	0.67	A PART	2	C	CMC	noun	singular
SERVED	4	0.7	WAITED ON	2	C	CMC	verb	past
SHORT	4	0.79	NOT ENOUGH OF	3	N		adjective	bare
START	4	0.92	TO SET SOMETHING MOVING	2	N		verb	bare
WALK	4	0.88	GO BY FOOT	1	N		verb	bare
FORM	6	0.71	KIND OR VARIETY	4	C	CMN	noun	singular
CENTER	8	0.69	COMMUNITY HOUSE	3	C	CMC	noun	singular
BENT	10	0.76	FORCED	3	N		verb	irregular past
CAST	12	0.73	A LIGHT COVERING OF COLOR	4	N		noun	singular
CENTER	12	0.86	POLITICALLY NEUTRAL	4	C	CMC	noun	singular
DRAW	12	0.63	TO GET, RECEIVE	2	N		verb	bare

Coding for Part of Speech (POS)

- Word meanings were coded for part of speech.

target_word	LWV_definition	POS
LIGHT	WHAT YOU SEE BY	noun
LIGHT	TO START THE FIRE	verb
LIGHT	NOT HEAVY	adjective
LIGHT	NOT SERIOUS	adjective
LIGHT	GRACEFUL	adjective
LIGHT	COME TO REST	verb
LIGHT	CHEERFUL	adjective
LIGHT	PALE IN COLOR	adjective
LIGHT	SMALL IN AMOUNT	adjective

Example Morphology Codes

- Word meanings were coded for morphology.

POS	Morphology	Explanation	Examples
Noun	gerund	Nouns derived from verbs by the suffix -*ing*	*saying: wise statement* *meaning: the sense of the words*
	singular	Singular count nouns[1] with no other morphology	*fox: dog-like animal* *gift: a present*
	plural	Plural count nouns with regular plural morphology	*foxes: dog-like animals* *gifts: presents (n)*
	irregular plural	Plural count nouns with irregular plural morphology (i.e., plural not formed by –*s*)	*geese: duck-like birds* *teeth: what you bite with*
	mass	Nouns without possible plural/singular distinction	*gas: gasoline* *tin: a material for cans*

Assigning Predictive Difficulty

- The TAVE was piloted in Grades 3-6
- Empirical estimates of word difficulty were obtained for 222 target vocabulary words.
  - Predictors included LWV grade level and percent, U value, lexile level, cognate status, and imageability.
- A regression model (referred to as model 1) was constructed to predict empirical item difficulties for the 222 words from item characteristics in our database.
- The regression model was used to calculate estimated difficulty values for all 14,000 words.

Assigning Predictive Difficulty

- Then these predicted difficulties were rescaled onto a developmental metric from 1 to 14, 000 (total number of word meanings in the final corpus), such that a student's score would directly correspond to the number of word meanings known by that student.

Assigning Predictive Difficulty

- Once the words were placed on a developmental metric the primary issue was to determine what developmental metric ranges were appropriate at a given grade level.
- The Typical Reader Lexile ranges were used to determine cut off points on our developmental scale at each target grade level (Grades 3-8).

Typical Reader Lexile values by grade		
Grade	25%	75%
3	330	770
4	445	810
5	565	910
6	665	1000
7	735	1065
8	805	1100

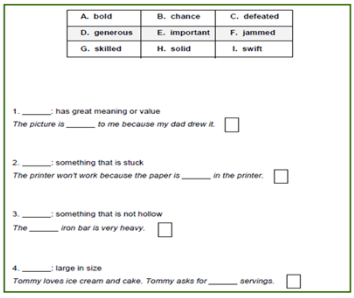
Example for Third Grade

- A subset of only the word meanings with lexile values between 330 and 770 was obtained.
- Next the mean and standard deviation of the developmental metric for that set of words was computed.
- Then the grade appropriate range on the dev. metric was considered to be this mean +/- 2 SD or 4211-6658 (min and max could not be used because of outliers)

Third Grade Dev Scale Descriptives (Lexile Range 330-770 Only)				
Variable	Mean	SD	Min	Max
Dev Scale	5434.74	611.8	4205.11	8268.25

Forms Development

- Example of a unit
- We controlled for part of speech within each unit
- The proportion of nouns (and other parts of speech) in the assessment reflected the proportion of nouns (and other parts of speech) in the database
- Three units made up a form



A. bold	B. chance	C. defeated
D. generous	E. important	F. jammed
G. skilled	H. solid	I. swift

1. _____: has great meaning or value
The picture is _____ to me because my dad drew it. ☐

2. _____: something that is stuck
The printer won't work because the paper is _____ in the printer. ☐

3. _____: something that is not hollow
The _____ iron bar is very heavy. ☐

4. _____: large in size
Tommy loves ice cream and cake. Tommy asks for _____ servings. ☐

Item Development

- LWV word meanings were modified to make them child friendly.
- Sentence stems were developed that conformed with certain Lexile levels.
  - Grade 3: Third grade lexile level
  - Grades 4-8: Fourth grade lexile level
- To vertically scale forms, two units from the grade level below one unit from the grade level above were randomly chosen to be used at the target grade level.

Example Units (Verb, 3rd Singular)

3rd grade

A. belongs	B. cages	C. grows
D. leaves	E. seals	F. suits
G. treasures	H. varies	I. wheels

1. _____: becomes
The weather _____ colder in the winter. 2

2. _____: changes often; becomes different
Carol _____ the games she plays. Carol plays soccer or cards. 4

3. _____: is right or acceptable for someone or something
This soft bed _____ me when I am tired. 3

4. _____: goes away from a place
Ana _____ school at 3pm. 1

8th grade

A. becomes	B. commands	C. features
D. nurses	E. processes	F. rises
G. tracks	H. voices	I. wears

1. _____: looks good on someone
Amy has a great new haircut. It _____ her. 4

2. _____: changes something by following certain steps
The factory _____ paper so it can be used again. 3

3. _____: gives special attention to something
The newspaper _____ an article about the cars that actors drive. 2

4. _____: grows bigger or higher
The bread dough _____ slowly until it is big enough to bake. 1

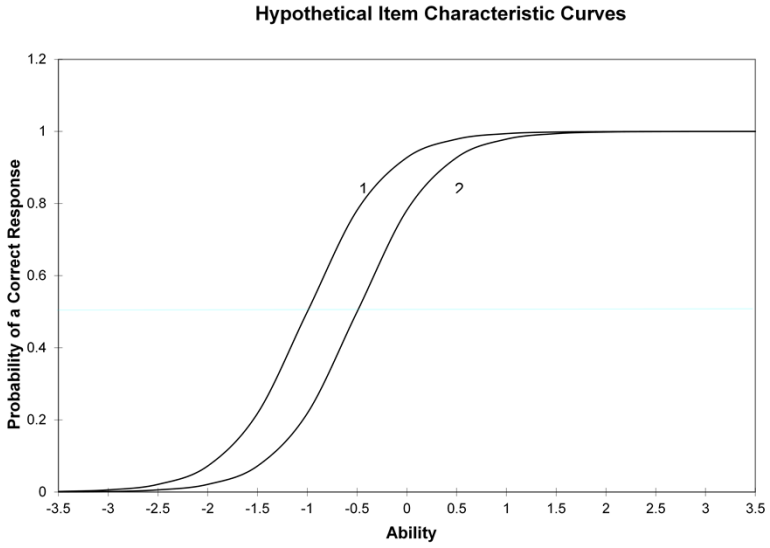
Research Questions

- Do the data fit the measurement model?
- To what extent do the empirical difficulties match the predicted difficulties? Can we generalize to the database/word corpus?
- What is the reliability of the TAVE?
- What is the relationship between the TAVE and other measures given?
- Does the TAVE perform in the same way for children of different language backgrounds?
- How do third grade children perform on the TAVE as compared to eighth grade children?

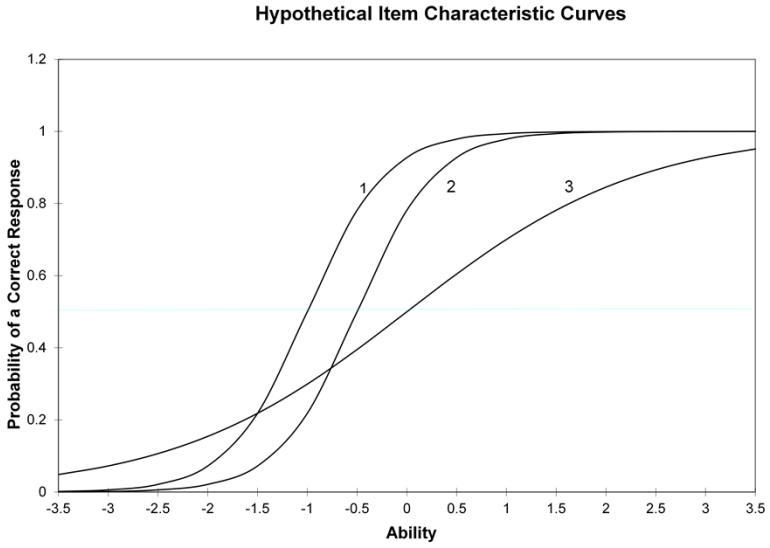
Do the data fit the measurement model?

- There are several critical features of the model
  - Vocabulary Knowledge represents a single ability
  - Once vocabulary knowledge is controlled, responses to different items are unrelated ("local independence")
    - That is, once vocabulary knowledge is controlled, a student's score on one item does not predict their score on another item
    - This is just a fancy way of saying that all of the information that is shared by pairs of items is explained by the dimension we call "vocabulary knowledge"
  - Items differ only in their difficulty and not in their relationship to Vocabulary Knowledge

Item Difficulty and Discrimination



Item Difficulty and Discrimination



Do the data fit the measurement model?

Dimensionality

- The first step is to see how much of the information in items can be captured by a single dimension
  - 40.3% of the information shared across items was explained by the first dimension
- Second step is to see how much more information can be captured by adding dimensions
  - Each of the next three dimensions explained < 1% of the remaining information (0.8%, 0.6%, and 0.6%, respectively)

Do these other dimensions relate to characteristics of the items?

- Although the amount of information in the second dimension is small, it appears that it may relate to how abstract the item is.
- Model results: $F(4,223) = 2.12$, $p=.08$, $R^2=.02$

Variable	Parameter Estimate	Standard Error	t Value	*p*	Standardized Estimate	Pearson's Correlation
Intercept	0.03	0.10	0.30	0.77	0	
Imageability	-0.02	0.01	-2.43	0.02	-0.16	-0.18*
LWV Grade Level	-0.002	0.003	-0.65	0.51	-0.05	-0.08
LWV Percent	0.06	0.12	0.53	0.60	0.04	0.08
Cognate Status	-0.01	0.01	-0.44	0.66	-0.03	-0.04

Do these other dimensions relate to characteristics of the students?

- Yes, there are relations between student characteristics and the second factor
  - Word knowledge measured on the GMRT correlated .25
  - Silent word reading fluency correlated .26
- Relations are negligible when vocabulary factor is controlled. See regression results below.

Outcome	Variable	Standardized Estimate 1	*F* value(R^2)	Standardized Estimate 2	*F* value(R^2)
Word Knowledge	Vocabulary Factor Score	.78*	3431.7 (.62)	.79*	1717.9 (.62)
	Residual Factor Score			-.02	
Silent Word Reading Fluency	Vocabulary Factor Score	.61*	1244.0 (.37)	.59*	627.6 (.37)
	Residual Factor Score			.05*	

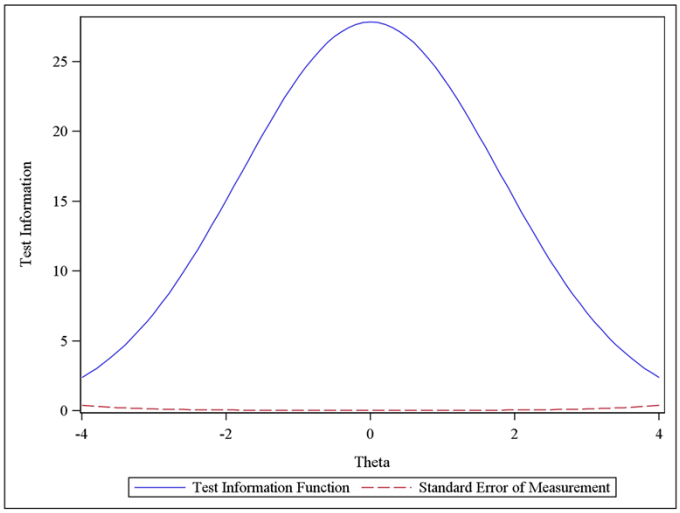
CREATE 34

All Gates analyses use extended scale score scores

To what extent do the empirical difficulties match the predicted difficulties?

- Empirical difficulty
  - Students in grade 3-8
  - Estimated from the unidimensional measurement model.
  - Twelve overlapping items between adjacent grades allowed all difficulty estimates to be on the same scale
  - Empirical estimates were correlated with predicted estimates obtained from pilot models
    - *r* = .59 between empirical and predicted difficulty estimates

What is the reliability of the TAVE?



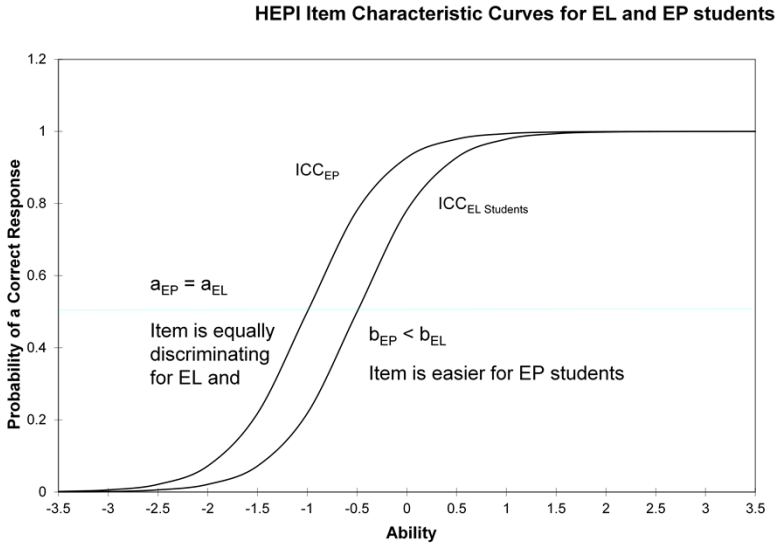
What is the relationship between the TAVE and other measures in the test battery?

- TAVE scores have reasonably high correlations with other measures in the battery
  - Gates-MacGinitie Word Knowledge $r = .78$
  - Test of Silent Word Reading Fluency, $r = .61$
- These correlations are in the expected direction
  - positive
  - larger for Word Knowledge than for Fluency

Does the TAVE perform in the same way for children of different language backgrounds?

- We always strive to develop tests so that they function equally well for students in different subgroups of our target population.
- In the case of TAVE, it is critically important that TAVE measure vocabulary equally well for EL and EP students.
- To examine this question, we employ a set of measurement procedures that look at how individual items function for EL and EP students.
- Specifically, we will look for evidence of ***differential item functioning*** between EL and EP students.

Item Difficulty and Discrimination



Differential Item Functioning (DIF)

- Differential Item Functioning (DIF) was examined in three ways
  - Using the Mantel-Hanzl significance test, which asks whether a particular item functions comparably
  - Using item characteristics to predict DIF
    - Point biseral correlations to predict DIF from individual item characteristics
    - Using logistic regression analysis to predict the presence of DIF from sets of item characteristics
  - Estimating item difficulty separately for EL and EP students and examining the effects of group (i.e., EL vs EP), item characteristics, and their interaction on item difficulty.

Mantel-Hanzl DIF test

- Using a criterion of 5%, 46 out of 228 (i.e., approx. 20%) of the TAVE items demonstrated DIF.
- 46 Target words:

ACID	CONTENT	LEAVES	SHIPS
ADDRESSED	COPY	LOAD	SINGULAR
AHEAD	CORNERS	MANNERS	STANDARDS
AROUND	DEFINITION	OBJECT	STIFF
ASSIGNED	ESTABLISHED	PIECES	STILL
BLEW	EXPOSED	PLAIN	SUSPECT
BREEZE	EXTENDED	PROBLEM	TWISTING
CAST	FINDINGS	RECEIVING	WEATHER
CEREMONY	FREQUENCY	RELATE	WEAVE
CODE	HESITATED	RELATIVE	WELL
CONCRETE	IMPORTANT	RULE	WINGS
CONTAIN	JAMMED		

Presence of DIF was unrelated to Item Characteristics

- We attempted to predict the presence of DIF from item characteristics, but they were not related.
- DIF was more weakly correlated with cognate status
  - DIF was more likely for cognates than non-cognates
  - DIF was less likely for cognates when other attributes were controlled.

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Point Biserial Correlation
Intercept	1.59	2.32	0.47	0.4945	
Imageability	0.30	0.20	2.31	0.13	-0.10
LWV Grade Level	-0.01	0.06	0.02	0.90	-0.02
LWV Percent	-1.12	2.71	0.17	0.68	0.04
Cognate Status	-0.42	0.34	1.60	0.21	0.08

Magnitude of DIF and its Relation to Item Characteristics

- On average item difficulties did not differ between EL and EP students.
- Item characteristic predicted item difficulty comparably for EL and EP students.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-1.01	1.13	-0.89	0.37
Group	0.13	1.59	0.08	0.93
Imageability	0.43	0.10	4.48	<.0001
Group X Imageability	-0.08	0.14	-0.61	0.55
LWV Grade Level	0.24	0.03	7.97	<.0001
Group X LWV Grade Level	-0.02	0.04	-0.39	0.70
LWV Percent	-3.00	1.31	-2.29	0.02
Group X LWV Percent	0.47	1.86	0.25	0.80
Cognate Status	0.08	0.16	0.46	0.65
Group X Cognate Status	-0.28	0.23	-1.19	0.23

CREATE 43

Note: intercept is EP kids as in EP = 0 ELL = 1 from a regression parameter standpoint.

How do children perform on TAVE as a function of grade and English language proficiency?

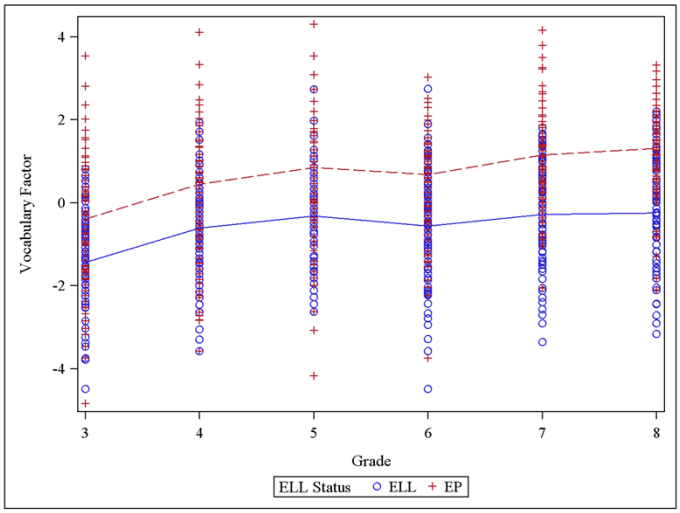
- Overall, EPs outperformed ELLs.
- Students in higher grades outperformed those in lower grades.
- Smaller differences were observed in middle school grades. However, a similar pattern was observed for GMRT Word Knowledge. Results may reflect sampling differences in grade cohorts.

	EP Students			EL Students			
Grade	N	Mean	Std Dev	N	Mean	Std Dev	Difference
3	149	-0.39	1.52	111	-1.44	1.10	1.05
4	194	0.45	1.38	107	-0.61	1.16	1.06
5	167	0.85	1.42	82	-0.32	1.15	1.17
6	170	0.68	1.03	176	-0.57	1.29	1.25
7	154	1.15	1.04	167	-0.28	1.30	1.43
8	115	1.32	1.00	130	-0.24	1.27	1.56

CREATE 44

Correcting for family-wise error, all 2 grade mean comparisons were significant excepts for the following pairs (8&7; 8&5; 7&5; 6&4)

TAVE by Grade and ELL Status

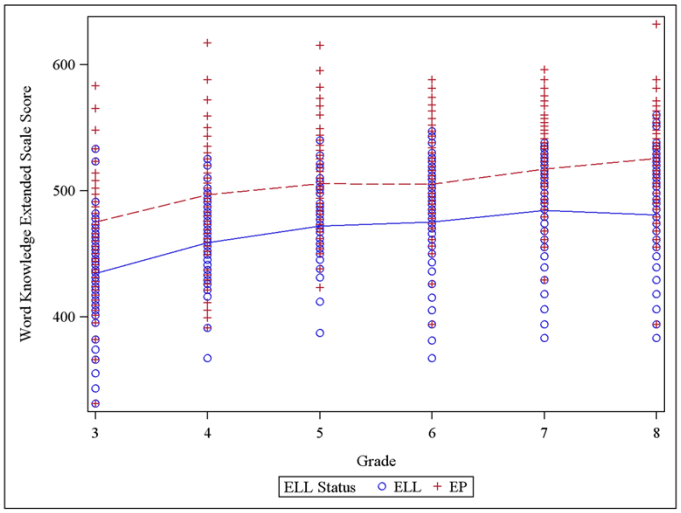


How do children perform on TAVE as a function of grade and English language proficiency?

- GMRT Word Knowledge was also examined by grade and ELL status.
- EPs outperformed ELLs.
- For the most part, higher grades outperformed lower grades.

	EP Students			EL Students			
Grade	N	Mean	Std Dev	N	Mean	Std Dev	Difference
3	147	475.2	45.7	110	434.3	31.9	40.9
4	189	496.7	43.1	107	458.6	25.9	38.1
5	167	505.7	36.3	80	471.8	23.9	33.9
6	167	505.0	28.5	174	475.2	29.7	29.8
7	146	517.1	28.2	158	484.2	33.6	32.9
8	115	525.4	30.5	128	480.6	38.7	44.8

Word Knowledge by Grade and ELL Status



Moving Forward

- Determine accuracy of predictive model – empirical estimates obtained from student testing will be compared to predicted estimates of difficulty.
- Model refinement – model parameters will be refined base on the larger set of data.
- Across grade equating – all grade level test forms will be put on a common scale, based on common across grades items.

Thank You!