Are Predictor Variables for Item Difficulty of L2 Sentence Repetition Consistent Between English and Chinese?

April 5, 2012 at LTRC 2012

Masanori Suzuki
Knowledge Technologies, Pearson

Overview

- Sentence Repetition
- Context of the Study
- Research Questions
- Predictor Variables
- Results
- Discussion & Summary

Sentence Repetition (a.k.a Elicited Imitation)

Test-takers repeat the sentence they hear

Once in a while, you find a really nice one

- Task integrating listening and speaking
- Used as a technique to elicit spoken language performance in FLA (e.g., Smith, 1970), SLA (e.g., Erlam, 2009), bilingualism (e.g., Radloff, 1991)
- Used in some high-stakes tests (e.g, PTE-Academic, Versant)
- Strong correlation with an overall spoken language proficiency

Sentence Repetition (a.k.a Elicited Imitation)

- The amount of linguistic units you can repeat correctly is dependent on the degree of your familiarity with vocabulary and structures of the language in question
- Not rote memorization once items become long enough (e.g., Vinther, 2002; Erlam, 2006, 2009)
- Miller (1956): 7 +/- 2

Context of The Study

- Spoken Chinese Test under development
- Comprised of 9 tasks
- One of the tasks is “Sentence Repeat”
- About 25 min (Sentence Repeat is about 3 min)

- What are the factors that affect the difficulty of sentence repeat items?

Potential Factors

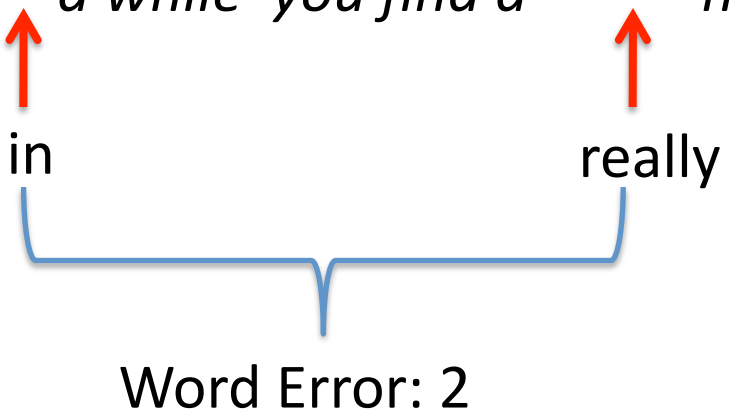
- What characteristics of item prompts potentially affect the difficulty of sentence repeat items?

- Length
- Speed of item recording
- Lexical
- Syntax
- Phonological characteristics of recording voices (accent)

Item Difficulty

Item: *Once in a while, you find a really nice one*

L2 learner response: *once* *a while* *you find a* *nice one*



- Number of word errors computed
- Partial-Credit Rasch Model

Previous Research

- Nair (2005) studied 18 linguistic and acoustic properties of 846 English sentence repetition items

Predictor variable	Correlation
log number of phonemes	0.75
log number of words	0.69
log duration of prompt	0.62
Avg. tirgram log likelihood	-0.60
log number of content words	0.60
Number of syllables	0.61
Phonemes per second	0.39

- Using multiple regression, predicted Rasch-based item difficulty (r=0.83)

Research Questions

RQ 1: What is the relationship of each of several individual variables with the Rasch-based difficulty values of sentence repeat items in Chinese?

RQ 2: What combination of variables can predict the difficulty of sentence repeat items in Chinese?

RQ 3: Are predictor variables consistent for sentence repeat items in English and in Chinese?

Items

- 458 sentence repeat items in Chinese
- Drafted by educated native Chinese item developers
- Syntactical structures adopted from authentic TV or conversation transcripts
- Vocabulary restricted to a corpus of apprx. 5,000 words
- Recorded by native Chinese speakers at natural, conversational pace

Item Example

- 要 下雨 了。

  It's going to rain.

  – 4 characters/syllables, 3 words

- 后来 他 又 去 了 一 次。 

  Afterwards, he went back again.

  – 8 characters/syllables, 7 words

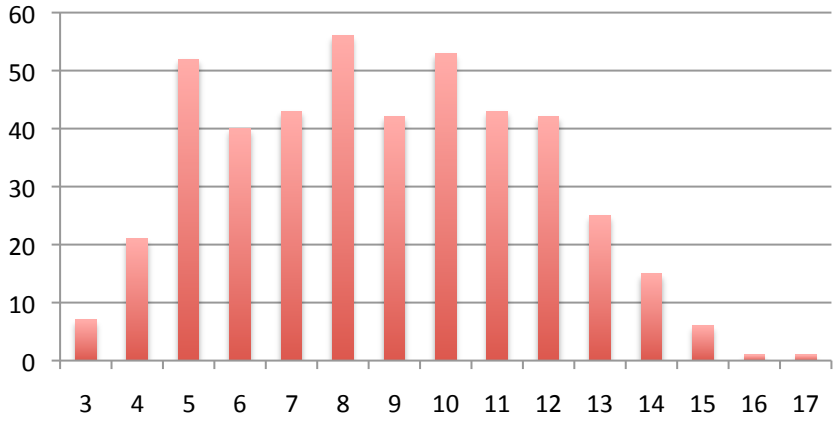
- 报纸 上 说 明天 下午 两 点 开始。

  The newspaper says that it starts at 2pm tomorrow afternoon.

  – 12 characters/syllables, 8 words

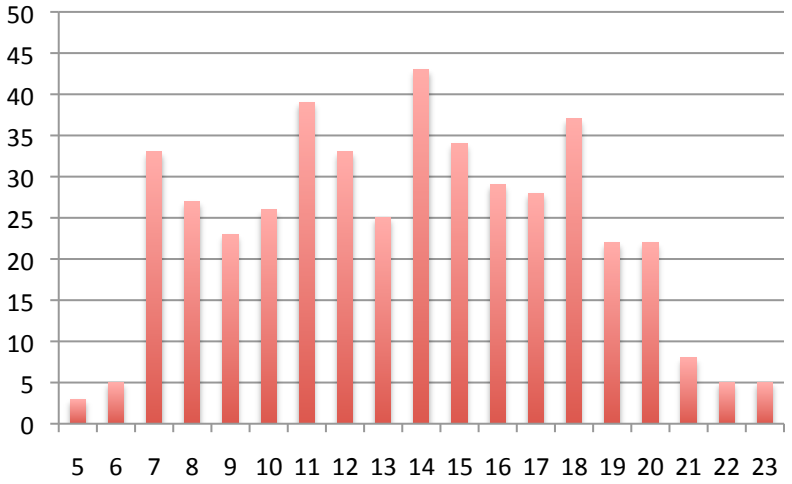
Number of Words and Characters distribution

Histogram: Number of Words



Mean: 8.72 words
(SD = 2.95)

Histogram: Number of Characters



Mean: 13.60 words
(SD =4.24)

Data Collection

- Time: Oct 2010 – Jan 2011
- Prototype version of the Spoken Chinese Test
- Repeat: 24 items per test, 3 minutes per test
- Delivery: On the phone or via computer
- Number of tests collected:
  - 511 tests from native speakers
  - 1,083 tests from learners of Chinese as a L2

Learner Data

- 1,083 tests from 715 learners of Chinese
- Mean age = 24.6
- Male-female = 52%-47%
  (1 % unreported)
- 39 L1s

L1	%
Japanese	26%
English	11%
Arabic	8%
Korean	7%
Spanish	7%

Predictor Variables Examined

- A total of 15 variables
- 7 Item variables
  - length
  - speed
- 8 Native performance variables

Item Variable: Length

1. Log number of words
2. Log number of syllables (=characters)
3. Log number of phonemes
4. Log duration of item prompts

Item Variable: Speed

- Computed based on item prompt recordings

5. Words per second
6. Phonemes per second
7. Syllables per second

Native Performance Variables

- Assumption: More difficult items may take more time for native speakers to respond

8. Leading silence: Mean
9. Leading silence: Median

10. Word duration: Mean
11. Word duration: Median

12. Inter-word silence: Mean
13. Inter-word silence: Median

14. Total speaking time: Mean
15. Total speaking time: Median

Analysis 1

- Final number of items analyzed = 447 items
- First, pairwise correlations between individual variables and Rasch-based item difficulty

Variable List

Length Variables

Predictor variable
Log number of phonemes
Log duration of item prompt
Log number of words
Log number of syllables
Words per second
Phonemes per second
Syllables per second
Leading silence: Mean
Leading silence: Median
Word duration: Mean
Word duration: Median
Inter-word silence: Mean
Inter-word silence: Median
Total speaking time: Mean
Total speaking time: Median

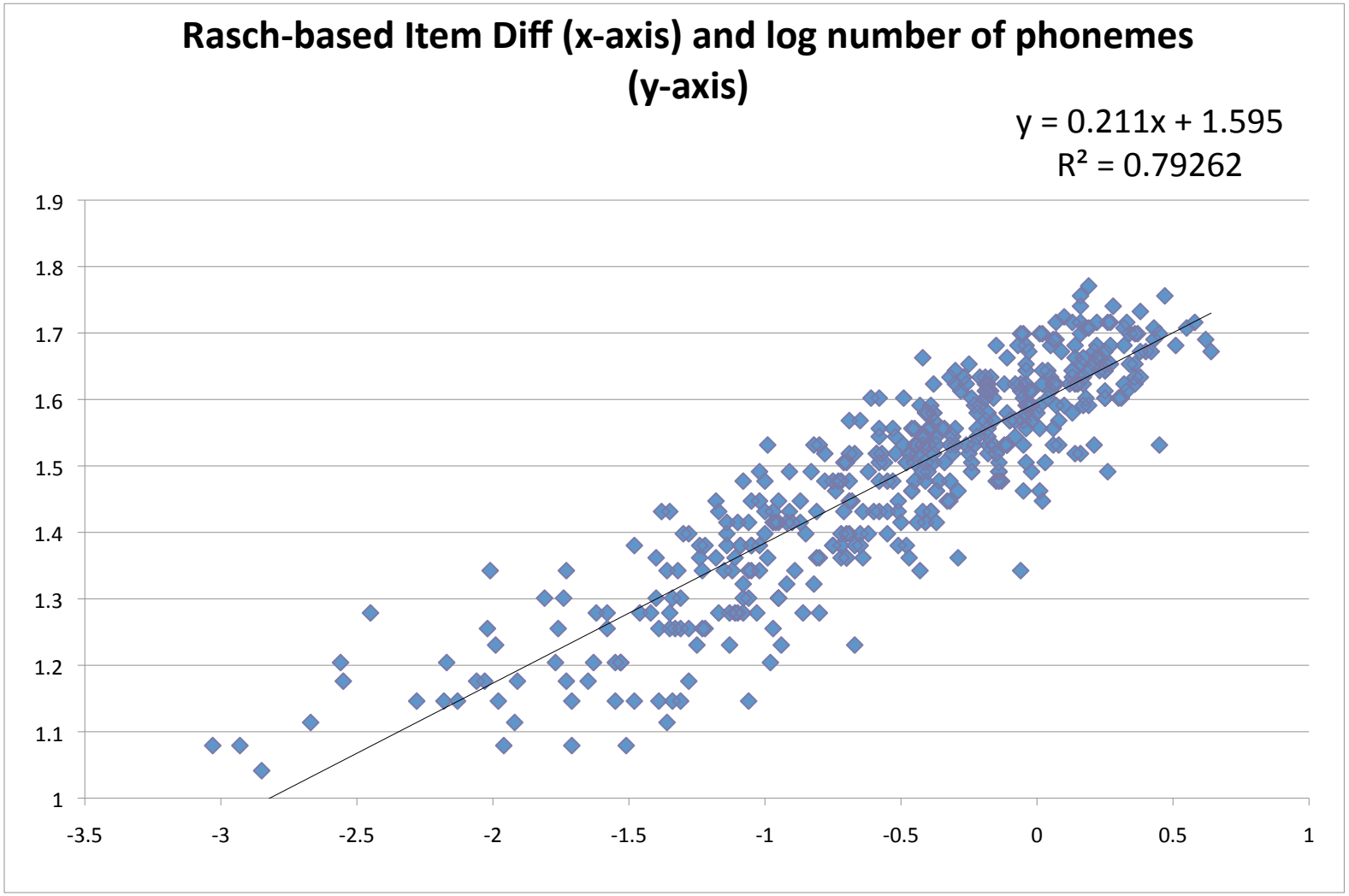
Result - Item Variable: Length

Type	Predictor variable	r
Length	log number of phonemes	0.89**
	log number of syllables	0.89**
	log number of words	0.86**
	Log duration of item prompts	0.82**

** p< 0.01

Result - Item Variable: Length

- All length variables (item properties) were highly correlated with Rasch-based item difficulty (estimated from learner performance)
- Much higher than the English sentence repeat items
  - log number of phonemes: r=0.75
  - log number of words: r=0.69


Rasch-based Item Diff (x-axis) and log number of phonemes (y-axis)
y = 0.211x + 1.595
R² = 0.79262
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1

r = 0.89

N=447

Variable List

Speed Variables

Predictor variable
Log number of phonemes
Log duration of item prompt
Log number of words
Log number of syllables
Words per second
Phonemes per second
Syllables per second
Leading silence: Mean
Leading silence: Median
Word duration: Mean
Word duration: Median
Inter-word silence: Mean
Inter-word silence: Median
Total speaking time: Mean
Total speaking time: Median

Result - Item Variable: Speed

- Computed based on item prompt recordings

Type	Predictor variable	r
Speed	Words per second	0.08*
	Phonemes per second	0.22**
	Syllables per second	0.14**

** p< 0.01, * p < 0.05

Result - Item Variable: Speed

- Phonemes per second correlated the highest among the three speed variables, but only moderately at r= 0.22

- In English, phonemes per second also had a moderate correlation (r=0.39)

Variable List

Predictor variable
Log number of phonemes
Log duration of item prompt
Log number of words
Log number of syllables
Words per second
Phonemes per second
Syllables per second
Leading silence: Mean
Leading silence: Median
Word duration: Mean
Word duration: Median
Inter-word silence: Mean
Inter-word silence: Median
Total speaking time: Mean
Total speaking time: Median

Native performance Variables

Result - Native Performance Variables

Type	Predictor variable	r
Native performance	Word duration: Mean	0.79**
	Word duration: Median	0.79**
	Total speaking time: Mean	0.79**
	Total speaking time: Median	0.78**
	Inter-word silence: Mean	0.15**
	Inter-word silence: Median	0.32**
	Leading silence: Mean	-0.14**
	Leading silence: Median	-0.01

** $p < 0.01$

Result - Native Performance Variables

- Word Duration and Total Speaking Time correlated highly (r = 0.79)
- Natives had to speak longer for longer sentences
- Another version of length variables

Analysis 2

- Multiple Regression Analysis
- Stepwise method in SPSS to find best combination of predictor variables

- Final number of items analyzed = 447 items

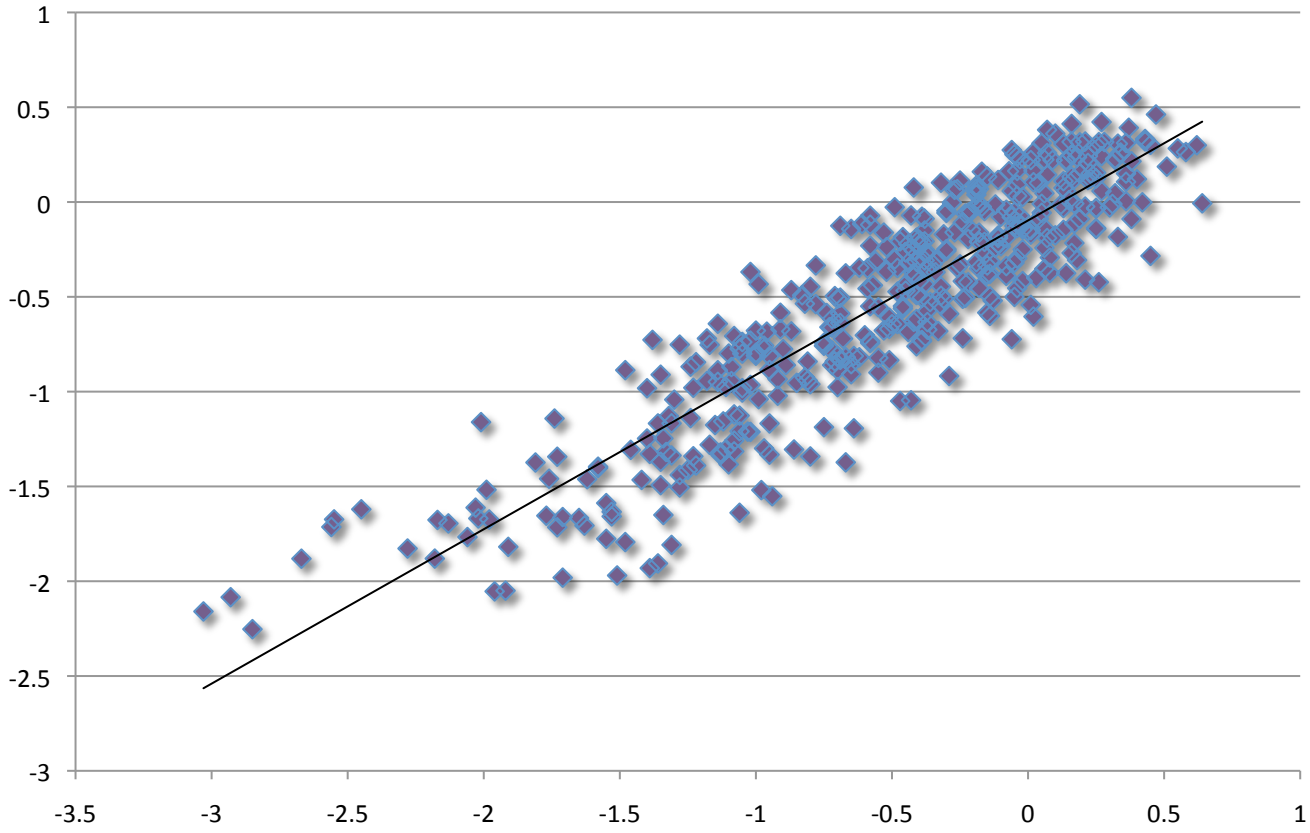
Multiple Regression

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.901[c]	.812	.811	.29188

Model	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	5.877	.187		31.444	.000
log_NPHONEMES	-2.633	.206	-.622	-12.766	.000
log_NWORDS	-1.238	.208	-.290	-5.959	.000
Syllable_per_second	-69.222	22.870	-.063	-3.027	.003

Result 2: Multiple Regression

Rasch Item Diff (x-axix) and Predicted Item Diff (y-axis)



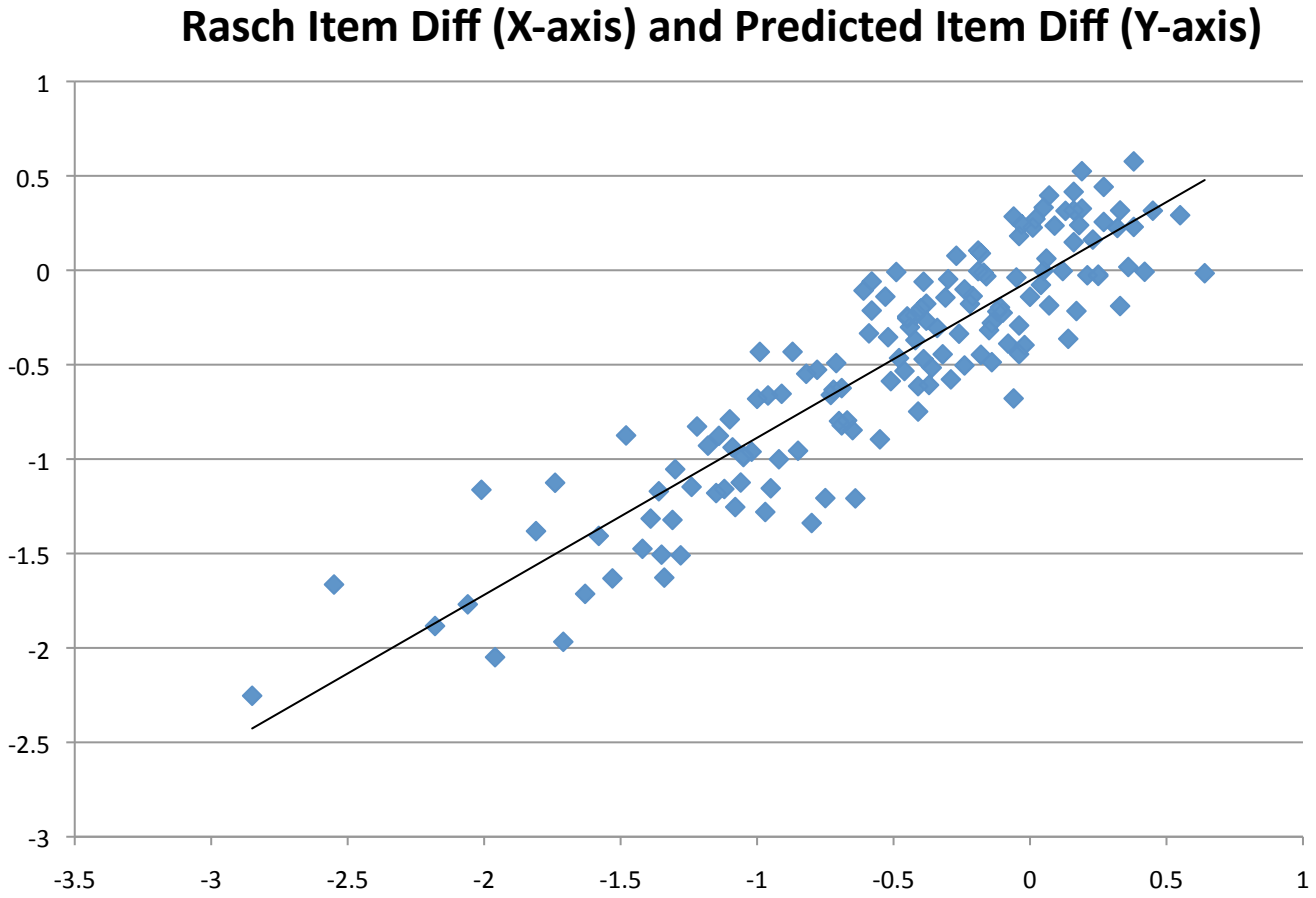
r = 0.90

Adjusted R square = 0.81

N = 447

y = 2.633(log_Nphonemes) + 1.238(log_Nwords) + 69. 222(Syllable per second) -5.877

Generalizability of the Model



N = 149

r = 0.90

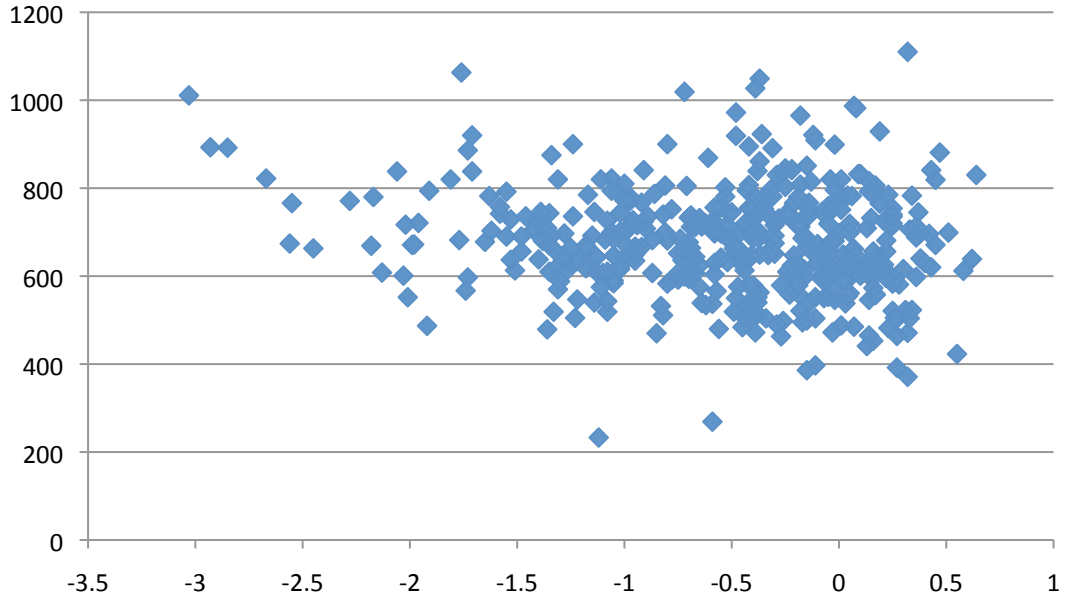
Adjusted R square = 0.81

Summary

- For Chinese, *log number of phonemes* and *log number of syllables* each correlated with Rasch-based item difficulty at 0.89 (79% of the variance)
- Best combination = three variables (*log number of phonemes, log number of words, syllables per second*)
- Regression model with all three variables predicted Rasch-based item difficulty (Rsq = 0.81)
- Regression model seems generalizable for Chinese

Summary (Cont'd)

- *Inter-word silence* or *Leading silence* variables did not correlate well with the item difficulty
- Natives started responding to most items within 1 seconds (min. 0.2 sec, max. 1.1 sec, *SD* = 0.12)



Summary (Cont'd)

- Speed variables (*phonemes per second*, *words per second*) are only moderately correlated both in English and in Chinese

  (English: r = 0.39, Chinese r = 0.22)

Discussion

- Length variables were found to be more highly correlated with Rasch-based item difficulty in Chinese than in English
- For both English and Chinese, *log number of phonemes* appears to be the best single predictor

	English	Chinese
Log number of phonemes	0.75	0.89
Log number of words	0.69	0.86
Log duration of prompt	0.62	0.82
Log number of syllables	0.61	0.89

- Observational, not by design

Possible explanation

- An American (5 syllables, 2words/units, 1 phrase)
- 我要打电话 (5 syllables, 4 words/unit, 1 sentence)

- Same number of syllables, but the number of information units is higher in Chinese
- This may be the reason why length variables and item difficulty are more strongly related in Chinese

Erlam (2009)

"With respect to the memorization issue, research that demonstrates the capacity of working memory is determined by the stored knowledge that already exists in the language, would suggest that those participants who had the ability to memorize stimuli were indeed those who had internalized language , and therefore, that their superior performance on the test was an indication of this." (p.90)

Limitation/Future Research

- Factor out item length, then analyze other variables
  - Can we get a different picture if item length is controlled?
- No structural or content variables were examined
  - Lexical measures (e.g., density, frequency)
  - Syntactic structures

Thank you!

Masanori Suzuki

masanori.suzuki@pearson.com