How Valid are Assessments of Vocabulary Knowledge? Comparing Rasch Vs. Item Facility

Liang Ye Tan, Stuart McLean, and Joseph P. Vitta
Momoyama University, Kindai University, and Kyushu University

Background

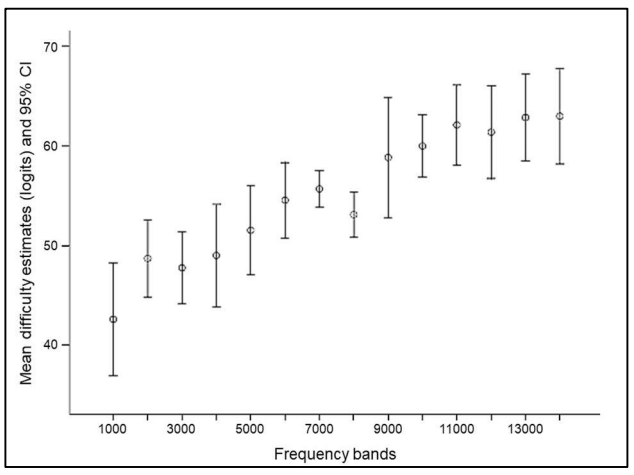
Rasch (or 1-IRT) modelling: Commonly used as a means of providing evidence of validity for vocabulary testing instruments

- E.g., VST (Beglar, 2010); LVLT (McLean et al., 2015); NGSL (Stoeckel et al., 2018); UVLT (Webb et al., 2017)

Premise behind using Rasch modelling:

- Lower frequency = more difficult
  - ⇨ If observations fit Rasch model well, then the test is "valid"
- But what does "valid" mean?
- Kane's (2006, 2013) Interpretation/Use Argument (IUA) framework on validity
- For valid inferences on vocabulary size or levels mastery, knowing a low frequency word extrapolates to knowledge of all other higher frequency words ± reasonable error margin

From Beglar (2010):

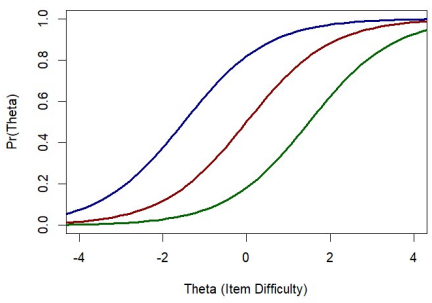


What is the Rasch model?

$$P(x_{ni};\ \beta_n,\ \delta_i) = \exp[x_{ni}(\beta_n - \delta_i)]/\gamma_{ni}$$

where $\gamma_{ni} = 1 + \exp[x_{ni}(\beta_n - \delta_i)]$ is the normalising factor (i.e., reduces the probability function to a probability density function = 1)

Model item characteristic curve (ICC):



- Total score of the individual/item is the sufficient statistic for deriving difficulty estimates
  - ⇨ So, how different are logits from item facility in the current context of vocabulary testing?

Research Question

Is the information provided by Rasch logits different from information provided by item facility in vocabulary tests?

- ⇨ Hypothesis: No significant difference
- ⇨ Analyses: General linear models (GLM) of lexical sophistication

Methodology

Participants & vocabulary test

- 82 Japanese university EFL learners
- Written meaning-recall vocabulary test, conducted in one seating using www.vocableveltest.org
- 150 items, responses typed in Japanese L1
  - 30 items from each of the first five 500-word bands from the NGSL
- Dichotomous rating (1 = correct, 0 = wrong)
- "Word knowledge" = flemma
  - ⇨ Responses with non-target meanings due to polysemy are also marked as correct

Conclusions

(a) In the absence of norm-referencing, Rasch logits = item facility

(b) Rasch logits and item facility produce similar results in GLM if there are no significant ceiling/floor effects

(c) However, using Rasch logits to operationalise difficulty may reduce probability of Type II errors in GLM

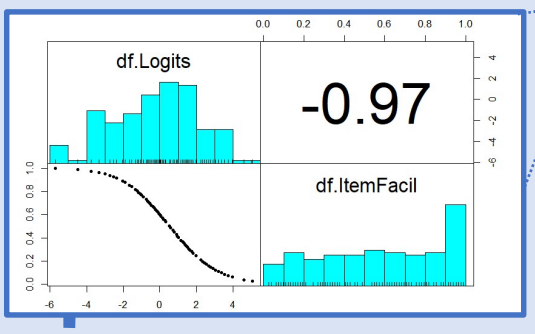
Results

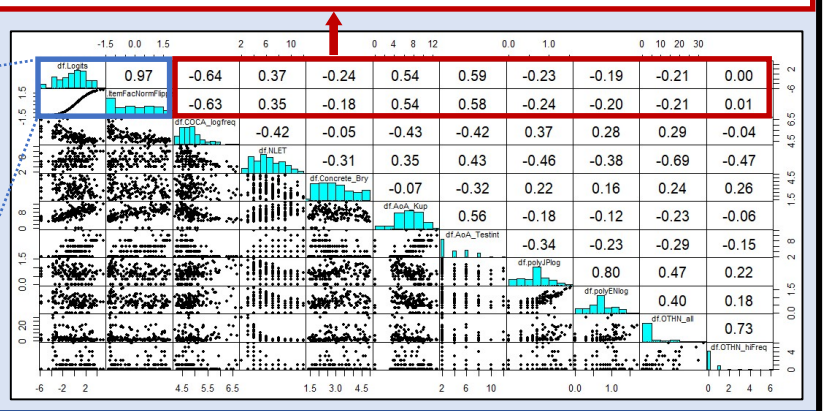
Comparing bivariate correlations between predictors and difficulty estimates in logits vis-à-vis item facility:

Bivariate correlation between difficulty & predictors

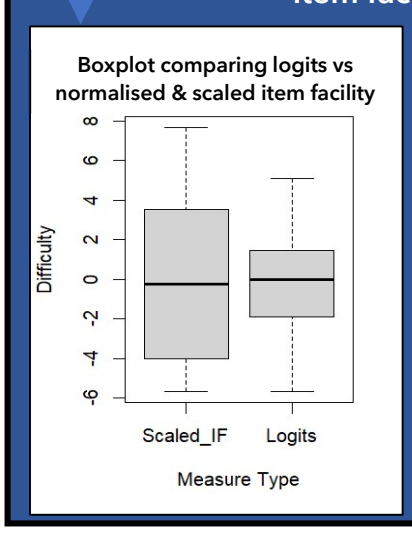
	Difficulty (in Logits)				Difficulty (in Item Facility*)			
Predictors	*r*	*F*	*df*	*p*	*r*	*F*	*df*	*p*
Log_freq	-0.64	100.50	1,148	<.001	-0.63	96.76	1,148	<.001
AoATest	0.59	76.34	1,144	<.001	0.58	73.99	1,144	<.001
AoAKup	0.54	61.64	1,148	<.001	0.54	60.61	1,148	<.001
NLET	0.37	23.66	1,148	<.001	0.35	20.98	1,148	<.001
Concrete_Bry	-0.24	8.67	1,148	.004	-0.18	5.02	1,148	.03
PolyJPlog	-0.23	8.13	1,148	.005	-0.24	8.78	1,148	.004
OTHN_all	-0.21	6.54	1,148	.01	-0.21	7.09	1,148	.01
PolyENlog	-0.19	5.28	1,142	.02	-0.20	5.99	1,142	.02
OTHN_hi	0.00	0.00	1,148	.99	0.01	0.03	1,148	.87

*Note: Difficulty measure in item facility has been normalised and sign flipped to facilitate direct comparison.





Comparing regression models with difficulty estimates in logits vs item facility:



RaschPars Model:

Difficulty = 9.50 – (2.15*Log_freq) – (0.43*Concrete_Bry) + (0.18*AoA_Kup) + (0.20*AoA_Test) + error

Variables	*b*	β	*SE*	*t*	*p*	*lmg*	Pratt
(Intercept)	9.50	-	1.86	5.11	<.001	-	-
Log_freq	-2.15	-0.46	0.30	-7.17	<.001	.25	.29
AoA_Test	0.20	0.22	0.07	3.02	.003	.16	.13
AoA_Kup	0.18	0.20	0.06	2.97	.004	.13	.11
Concrete_Bry	-0.43	-0.18	0.14	-2.96	.004	.04	.04

F[4,141] = 48.4, p < .001, multiple R^2 = .58, adjusted R^2 = .57

IFPars Model:

Difficulty = 3.01 - (0.81*Log_freq) + (0.07*AoA_Kup) + (0.11*AoA_Test) + error

Variables	*b*	β	*SE*	*t*	*p*	*lmg*	Pratt
(Intercept)	3.01	-	0.72	4.17	<.001	-	-
Log_freq	-0.81	-0.42	0.13	-6.40	<.001	.24	.26
AoA_Test	0.11	0.30	0.03	4.19	<.001	.17	.17
AoA_Kup	0.07	0.20	0.03	2.74	.007	.14	.11

F[3,142] = 56.2, p < .001, multiple R^2 = .54, adjusted R^2 = .53

Discussion

IRT paradigm vs CTT paradigm

- IRT: Item-invariance allows use of test items to place individuals on an ability continuum
  - ⇨ This is based on norm-referencing
- CTT: Observed score = True score + Error
  - ⇨ Relative ability/difficulty is based on references to other groups within the sample pool
- De Wilde (2023): Appropriate use of test-interpretation through referencing age-norms

References

1. Andrich, D. (1988). Rasch Models for Measurement. SAGE Publications, Inc.
2. Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101–118.
3. De Wilde, V. (2023). The auditory picture vocabulary test for English L2: A spoken receptive meaning-recognition test intended for Dutch-speaking L2 learners of English. *Language Teaching Research*, 1-31.
4. McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741–760.
5. Stoeckel, T., Bennett, P, & Ishii, T. (2018). A Japanese-English bilingual version of the New General Service List test. *JALT Journal*, *40*(1), 5.
6. Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics*, *168*(1), 33–69.
7. Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64).
8. Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores: Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73.