Psychometrics for educational assessment: The Test de connaissance du français®

Christophe Lalanne, Marianne Mavel, Pascal Bessonneau, Marina Esposito-Farèse

Centre international d'études pédagogiques
Département évaluation et certifications
Sèvres, France, www.ciep.fr
✉ lalanne@ciep.fr

Background

The Test de connaissance du français

The TCF is administered by the CIEP and is a recognized certification of the French Ministry of Education. It is composed of:

- 3 compulsory parts: Listening, Grammar and Reading (80 multiple choice items of varying difficulty and content)
- 2 additional parts: Speaking and Writing (oral interaction and small or more elaborated constructed responses on selected texts)

How do psychometry contribute to assessment quality?

All items are pretested on a representative sample of candidates, before being calibrated to allow for score delivering (Fig. d).
This process obeys standard guidelines for test construction and management.
Scores are equated, thus ensuring their equivalence between the different test forms that are delivered to our different passation centers (Fig. a).
This is the basis for quality control and fairness.

Methodology

Classical Test Theory (CTT)

CTT[2] allows to study items functionning, as well as scores and test reliability.

- Observed *P*-value (proportion of responses correct) relates to item difficulty (Fig. b and f).
- Item-test correlation (key and distractors) reflects the discriminative power of each item (Fig. b and g).
- Cronbach alpha provides a means to assess internal consistency.

$$x_i = \tau_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Item Response Theory (IRT)

IRT provides both a theoretical framework[3,4] and new technical innovations for studying subjects' behavior faced with items of varying difficulty.

- Item difficulty and person ability can be estimated along with additional parameters related to test content and/or individuals characteristics.
- Measurement model adequation or statistical fit can be tested using classical statistics (mostly χ^2distributed[5], Fig. c and i).
- Standard error of measurement (SEM Fig. h)) indicates the degree to which we can be conident in estimated person scores.

(1 PL Rasch Model)

$$\mathrm{P}(X_{vi} = x) = \frac{\exp(x(\theta_v - \beta_i))}{1 + \exp(\theta_v - \beta_i)}, \quad x \in \{0, 1\}$$

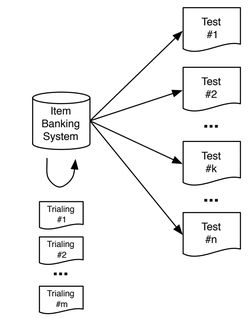
$$\log \frac{\mathrm{P}(X_{vi} = 1)}{\mathrm{P}(X_{vi} = 0)} = \theta_v - \beta_i$$

The probability of answering correctly depends on the distance between subject's ability and item difficulty

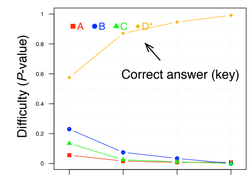
$$I(\theta) = \sum_{i=1}^{k} I_i(\theta; b_i)$$

Each item contributes equally to the total Test Information (Fig. h)

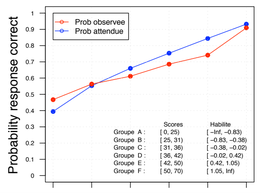
a. Test construction: Trialing and Item Banking System.



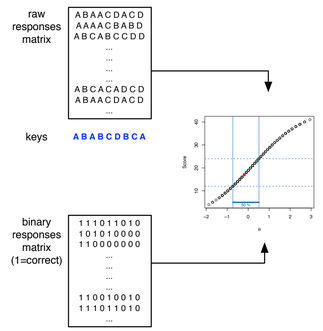
b. Distribution of reponses for a given item (4 alternative forced choice).



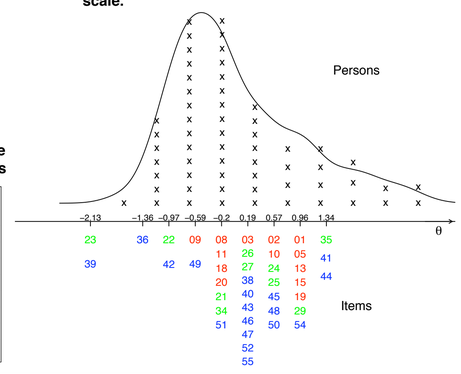
c. Expected and observed response probability for increasing abilities



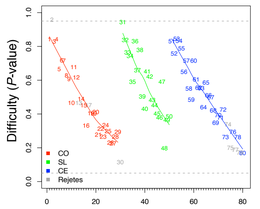
d. The Equating Procedure: Converting raw scores to calibrated scores.



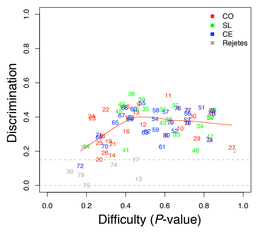
e. Item-Person Map: Localization of items and subjects on the (calibrated) measurement scale.



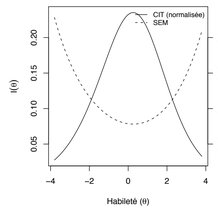
f. Items difficulty as a function of rank.



g. Items difficulty and discrimination.



h. Test information and Standard Error of measurement are invertely related.



The validation stage

Several filters are applied to the data, in an iterative manner: Inclusion criteria for subjects and items (3 stages).

- Remove subjects too young or those who do not provide demographic information
- Remove subjects with more than 5 non-response at the beginning of the test and flag those who do not complete the whole test ①

- Remove items with too much missing values ②
- Remove items that are too easy or too difficult (.5 < *P*-value < .95)
- Remove items likely to be 'unbalanced' (one or more distractor(s) never choosed)
- Remove items that are not discriminative enough (point-biserial coefficient < .15, responses to distractor positively correlated to total score)
- Remove items that weaken Cronbach α

- Remove items not well represented on the hypothesized unidimensional construct (first principal component)
- Remove items that misfit the Rasch Model ((un)standardized INFIT and OUTFIT values) *if any, Model is fitted again*
- Remove items that are biased toward one category of the sample (age, sex, mother tongue) *if any, Model is fitted again* ③

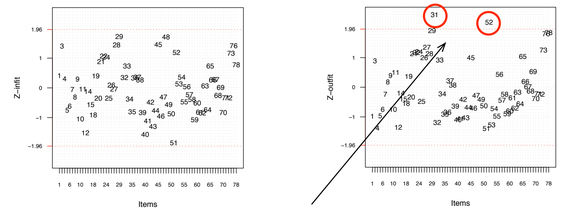
Exclusion rate: 25-35% (80 items, 500 subjects min.)

Reliability and items fit

CTT criteria give us with items which potentially allow to locate a given person on a calibrated scale of measurement. We aim at giving a *reliable* score (i.e. with minimum error) when estimating a person's proficiency level, while ensuring the *validity* of the measurement instrument.

The Rasch Model[6] is not only a statistical model, but also a measurement device assuming local (stochastic) independance, unidimensionality and a monotone relationship between the probability of correctly answering an item and the individual's ability.

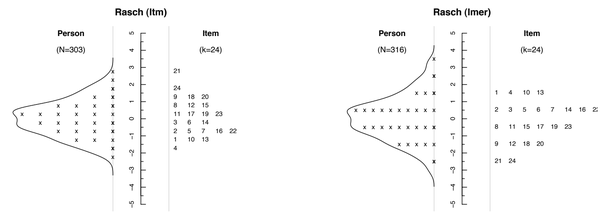
i. Standardized residuals (Z-INFIT and Z-OUTFIT) for the Rasch Model.



Person are more likely to give unexpected answer

Marginal Likelihood Estimation can be used to uncover model parameters.[7] This allows for a comparison between purely IRT software and a more 'classical' approach standing on Mixed-Models[8](GLMM, Fig. j, right panel).

j. Estimated parameters using Rasch Model (MML) and GLMM.



Simulations done with the R Open Source statistical package.
Benchmarking with the SAS System and other dedicated IRT software (e.g. WinSteps, Conquest) leads to comparable conclusions.

Ongoing Projects

- Analyzing Differential Item Functionning with a GLMM and comparison with usual approaches (SIBTEST, Mantel-Haenszel (with/whth standardization))
- Testing for the lack of fit using various statistics[5]
- Reliability study for constructed response marking using Balanced Incomplete Block Design
- Computerized adaptive testing: Mixing both IRT and NL binary programming

References

[1] More information available on www.ciep.fr/tcf

[2] Nunnaly, J.C. and Bernstein, I. (1994). *Psychometric Theory* (3rd Ed.). McGraw-Hill.

[3] Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications.* Boston: Kluwer-Nijhoff Publishing.

[4] Van der Linden, W.J. and Hambleton, R.K. (1997). *Handbook of modern item response theory.* Springer.

[5] Flieller, A. (1994). Méthodes d'étude de l'adéquation au modèle logistique à 1 paramètre. *Mathématiques et Sciences Humaines*, *127*, 19-47.

[6] Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980). Chicago: The University of Chicago Press.

[7] Baker, F.B. and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques.* CRC Press.

[8] De Boeck, P. and Wilson, M. (2004). *Explanatory Item Response Models: a Generalized Linear and Nonlinear Approach.* Springer.