An Item Response Theory (IRT) approach to check correspondence between cut-off scores and maximal test information in French pilot selection

Nadine Matton, Michel Veldhuis, Stéphane Vautier

17 An Item Response Theory (IRT) Approach to check Correspondence between cut-off Scores and maximal Test Information in French Pilot Selection

N. Matton[1]; M. Veldhuis[2] and S. Vautier[2]*

[1]ENAC, France; [2]University of Toulouse, France
*Corresponding author. E-mail: nadine.matton@enac.fr

Introduction

Cognitive ability test scores are widely used in selection procedures in aviation for hiring pilot trainees or ATC trainees (e.g., Damos, 1996; Burke, Hobson & Linsky, 1997; Martinussen & Torjussen, 1998; Sommer, Olbrich & Arendasy, 2004; Matton, Vautier & Raufaste, 2009). In Europe, the underlying theory implicitly used in this context is the Classical Test Theory (CTT, Gulliksen, 1950; Lord & Novick, 1968). Following this theory, the observed score variable (*Y*), usually the sum of elementary scores for each item, is construed as the sum of a true score variable (*T*) and an error variable (*E*), $Y = T + E$. In CTT, measurement precision is generally assessed through *reliability* indexes. Considering scores of a group of participants, the reliability of a test is defined as the proportion of true variance (var(*T*)) on observed variance (var(*Y*)). Reliability cannot be computed directly (as *T* is a latent variable) and can only be estimated under some hypotheses (e.g., when two tests are supposed to be parallel[1], reliability can be computed as the correlation between both score variables). Moreover, in CTT, reliability is assumed to be constant whatever the score level.

In a more modern psychometric theory, item response theory (IRT, Rasch, 1960; Birnbaum, 1968), the focus is on the response on each item instead of on the test. Furthermore, the measurement precision is assessed by the level of *information* that is provided by each item, with the idea that the degree of information depends on the level of the respondent's *ability*, defined as the latent psychological dimension assessed by the test. The key idea in IRT is that the probability of success of an item depends on the level of ability of the respondent. Generally IRT models assume an S-shaped relationship (see Figure 1, left panel) depending on one, two or three parameters being characteristics of the item (e.g., difficulty or discrimination parameters). Classically, the difficulty corresponds to the location of the inflexion point of the curve (the more this point is on the left of the theta axis, the easier the item) and the discrimination corresponds to the steepness of the curve at the inflexion point (the steeper the curve, the more discriminant the item). The information given by an item is defined as the precision of measurement of the estimated ability and depends on the item parameters as well as on the level of ability (see Figure 1, right panel). It is also inversely related to the standard error of the ability level estimation.

[1] Two test scores, Y_1 and Y_2, are parallel if and only if $T_1=T_2$, $\text{var}(E_1)=\text{var}(E_2)$, and $\text{cov}(E_1, E_2.)=0$.

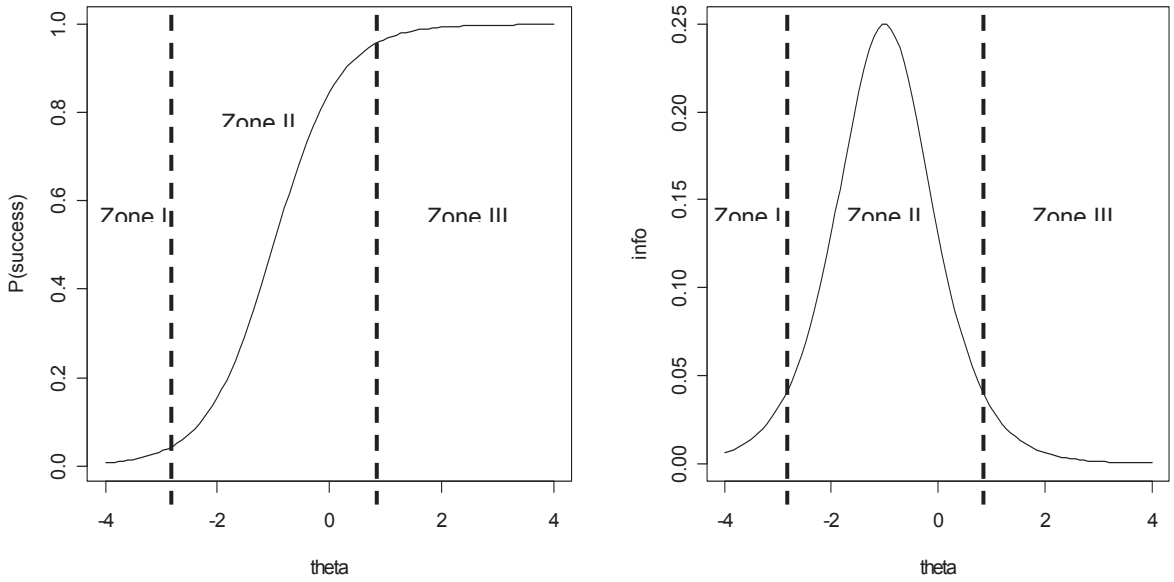

***Figure 1*. An example of item characteristic curve showing the relationship between the ability (theta) and the probability of success of a given item (left panel). An example of item information curve showing the relationship between the ability (theta) and the level of information of a given item (right panel).**

Pragmatically, three ability-zones are of interest:

- low-ability respondents (zone I) are very likely to fail the item. This item would not discriminate among them. Correspondingly the amount of information given by the item is relatively small for this subgroup and the ability level would not be precisely estimated.
- for medium-ability respondents (zone II) the probability of success depends closely on the ability level, thus the item would discriminate among them and the amount of information provided by the item in this ability range is large. Consequently the precision of estimation of the ability level would be high.
- high-ability respondents (zone III) are very likely to succeed the item. Therefore the amount of information provided by the item is low for these examinees.

Moreover, each item will be characterised by a different item information curve, with a different location of the peak of information. By summing the different item curves, we obtain the test information curve that highlights the ability level that is the most precisely assessed by the test.
In a context of selection it seems interesting to estimate the level of ability that is most precisely assessed by a test and to compare this level to the cut-off score that is used and possibly (iii) to optimise the test by selecting items that are more informative is the cut-off zone or creating new items. In the present study we analysed data from ten cognitive ability tests used in the French national pilot trainee selection (ENAC).

Method

Population. The 577 applicants that took the ten cognitive ability tests are the population of interest in this study. They were all applicants for the yearly selection (2009) for entry in the French national pilot training school (male n = 86,7%, median age = 20 (18-31)).

Material. All participants took the same test battery, which comprised the following ten cognitive ability tests: three abstract reasoning tests (RS1, RS2 and RS3), three verbal ability tests (VER1, VER2 and VER3), two spatial reasoning tests (SPA1 and SPA2), a mechanical comprehension test (MEC) and a numerical ability test (NUM).

Model. As all tests used in this selection procedure consist of multiple choice questions, we expected the three parameter logistic (3PL) model to fit the data. The 3PL model has in addition to the discrimination parameter and the difficulty parameter, a so-called pseudochance parameter. This pseudochance parameter takes into account the fact that test takers can find the answer by guessing. We fitted this model with the program MULTILOG (Thissen, 1991; 2001) with the following restrictions; (1) all item discrimination parameters were constrained to be equal in order to conserve a unique ordering of the items on the ability scale (by doing this the ICC's cannot cross each other, thus the ordering of the items does not depend on the ability), and (2) the pseudochance parameter was set to be equal to the probability of guessing the right answer on an item. These restrictions qualify this model as a model of the Rasch-family (1960) which permits a direct comparison between expected and observed scores.

Results and Discussion

Fit analysis. There are multiple fit indexes that have been proposed in the literature. We chose one commonly used IRT fit statistic in organizational settings (see Table 1), the adjusted $\chi 2$ to degrees of freedom (adj. $\chi 2$/df) ratio test (Drasgow, Levine, Tsien, Williams and Mead, 1995). Given the threshold for acceptable fit of 3.0, the model fits to six of the ten tests: MEC, SPA2, RS1, RS3, VER1, VER2. Scores of tests SPA1, RS2, NUM and VER3 and are not well adjusted to the 3PL model with constraints.

Table 1. Fit indexes of the 3PL model for the ten cognitive ability tests.

Test	adj. $\chi 2$/df
MEC	1.1
SPA1	48.7
SPA2	1.2
RS1	0.2
RS2	86.6
RS3	2.5
NUM	38.2
VER1	0.5
VER2	0.7
VER3	21.0

Test Information Curves. Figure 2 displays the test information curves for the six tests that fitted correctly to the IRT model. All peaks of information are located in the zone of negative thetas. As the ability theta are normalized, negative thetas correspond to thetas below the mean, i.e. to (relatively) low-ability applicants.

At this selection stage, the current ENAC policy is to eliminate applicants that have scores under *stanine 3* for at least one psychological dimension. Each dimension is assessed by an aggregation of several tests scores, therefore it is interesting to evaluate the level of information given by each test in the cut-off zone

between *stanine 3* and *stanine 4*, i.e. around *theta* = -0.75 (estimated IRT *theta* are usually very close to standardised scores).

SPA2, RS3 and MEC are well calibrated tests for the ENAC policy. On the contrary, RS2, VER1 and especially VER2 are mostly informative in the very-low-ability zone. These three last tests could be optimised by eliminating items whose peak of information is located in very low *thetas*, and creating new items that would have their peak of information around *theta* = -0.75. One practical way of creating such new items could be to study the characteristics of the items whose information level is maximum around *theta* $\in$ [-1.0;-0.5] and design analogous items.

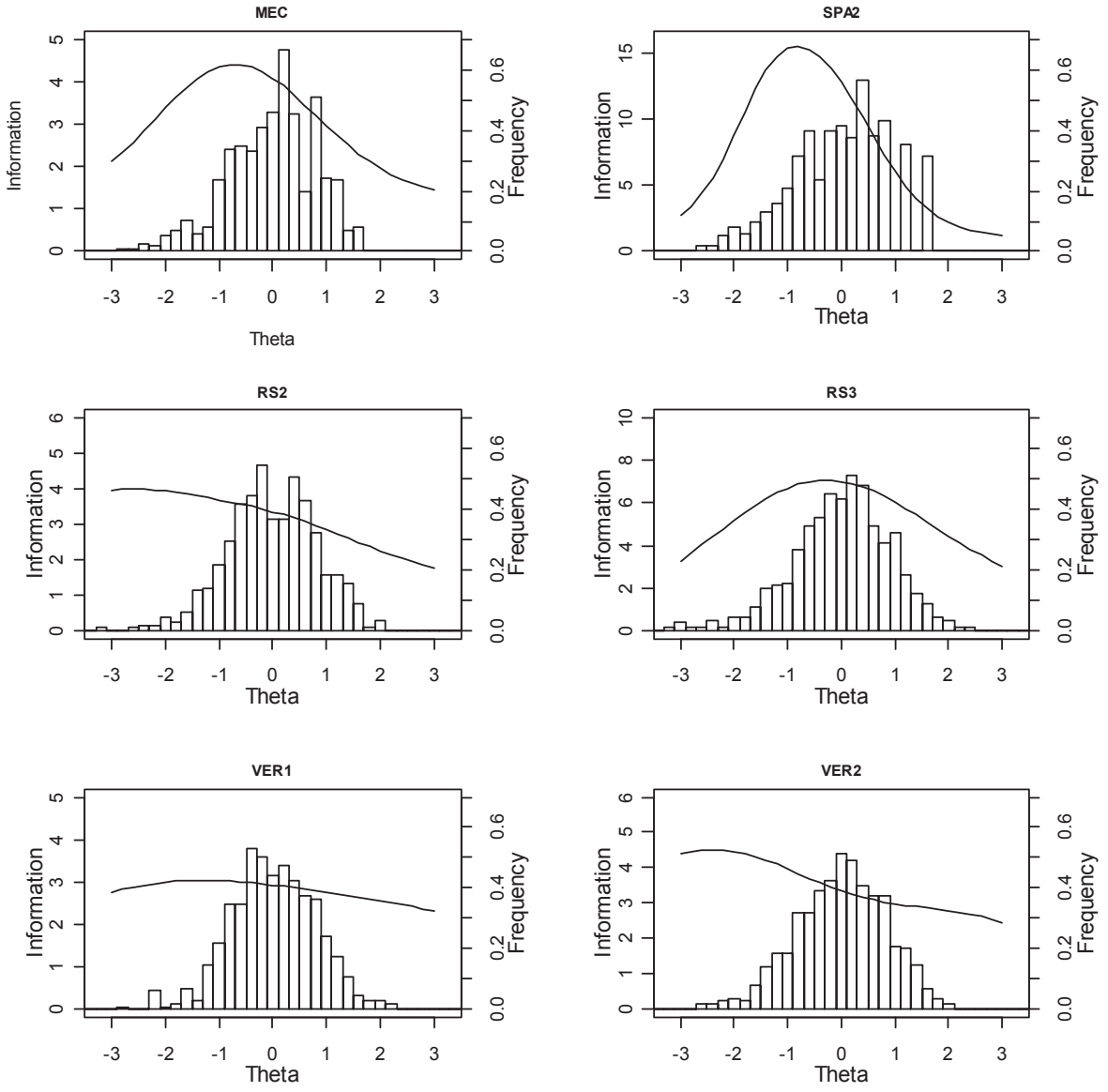

***Figure 2*. Information test curves and histogram of theta distribution for each of the six cognitive ability tests that were correctly fitted by our model.**

This study illustrates a possible use of IRT modelling in a selection setting. One of the most widely used applications of IRT in ability testing is the use of item banks and adaptive testing, which has been proven to

be efficient and cost-effective. In the case of a national exam, as is the case at the ENAC, where all applicants have to take rigorously the same tests with the same items, instead of adapting the testing material to the applicants, an alternative is to adapt the testing material to the selection policy.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental tests scores*. Reading, MA: Addison-Wesley.

Burke, E., Hobson, C., & Linsky, C. (1997). Large Sample Validations of Three General Predictors of Pilot Training Success. *International Journal of Aviation Psychology, 7*(3), 225.

Damos, D. L. (1996). Pilot Selection Batteries: Shortcomings and Perspectives. *International Journal of Aviation Psychology, 6*(2), 199--209.

Drasgow, F, Levine, M.V., Tsien, S., Williams, B.A., & Mead, A.D. (1995). Fitting

polytomous item response theory models to multiple choice tests. *Applied Psychological Measurement, 19*, 143-165.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental tests scores*. Reading, MA: Addison-Wesley.

Martinussen, M., & Torjussen, T. (1998). Pilot Selection in the Norwegian Air Force: A Validation and Meta-Analysis of the Test Battery. *International Journal of Aviation Psychology, 8*(1), 33.

Matton, N., Vautier, S., & Raufaste, E. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence*, 37, 412–421.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogische Institut.

Sommer, M., Olbrich, A., & Arendasy, M. (2004). Improvements in personnel selection with neural networks: a pilot study in the field of aviation psychology. *International Journal of Aviation Psychology, 14*, 103--115.

Thissen, D. (1991, 2001). MULTILOG 7.03 user's guide: Multiple categorical item analysis and test scoring using item response theory. [Computer software and manual]. Chicago: Scientific Software International.