Using IRT to evaluate measurement precision of selection tests at the french pilot training

Michel Veldhuis, Nadine Matton, Stéphane Vautier

Running head: USING IRT IN THE FRENCH PILOT SELECTION

Using IRT to Evaluate Measurement Precision of Selection Tests at the French Pilot Training

Michiel Veldhuis[1]

Nadine Matton[2]

Stéphane Vautier[1]

[1]Université de Toulouse, France

[2]École Nationale d'Aviation Civile, France

Correspondence should be addressed to: Stéphane Vautier, Centre d'études et de recherche en psychopathologie, Maison de la Recherche, 5 allées A. Machado, 31058 Toulouse Cedex 9, France (e-mail: vautier@univ-tlse2.fr).

Abstract

In pilot selection settings decisions are often based on cut-off scores. In item response theory the measurement precision of a test score can be evaluated by its degree of information. We investigated whether the maximum of test information corresponded to the cut-off zone for ten cognitive ability tests of the current French civil air transport pilot selection (n = 577). An item response theory model was fitted to the data. The six fitting tests' test information curves were generally maximal in the cut-off score zone. The absolute level of precision could nevertheless be improved for some tests.

Keywords: item response theory (IRT); pilot selection; cognitive ability tests; test information; cut-off score; three parameter logistic (3PL)

The selection for *ab initio* pilot training is known as one of the most difficult selections to pass for students. Indeed several civil pilot training schools have rejection rates of about 95% of the candidates. All these schools use test batteries with cognitive, personality, specific skills tests followed by interviews to select the students. The reliability and validity of these tests are often investigated as they have to be very efficient in order to make a sound decision about acceptance or rejection of the future pilot. In selection situations mistakes are always just around the bend. The tests used in selection for aviation training in Europe have usually been investigated within the framework of Classical Test Theory (CTT; e.g. Damos, 1996; Burke, Hobson & Linsky, 1997; Martinussen & Torjussen, 1998; Sommer, Olbrich & Arendasy, 2004; Matton, Vautier & Raufaste, 2009; in press). CTT is the theory that introduced the notions of the latent/unobservable true score (T), error (E), and the observable composite test score (X) with the simple linear model $X = T + E$. This model exists since the 1920's and is still very often used in psychological research. However since the 1960's there exists another modern psychometric theory, based on the analysis at the item level, namely Item Response Theory (IRT).

IRT comprises a collection of modeling techniques for the analysis of item level data obtained to measure inter-individual variation (Orlando & Reeve, 2007). The relationship between item performance and ability is the main focus in IRT, where the ability refers to a unidimensional latent trait that characterizes each subject regarding the psychological dimension assessed by the test. This latent trait ability is generally noted as theta (θ) and depicts every subject's ability level with a mean of 0 and a standard deviation of 1 (Schlessman, 2009; Embretson & Reise, 2000). Logistic IRT models have the particularity that the probability of having a correct response on an item is a logistic function of the "ability" level. This function is called the item characteristic curve (ICC) and specifies for every item how the probability of a correct response varies over the ability scale. In most IRT models the ICC is characterized by parameters for the difficulty and the discrimination of each item. If an item is efficient at measuring a given level of ability the curve will be relatively steep around this level, which means that it discriminates well between test takers of different ability levels. *Item discrimination* depicts the steepness of the slope of the ICC. *Item difficulty* represents the location of the centre of the ICC, i.e. the point of infliction where the probability of a correct response generally equals 50%. These parameters are defined in relation to the latent ability scale and are theoretically independent of the group ability distribution. Using these ICC's one can estimate the abilities of the examinees through a maximum likelihood procedure. For each ability estimate, the variance of the estimator can be computed. By definition, *information* at an ability level is the reciprocal of this variance (Baker, 2001). The more precision in estimation, the more information about the estimated quantity is obtained. In IRT the amount of information that an item provides on differences in ability is provided by the steepness of the slope of the ICC at an ability level. This *item information* is maximal at the ability level that corresponds to the item difficulty. As every item has a different difficulty the item information function is also different for every item. By summing the item information for all items in a test one obtains the *test information*. Test information shows how the measurement precision of the test differs over the ability scale. In CTT such an approach does not exist, the next best thing to it would be the test's reliability that only gives an overall estimate of the measurement precision and that is supposed to be constant over the whole ability scale.

In short, IRT models explicitly posit how the item responses depend on the measured ability, whereas CTT uses primarily the composite score and not the relation between item response and ability. It has been argued that IRT has many advantages over CTT (e.g., Embretson & Reise, 2000; or Hambleton, Swaminathan & Rogers, 1991). In CTT the composite score of all items is the main focus, whereas IRT specifically singles out the item scores. CTT provides true scores, *p*-values (difficulty) and item-test correlations (discrimination) that are dependent on the test and examinees, which limits the usefulness of these statistics. On the contrary in IRT the person and item parameters are not test or group dependent. These parameters permit the researcher to calculate a latent trait ability estimate unlinked to a specific test.

The purpose of this study was to examine the psychometric properties of ten tests used in the selection for aviation training at the École Nationale d'Aviation Civile (ENAC) the French national pilot training in Toulouse, France. The ENAC receives about 2000 candidates each year, of which around 700 take the psychological test battery, of which around 200 pass an interview and of which circa 50 are in the end accepted in the pilot training program. Within the framework of CTT some of these tests have already been examined in detail (e.g. Matton, Vautier & Raufaste, 2009), but an IRT approach has as of yet not been used. In order to accomplish the selection of the best future pilots, as in the fact that they are without any weaknesses, the ENAC eliminates those candidates that lag significantly behind on at least one psychological dimension, based on unpublished predictive validity studies. A way to find out if the tests' item characteristics are in concurrence with this strategy is by verifying if the most information is obtained at the ability level that would correspond to the cut-off zone. A three parameter IRT model (3PL) will be fitted to the data, the estimated parameters can be used to establish if the maximal information of a test is indeed at the cut-off zone ability.

Method

Participants

The data used in this study come from the actual selection process at the French national pilot training school (ENAC) of the 2009 session (male n = 500, female n = 77, median age = 20 (18-31)). After a first selection round with three tests (English language, mathematics and physics) 577 of the 2100 candidates remained. These 577 candidates that continued to the following step of ten cognitive ability tests are the population of interest in this study. Following the ten cognitive tests 232 candidates were selected for an interview, a group exercise and an English expression test. In the end 60 potential pilots (~3%) were selected for the training: 48 for *ab initio* training and 12 for flight training.

Materials

All participants underwent the same test battery of cognitive ability tests on a single day. The tests can be classified by the type of ability that is needed to accomplish the task. The data of the ten multiple choice tests that measured space relations, logical reasoning, arithmetics, verbal comprehension are used in this study. These same abilities are esteemed to be of great importance in the training to become and the future work as a pilot (e.g. Martinussen & Torjussen, 1998; Goeters, Maschke & Eißfeldt, 2004) and are used in most pilot selection procedures.

For reasons of confidentiality, the tests used in the selection procedure are not fully described in the present paper. On request, the corresponding author will answer any questions regarding these tests. The ten multiple choice tests consisted of three spatial ability tests (SPA1, SPA2 and SPA3), three logical reasoning tests (RS1, RS2 and RS3), three verbal ability tests (VER1, VER2 and VER3) and one 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 (Birnbaum, 1968) to fit the data. The 3PL model has in addition to the *discrimination* parameter and the *difficulty* parameter, a so-called *pseudo-chance* parameter. This pseudo-chance parameter takes into account the fact that test takers can find the correct answer by guessing. We fitted this model with the program MULTILOG 7.03 (Thissen, 1991; 2001). This program uses the MML (Marginal Maximum Likelihood) method to estimate the parameters. To fit the model we used the following restrictions; all item discrimination parameters were constrained to be equal in order to conserve a unique ordering of the items on the ability scale, and the pseudo-chance parameter was set to be equal to the probability of guessing the right answer on an item, i.e. 1/k, where k is the number of possible answers. Setting the discrimination parameters to be equal has as a consequence that the ICC's cannot cross each other. If this parameter would have been left free, the ICC's could have crossed and lead to interpretation issues: the ordering of the difficulty of the items would change depending on the ability of a person. For example item A could be easier than B for someone with a 'low' ability whereas for someone with a 'high' ability the opposite would occur. On a test with items that measure a single ability, this would be difficult to comprehend.

In Europe psychologists rarely use the three parameter logistic model, whereas in the U.S. it is the most used IRT model. This cross-oceanic difference comes from the fact that 3PL has been developed in the U.S. when in Europe Rasch's model (1960) was in use. European psychometricians have found that the Rasch model has several theoretical and mathematical advantages over the 3PL, which caused the use of 3PL, notwithstanding its practical merit, in Europe to be extremely limited. The 3PL model has for example been found to be unidentified if one does not add the hypothesis that the ability is normally distributed in the population (Maris, 2002). Nonetheless we chose to use this model as the model is effectively identified and thus useful when the hypothesis that the ability is normally distributed is added. This modus operandi would not pass the scrutinizing test of a pure psychometrician as a hypothesis is added that cannot be tested (Verhelst, personal communication), nonetheless it's fairly common to suppose that cognitive ability is normally distributed in large populations (as is the case in this study, n = 577). For practical reasons the 3PL is nonetheless used in this study.

As an IRT model is merely a model of the data, it is important to assess how well fit it fits the data. Every model is essentially an oversimplification of reality and as such does not fit the data completely. Several different measures of goodness of fit exist of which most are χ^2 distributed, there is no real consensus in the psychometric community about which measure to use. We used the ratio of the adjusted χ^2 to the degrees of freedom (Drasgow, Levine, Tsien, Williams & Mead, 1995). When this ratio exceeds 3.0 the model is considered to misfit the data (Schlessmann, 2009), that is, useless as a means to interpret how the data was generated..

Analyses

Missing responses were treated as failed. After fitting the model and estimating the item parameters, the item information curves were calculated. In summing the item information curves for each item one obtains the test information curve. The maximum of the test information curve was compared to the region of the cut-off score. The students that fell in the third stanine or lower for any dimension of the tests were eliminated from the process, from the fourth onward the students were selected for the ensuing selection procedure. The corresponding standard (z) cut-off score would be -0,75. As the ability scale in the 3PL model is set to be standard normally distributed the location of the maximum of the test information function on the ability scale should ideally fall in the region of theta (θ) = -0,75. This would mean that a test measures the ability most precisely in the cut-off zone. If the maximum of a test is between θ = -1.5 and θ = -0.5 and the degree of information is relatively high, we consider the test to be precise enough to identify the least performant candidates. If the test's maximal information point does not coincide with the cut-off point, but the degree of information is still high at this point, the test is also considered good enough. If both the maximum and the degree of information are non-satisfactory the test is not sufficiently precise.

Results

Given the threshold of 3.0 for the adjusted χ^2, the fit of six of the ten tests was satisfactory (Table 1).

The unsatisfactory fit of the four models (SPA1, RS1, NUM, VER3) could be due to the imposed time limit on the tests. These tests are relatively long and become more difficult towards the end, causing the candidates to fail to respond to the last few items. As a consequence the observed probability to give the right answer on the last items approaches zero, whereas the minimal probability in the model is 0.125 (the *pseudo-chance* parameter/ lower asymptote). The differences between these proportions being the basis for the adj.χ^2/df measure explains the unsatisfactory fit (Drasgow et al., 1995). Indeed the proportion of mean missing responses is for all fitting tests less than 10 % (except RS3) and for all non-fitting tests at least 22 % (see Table 2).

This indicates that more than a fifth of the items were left without response in these tests, whereas the fitting tests had at the most a tenth of missing item responses. The fitting test with a relatively high ratio (RS3) is an exception, this test is very long (n items = 105) and candidates often skip a number of items within the test, which does not influence the estimation procedure as much as a series of missing item responses at the end of a test.

The test information curves for the six fitting tests are displayed in Figure 1. The amount of information provided by every test is different (max = 16, min = 3). All information curves have their maximum in the zone of negative thetas. This means that all tests give the maximal information for students in the lower halve of the ability distribution. In Table 3 all tests are displayed with their maximal information.

The tests that give maximal information near the cut-off zone are also those that have the highest degree of information (I(θ=-0.75) = 15.0 (SPA2); 6.5 (RS3); 4.0 (MEC)). The information provided by the RS2, VER1 and VER2 changes very little over the ability scale. In this case the peak of information is not the matter of interest, the amount of information that these tests provide around the cut-off zone is. For example the comparison of the amount of information provided by these tests to the information provided by the SPA2 suggests that the amount of information around the cut-off point could be improved for the RS2, VER1 and VER2. The amount of information that these tests give at the maximal information point does not differ much from the amount of information in the cut-off zone, even though the maximal information is not at all inside this zone (for example the RS2; I(θ=-0.75) = 3.7 whereas its maximal information is 4.0 at θ = -2.40). All fitting tests provide a considerable amount of information about candidates in the cut-off zone.

Discussion

For most of the tests used in the selection procedure at the ENAC the maximal information on the test is obtained at the corresponding cut-off zone on the ability scale. Around the cut-off point in ability, the ability is measured with the most precision, providing evidence that the tests used at the ENAC are congruent with the policy used to select future pilots for training.

By fitting a model such as an IRT model to the data we are in fact developing a metaphysical theory in order to interpret the data (Cliff & Keats, 2003). After the positing of the theory one has to verify if the theory is compatible with the data (e.g. if the model fits the data). The metaphysical entities at hand are the ability (theta) and the probability of endorsing an item given a certain ability (ICC). The theory (IRT) states that the probabilities follow logistic functions with the specificity that the functions cannot cross each other permitting a unique ordering of the items on the ability scale.

For four tests our models did not fit the data. In order to find out whether the characteristics of these tests would better be taken into account by another model, we fitted exploratory the same model without the *pseudo-chance* parameter. By doing this two of the formerly non-fitting models (the RS1 and the SPA1) did fit the data. As the *pseudo-chance* parameter was fixed to zero, the probability to respond correctly to a multiple choice item could approach and even be equal to zero. This is theoretically and practically impossible when the candidate has the possibility to respond to the item. A candidate, in addition to being prepared for the test, also has the possibility to respond by chance ensuring that the probability to find the correct response will always be greater than 0. So what is the point in changing the model? It actually shows how the original 3PL model is based on the assumption that all candidates had the possibility to answer all items. As all tests in this study were subjected to a severe time limit, most candidates did not reach the end of the tests. This caused a lot of missing responses on the final items of most tests. In this pilot selection situation the number of correct responses was used as the ability score, omitted items or missing responses were considered as incorrect responses. By doing this the school hopes to ensure that candidates respond quickly and accurately, and that they measure a right mix of accuracy of the ability and speed. In this study the omitted responses were considered as incorrect responses. As the time limit caused most final items to be left without response, the probability to correctly respond to these items was indeed lower than the *pseudo-chance* parameter of the 3PL model proposes. Nonetheless most models had a satisfactory fit, probably due to the fact that these tests had less missing data than the four non-fitting tests.

Changing the model by fixing the *pseudo-chance* meter to zero is theoretically unasked for, but another interpretation of omitted and missing responses could change the way the test score is obtained. Verhelst, Verstralen and Jansen (1997) propose two types of missing responses. First, missing responses on items in the beginning of the test followed by at least one response on a latter item, and secondly a series of missing responses at the end of the test without any response on latter items. By using the number correct as sole indicator of performance two persons, who respond to a different number of items with the same proportion of correct answers, will obtain the same test score. This approach prejudices persons working slowly but accurately. If speed and accuracy in responding to test items are of equal importance, as it is in this case, it could be interesting to interpret the two types of missing responses differently: The responses on items at the end of the test as not reached/ missing, and the responses on items in the middle of the test as incorrect. In addition to these two types of missing responses they proposed a logistic model for time-limit tests that uses speed and accuracy to come to an ability score (Verhelst, Verstralen & Jansen, 1997). The computer program OPLM (Verhelst, Glas & Verstralen, 1995) offers the possibility to the researcher to use these two types of missing responses as well as the logistic model. In this study we did not use this approach as our sole objective was to find out whether the theoretically most appropriate 3PL model could help to find out if the tests' maximal information was in the zone of cut-off. Further research into the interpretation of missing responses can be of great interest for practitioners that use time-limited tests to measure speed and accuracy/ability at the same time.

Test and item information can be used in many ways. If a test does not give enough information in the target zone, one can improve the test's characteristics. A way to improve these is by selecting items that do give the information in the ability area that is asked for. Then these items have to be examined in order to be able to construct analogous items and as such maximizing the information in this area. Another approach would be to delete the existing items that give very little information or not in the right area (for a complete overview see Embretson & Reise, 2000). It can also be used for computerized adaptive testing (CAT). In CAT the items that are used in a test come from an item bank, the items that an applicant answers correctly influence which items will be proposed. The test is adjusted to the subject's ability level, while he is taking the test, by calculating which items give the most information. This method consequently permits to evaluate the ability level more precisely. Thanks to this procedure, tests can be shortened up to 50 % with equal or better measurement precision (Embretson & Reise, 2000). Unfortunately CAT is not possible at the ENAC as the selection procedure is nationalized and has to be exactly the same for all candidates. The tests can thus not be adjusted to the candidates' ability level. In a private selection procedure CAT can effectively be done and improve the outcome of the selection.

In conclusion, IRT modelling has been shown to be of value in selection settings in aviation psychology. In many distinct areas, from educational testing, via medical diagnoses to personality assessment, IRT models and other latent trait models have replaced the CTT approach. In Aviation Psychology CTT still has the upper-hand though. Illustrative is the fact that the IJAP has published just one research article that used IRT in its twenty years of existence (Mulqueen, Baker & Key Dismukes, 2002). Odd as IRT offers a framework that could shed new light on many different subjects in aviation psychology. The *ab initio* pilot training selection methods in the present article, but also the evaluations of pilot performance could be investigated within the IRT framework.

References

Baker, F. (2001). *The Basics of Item Response Theory.* ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. & Novick, M. R., *Statistical Theories of Mental Test Scores*, (pp. 397-424), 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*, 225-234.

Cliff, N., & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Damos, D.L. (1996). Pilot selection batteries: Shortcomings and perspectives. *International Journal of Aviation Psychology, 6*, 199-209.

Drasgow, Fl, 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.

Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Goeters, K-M., Maschke, P., & Eißfeldt, H. (2004). Ability Requirements in Core Aviation Professions: Job Analyses of Airline Pilots and Air Traffic Controllers. In Goeters, K-M. (Ed.) *Aviation Psychology: Practice and Research (pp. 99-119).* Hampshire: Ashgate.

Hambleton, R.K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: SAGE Publications.

Maris, G. (2002). Concerning the identification of the 3PL model. (retrieved at: http://www.cito.com/research_and_development/pyschometrics/publications.aspx)

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,* 33-45.

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.

Matton, N., Vautier, S. & Raufaste, E. (in press). Test-Specificity of the Advantage of Retaking Cognitive Ability Tests. *International Journal of Selection and Assessment.*

Mulqueen, C., Baker, D.P. & Key Dismukes, R. (2002). Pilot Instructor Training: The Utility of the Multifacet Item Response Theory Model. *International Journal of Aviation Psychology, 12,* 287-303.

Orlando-Edelen, M. & Reeve, B.B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16,* 5-18.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Danish Institute for Educational Research.

Schlessman, B.R. (2009). *Type I error rates and power estimates for multiple item response theory fit indices.* Dissertation, Wright S.U.

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.

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *OPLM: One Parameter Logistic Model*. Computer model and manual. Arnhem Cito.

Verhelst, N.D., Verstralen, H.H.F.M., & Jansen, M.G.H. (1997). A logistic model for time-limit tests. In: van der Linden & Hambleton (Eds.). *Handbook of Modern Item Response Theory.* (pp. 169-185). New York: Springer.

Table 1

Model-data fit - The ratio of the adjusted χ^2 to the degrees of freedom for all tests

Test	**Adj χ^2/df**
MEC	1,12
SPA1 [a]	48,70
SPA2	1,20
RS1 [a]	86,63
RS2	0,19
RS3	2,50
NUM [a]	38,16
VER1	0,68
VER2	0,46
VER3 [a]	21,01

[a] the adjusted χ^2/df for this test is higher than the cut-off value of 3.0

Table 2

Number, mean, and standard deviations of missing item responses for all tests

Missing data	**n items**	**Mean proportion missing responses**
MEC	30	4%
RS2	30	8%
RS3	105	17%
SPA2	40	8%
VER1	50	6%
VER2	43	4%
RS1	*36*	*30%*
NUM	*30*	*38%*
VER3	*30*	*22%*
SPA1	*30*	*28%*

Notes: In *italics,* under the dotted line, the non-fitting tests; SD: Standard Deviation

Table 3

Information and standard error of ability estimation at the cut-off point, range of information, and location of the ability level with the maximal information for all tests

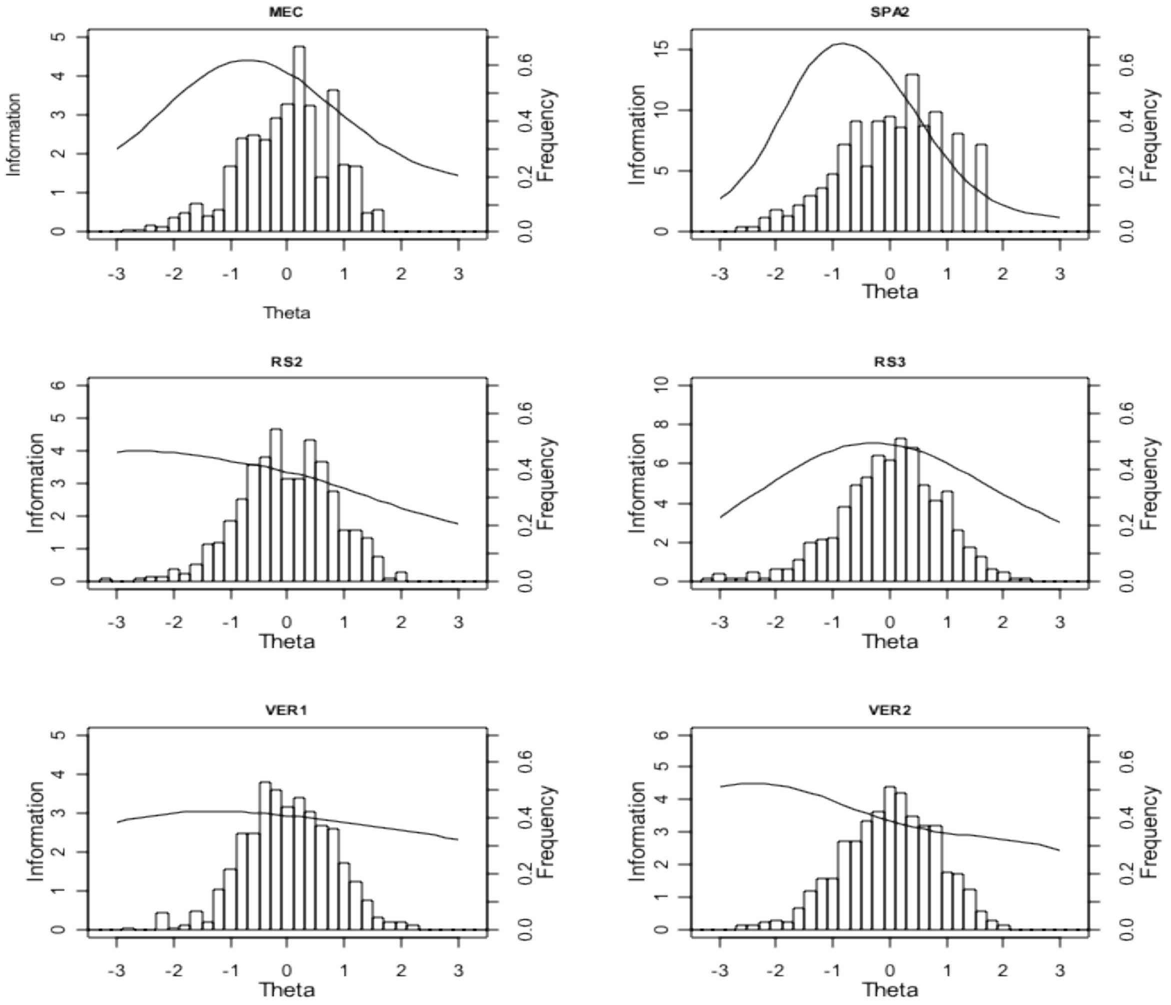
Test	**I (θ=-0.75)**	**SE (θ=-0.75)**	**Information [min,max]**	**θ with maximal information**
SPA2	15.0	0.26	[2.0 - 16.0]	-0.80
RS3	6.5	0.40	[3.0 - 7.0]	-0.37
MEC	4.0	0.50	[2.0 - 4.5]	-0.80
RS2	3.7	0.52	[2.0 - 4.0]	-2.40
VER2	3.5	0.53	[2.5- 4.5]	-2.40
VER1	3.0	0.58	[2.4 – 3.0]	-1.30

Notes: I: Information; SE: Standard Error

Figure Caption

Figure 1.

The test information curves and the theta distribution for the six fitting tests


MEC
SPA2
RS2
RS3
VER1
VER2
Information
Frequency
Theta
-3
-2
-1
0
1
2
3
0.0
0.2
0.4
0.6
0
1
2
3
4
5
6
8
10
15