Statistica & Applicazioni
Vol. IV, n. 2, 2006

Statistical calibration of psychometric tests

Francesca De Battisti – Silvia Salini[§]
Alberto Crescentini[‡]

Summary: *A calibration procedure is generally performed in order to correctly translate the personal traits observed through a psychometric test into numerical values. The calibration process ensures the objectivity of the measure instruments. Psychological measures are usually of indirect type, they are obtained as a result of a statistical inference process. Statistical calibration makes use of particular models, based on the inversion of the previous mentioned indirect measures. The Rasch model can be considered one of this model.*

Keywords: *statistical calibration, intelligence test, Rasch analysis.*

1. Introduction

Calibration is the process whereby the scale of a measuring instrument is determined or adjusted on the basis of a proper experiment. Statistical calibration is a kind of inverse prediction (Sundberg, 1999). In this paper we consider the calibration of a psychometric measuring instrument. In psychometric field classical calibration models can not be applied since the true unknown measure is latent and unobservable. The standard methods (Spearman, 1904; Thurstone, 1938), used in order to get a measure of some psychological attributes, are based on a direct approach, while manifest observed values are indirect measures of the psychological attributes. In psychometric applications the indirect approach is almost exclusively considered.

In section 2 a brief history of the well known Intelligence Test is presented. In Section 3 some details about statistical calibration are given. In

[§] Dipartimento di Scienze Economiche, Aziendali e Statistiche - Università degli Studi di Milano – via Conservatorio, 7, 20122 MILANO (e-mail: francesca.debattisti@unimi.it – silvia.salini@unimi.it).

[‡] Dipartimento di Psicologia – Università Cattolica del Sacro Cuore – Largo Gemelli, 1 20143 MILANO (alberto.crescentini@unicatt.it).

Section 4 the Rasch model is described. In Section 5 the Rasch analysis is applied to an Intelligence Test and the appropriateness of the relating model is discussed.

2. Psychometric tests to measure the intelligence

The origin of the attribute "mental test" is traditionally connected to the work of Cattell (1890) who used the experimental method to measure psycho physic reactions and cognitive elementary processes.

In the psychological field, intelligence is one of the most important dimension considered and, since the beginning of the last century, several instruments were made available for different purposes. With reference to the intelligence appraisal we can found instruments built for different population: for instance, the child evaluation (an example is the Binet-Simon (1905) scale) or test for military selection (an historical example is the Army test).

What intelligence is, and consequently the ability of its evaluation, still plays an important role in the nowadays scientific debate (Gould, 1996): the hierarchical theory and the multifactor theory are the main categories. The European scientific community pays more attention to the first one (see, for example, the work of Spearman, *inter alia* 1904 and 1927); the second approach, started with Thurstone's (1938) work, has an important impact on the North American studies.

Both theories agreed on the existence of an entity, called *g* factor (from general) which is a dimension of intelligence poorly influenced by culture. Some authors said that it is inherited or inborn; others agreed on the existence of the *g* dimension but refused the idea of inheritance of intelligence.

The first tests expressly made to evaluate *g* are due to Raven (1940) and Cattell (1940), although they started from different assumptions. The Raven test, called Progressive Matrices (PM), was widely used in Great Britain during the WWII for selection of soldiers with the aim of evaluating subjects without making use of language. The Cattell test, firstly called Culture Free and successively Culture Fair, evaluates subjects using items not suffering from the influence of their different cultures. A test with characteristics similar to the Raven one was built by Anstey and Illing (Anstey, 1955) expressly made for military selection in Great Britain and to retest the diagnosis made with PM. It is based on domino cards and shows an higher saturation in the *g* factor (Vernon, 1947). Each item is made by a logical sequence of domino cards, and the subjects have to write the two missing numbers of the last card. The most widely used domino test in Europe is the D 48, that is a French version of the Anstey and Illing instrument made by Pichot (1949), translated into Italian in 1954. Actually it is used also in USA for trans cultural studies (Domino, 2001). The D 48 test is made for subjects from the age of twelve. There are 48 items, 4 of them devoted to the training of subjects. In the original version the respondent must write the numbers to complete the sequence, as in the Anstey and Illing instrument. There is a fixed time limit and the total score is calculated as the number of right answers given by the subject. The items are organized in sequences that follow the same logic process and spatial representation, selected from the following five different structures: six cards disposed on two lines, nine cards disposed on three lines, a rose of five or eight cards, cards with spiral and with oval disposition (*Les Editiones du Centre de Psychologie Appliquées*, 2000). Each sequence has a growing difficulty related to the different logical process: spatial, numerical, arithmetical and mixed items.

3. Statistical calibration

In statistical calibration the two following measures are considered: the *standard measure* (**X**), which is expensive, accurate and not easy to reach; the *test measure* (***Y***), obtained by the measurement instrument, which is cheaper, less accurate and easier to reach. The calibration experiment starts with an initial sample of *n* observations (x_i, y_i). Classical calibration theory assumes that the test measure ***Y*** (stochastic variable) is linked with the standard measure **X** (not stochastic) through a linear model, whose parameters are estimated by the observations.

In the prediction experiment it is possible to estimate the true unknown measure **X** when the *test* measure ***Y*** is observed by inverting the linear model (Brown, 1993). This approach is also called *inverse regression* (Sundberg, 1999).

The literature on calibration deals with the mathematical problem of inversion, the statistical properties of the estimators obtained with the inversion, the extension to the multivariate context (Salini, 2003).

Statistical calibration models are not properly valid in psychological applications, since psychology deals with unobservable variables for which a calibration experiment is not available. On the contrary, statistical models with a structure formally similar to the linear models adopted for the calibration experiment are frequently considered in psychometric applications: latent attributes are expressed by a linear combination of the answers to a battery of items in a questionnaire (Boncori, 1993). The main weakness of this approach is that the linear weights are not obtained following a rigorous estimating procedure. Furthermore, the goodness of fit of the model cannot be evaluated, being the true measure not observable. The third, and more relevant, weakness depends on the implicit assumption that the psychological characteristic (***Y***) should be defined as a function of the ($\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_n$). This natural asymmetry is actually reversed, depending the answers to the items on the psychological characteristic and not vice versa. For example, the intelligence quotient of a subject is given and the results in an intelligence test hangs on the intelligence of the subject that carries out the test.

In order to proceed in a correct way, a model should be available such that the answers to the items ($\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$, …, $\boldsymbol{Y}_\mathrm{n}$) be functions of some psychological characteristic (**X**). Since **X** is not perceivable, the direct model cannot be estimated, thus it cannot be inverted.

Observe that models defining the measure of a latent variable by indirect measures of manifest variables exist and they can be considered statistical calibration models: the most important one is the Rasch model (Rasch, 1960).

4. Rasch model: an overview

In 1960 Georg Rasch stated that the answers to an item depend on two independent factors: the ability of the subject and the intrinsic difficulty of the item. He proposed an item-response model, allowing to measure both the item difficulty and the subject ability along a shared continuum. In the dichotomous case the model expresses the probability of right response by the following relation:

$$P(x_{ij} = 1) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \tag{1}$$

in which x_ij is the answer of subject *i* (*i*=1,….,*n*) to item *j* (*j*=1,….,*k*), θ_i is the ability of the subject *i* and β_j is the difficulty of the item *j*.

The model has a unique set of properties, making it an ideal tool for testing the validity of ordinal scales (Wright and Linacre, 1989).

The Rasch model uses a particular unit measure, called *logit* (Wright and Stone, 1979). With the transformation from *raw scores* into *logits* (or from ordinal-level data into interval-level data), the parameters θ_i and β_j can be expressed in the same unit measure, just the *logit*, thus they can represent subjects and items on a shared continuum respectively.

The Rasch model produces person-free measures and item-free calibrations, abstract measures that overcome specific person responses to specific items at a specific time. This characteristic, unique to the Rasch model, is called *parameter separation*. Thus, Rasch parameters represent a person ability as independent of the specific test items and item difficulty as independent of specific samples (Wright and Masters, 1982). Necessary information to estimate θ_i and β_j is respectively contained in the number of items got through by the subject *i* ($\mathrm{r_i}$) and in the total number of correct answers for item *j* (s_j). So, the scores r_i and s_j represent *sufficient statistics* for the parameters θ_i and β_j.

The model is probabilistic, not deterministic, defining for each subject/item interaction the probability of right answer. The model is prescriptive, not descriptive. This also allows to estimate the precision of the computed measure of difficulty/ability. It requires unidimensionality, that is all the items measure only a single construct, and local independence, that is, conditionally to the latent trait, the responses to a given item are independent from the responses to the other items.

It can be noticed that the difference between ability/difficulty latent traits and x_{ij} manifest variables is both metric as conceptual. Latent traits are not observable, not stochastic and expressed in a numerical interval scale. Manifest variables x_{ij} are observable, stochastic and expressed in a ordinal (dichotomous) scale. The model, as formulated, is in all respects a calibration model in which a multivariate measure obtained through a measuring instrument (the psychological test) is linked by a direct relation to a latent unknown measure (the ability of the subject). According with the typical terminology of the calibration context, an indirect *test* measure is observed to get the true *standard* measure. The probabilistic statement of the model makes it possible the construction of goodness of fit tests on the complete model as well as on the single items, which constitute the calibration of the psychological measurement instrument.

5. Application: calibration of an Intelligence Test

5.1 *Preliminary analysis*

The Rasch model will be applied to data collected for 958 subjects participating in 2002 to the selection procedure for the application to the Psychology degree course of the Catholic University of Milan. Data refer to a closed version of the D 48 Intelligent Test, built for selection on large number of candidates. This version uses the original 44 items but respondents, instead of writing the numbers in each empty card, are requested to choose the right answer in a set of five cards, where the correct one is always present. In case of correct answer the score 1 is assigned, null otherwise[1]. This closed version makes the instrument easier than the open one. For this reason, fixed time limit of 18 minutes is assigned (25 in the original version), so that it may become more difficult to complete all the

[1] It is important to note that the score 0 identifies indifferently wrong answer and non response too. In this context, the formalities for carrying out the test, given to the subjects, inform them that a non response is a wrong answer. This is a basic rule for the construction of this psychometric test.

items. The subjects can choose the question resolution order, but usually they tend to proceed in a sequential way (Crescentini et al. 2003). In figure 1 the frequency distribution of raw scores is shown. The majority of the subjects presents raw score in the 25-35 range, with minimum 0 and maximum 44. This distribution is in according with previous studies regarding people with a secondary school degree (Cristante and Borgatti, 1975; Boncori, 1987). A few candidates have raw score less than 15 or greater than 40, coherently with the idea expressed by Bruni (1966) on time limitation.

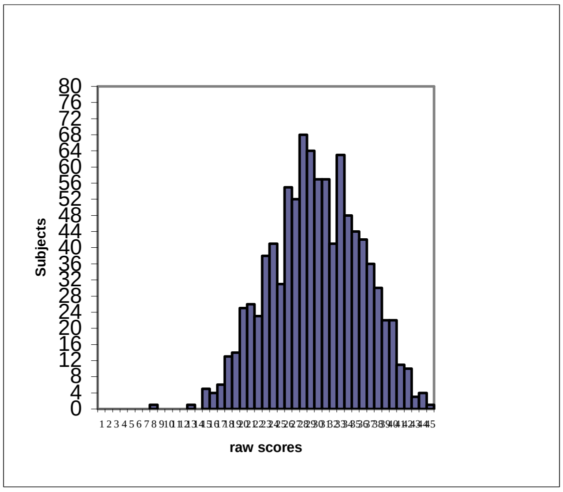


Figure 1. *Frequency distribution of raw scores for 44 items*

The Dichotomous Rasch Model, *Simple Logistic Model* (Rasch, 1960), is available in the computer program RUMM (Rasch Unidimensional Measurement Models) by Andrich, Sheridan, Lyne and Luo (2000). It produces scale-free subject measures and sample-free item difficulties (Andrich, 1988; Wright and Masters, 1982). The items are calibrated from easy to hard and the subject measures are aligned, on the same scale, from lower to higher.

Figure 2 shows the classical “Rasch ruler” (also called the “Item map”) obtained for our data. The vertical dashed line represents the ideal less-to-more continuum of “level of intelligence”; for simplicity, we prefer to use the term intelligence instead of the more correct “estimated intelligence”. Items and subjects share the same linear measurement units (logits, left column). Conventionally, the average item difficulty (as for intelligence we will use “difficulty” instead of “estimated difficulty”) is set to 0. On the right of the dashed line, the items are aligned from easiest to hardest, starting from the bottom. Along the same ruler, on the left, the subjects are aligned in increasing order of intelligence from bottom to top. Each X symbol represents 6 subjects. One subject reaches the extreme score of 44; it is omitted from the analysis since, according to the Rasch model, his/her ability cannot be estimated.

Subject scores range from –1.4 to 4.8 logits, while item locations from –4.2 to 2.6. Thus we observe a spread in difficulty of almost 7 units and more than 6 in intelligence. The measure of the intelligence obtained by this set of items seems reliable being the range wide enough. If all the items have the same characteristics, the probabilities of the answers' profiles are similar giving no raise to a *continuum*, but only a point. The range of items does not match completely the range of intelligence scores. There is a lot of subjects at the upper end of the scale and there are not subjects at the lower end. Furthermore, 84 subjects have a level of intelligence higher than the most difficult item (from 2.6 to 4.8 logits) and 12 items have a difficulty easier than the less intelligent subject (from –1.4 to –4.2 logits).

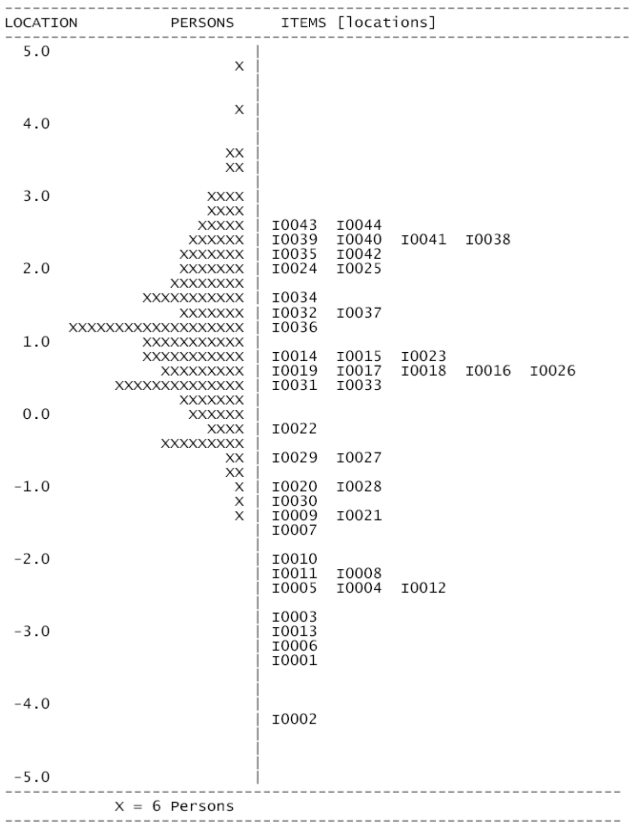
Thus, it seems that the item difficulties are not appropriately targeted to the subjects (only 874 of out 958 intelligence measures (91%) are "covered" by item difficulty). The first part of the scale is too easy, but this fact is coherent with the logic of the heating exercises.

Furthermore items are well spanned and spaced throughout the continuum. This can be taken as an indicator of accuracy. With the "same" increase of intelligence level there is the "same" increase in the total raw score. This is not completely true, because there is a *potential redundancy* when a lot of items are on each tick; so, when a particular level of intelligence is achieved an increase of 4 to 5 marks (as many items on the same tick) could be in the total raw score.

5.2 *Some problems: item redundancy and time effect*

This analysis outlines some potentially redundant items: those with the same difficulty level, that in the graph are on the same line (e.g. 19, 17, 18, 16, 26; or 39, 40, 41, 38). The redundant items are always part of the same sequence (group of items with the same logic process and spatial representation) so we can affirm that the difficulties are connected with the logical process that lay behind the item construction.

The sequence starting with the item 14 and terminating with the item 26 shows some redundancies. From a calibration perspective we can say that some of them should be eliminated or changed, but analysing the items from a psychometric perspective we find some problems. The difficulties of the items are connected, as we already observed, with the number of right answers given by the subjects. The items 16 and 17 (see Figure 3) have the same difficulty (the location value is 0.721 for item 16 and 0.689 for item 17, corresponding respectively to 579 and 586 right answers); but if we observe the number of non responses there is a great difference that has to be



```
-----------------------------------------------------------------
LOCATION          PERSONS      ITEMS [locations]
-----------------------------------------------------------------
  5.0                        |
                           X |
                             |
                             |
                           X |
  4.0                        |
                             |
                          XX |
                          XX |
                             |
  3.0                   XXXX |
                        XXXX |
                       XXXXX | I0043  I0044
                      XXXXXX | I0039  I0040  I0041  I0038
                     XXXXXXX | I0035  I0042
  2.0                XXXXXXX | I0024  I0025
                    XXXXXXXX |
                 XXXXXXXXXXX | I0034
                     XXXXXXX | I0032  I0037
         XXXXXXXXXXXXXXXXXXX | I0036
  1.0            XXXXXXXXXXX |
                 XXXXXXXXXXX | I0014  I0015  I0023
                   XXXXXXXXX | I0019  I0017  I0018  I0016  I0026
              XXXXXXXXXXXXXX | I0031  I0033
                     XXXXXXX |
  0.0                 XXXXXX |
                        XXXX | I0022
                   XXXXXXXXX |
                          XX | I0029  I0027
                          XX |
 -1.0                      X | I0020  I0028
                           X | I0030
                           X | I0009  I0021
                             | I0007
                             |
 -2.0                        | I0010
                             | I0011  I0008
                             | I0005  I0004  I0012
                             |
                             | I0003
 -3.0                        | I0013
                             | I0006
                             | I0001
                             |
                             |
 -4.0                        |
                             | I0002
                             |
                             |
                             |
 -5.0                        |
----------------------------------------------------------------
            X = 6 Persons
----------------------------------------------------------------
```

Figure 2. *Item map*

explained: there are 2 non responses on the first item and 122 on the second one. This sequence is made by items disposed on a rose configuration and the target (position of the answer) is disposed skewed. In the set of possible answers for the first three items (from 14 to 16) we have the correct one, and also the answer that is the mirror of this one; on the item 17 the mirror opportunity is absent. Analysing the wrong answer in the first three items we find that most of them are on the mirror choice. We can suppose that if a subject chooses a response strategy he perseveres in his error; when he has to respond to the item 17 and there is not the answer coherent with this strategy he prefers to make no choice. The different logical sequences need items on the same level to verify the acquisition of the logical process. In this sequence the item 17 gives us the opportunity to verify our hypothesis.

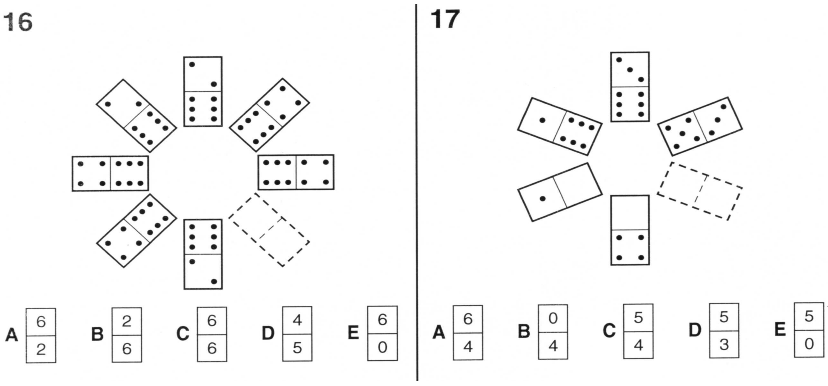


Figure 3. *D48 Test, item 16 and item 17*

Since time limit plays an important role in the performance of subjects (Csonka, 1973), some authors (Bruni, 1966) suggest to give more time up to 45 minutes. Relating to the non responses we find that from item 35 to the end of the test more than 300 subjects gave no answer. We may suppose that many of them have not received enough time to complete the test. To further analyse this dimension we cut the test on item 35: figure 4 represents the frequency distribution of the raw scores for the remaining 34 items and figure 5 represents the corresponding item map.

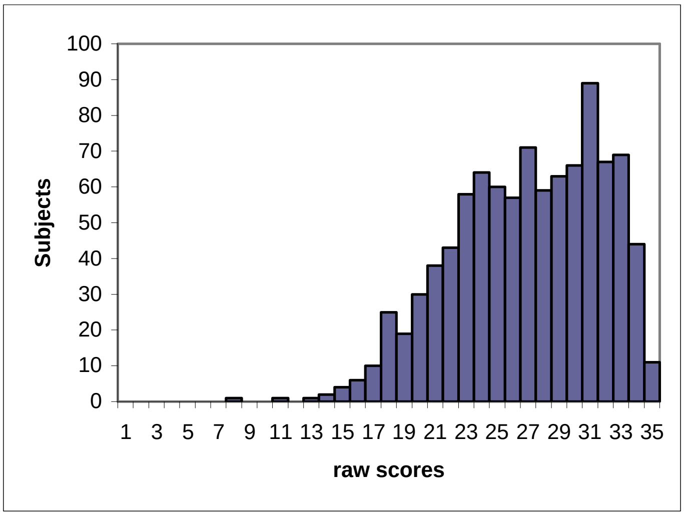
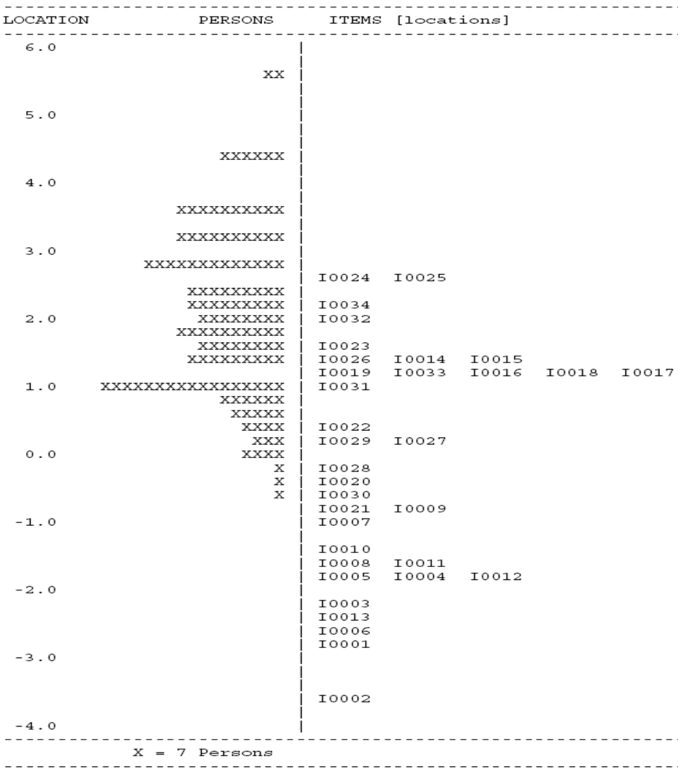


Figure 4. *Frequency distribution of raw scores for 34 items*



```
------------------------------------------------------------------
LOCATION           PERSONS      ITEMS [locations]
------------------------------------------------------------------
  6.0                         |
                              |
                           XX |
                              |
                              |
  5.0                         |
                              |
                              |
                       XXXXXX |
                              |
  4.0                         |
                              |
                   XXXXXXXXXX |
                              |
                   XXXXXXXXXX |
  3.0                         |
                XXXXXXXXXXXXX |
                              | I0024  I0025
                    XXXXXXXXX |
                    XXXXXXXXX | I0034
  2.0                XXXXXXXX | I0032
                   XXXXXXXXXX |
                     XXXXXXXX | I0023
                    XXXXXXXXX | I0026  I0014  I0015
                              | I0019  I0033  I0016  I0018  I0017
  1.0       XXXXXXXXXXXXXXXXX | I0031
                       XXXXXX |
                        XXXXX |
                         XXXX | I0022
                          XXX | I0029  I0027
  0.0                    XXXX |
                            X | I0028
                            X | I0020
                            X | I0030
                              | I0021  I0009
 -1.0                         | I0007
                              |
                              | I0010
                              | I0008  I0011
                              | I0005  I0004  I0012
 -2.0                         |
                              | I0003
                              | I0013
                              | I0006
                              | I0001
 -3.0                         |
                              |
                              |
                              | I0002
                              |
 -4.0                         |
------------------------------------------------------------------
            X = 7 Persons
------------------------------------------------------------------
```

Figure 5. *Item map for 34 items*

So we find a ceiling effect. The modified scale seems to be easier for the sample. This is coherent with the hypothesis of a difficulty connected with the presence of a time limit, although a complete evaluation of this hypothesis is not made.

Since the value of the person parameter θ may be used to select a subsample of subjects with the best performance, we choose subjects with a θ value more than 3[2]. We analyse the answer profile of these subjects and we make a comparison between this subgroup (subjects with the best performance) and the complete sample (958 subjects). In particular, we focus on the last items to check if they discriminate between clever subjects and the others. Observing only the items from 35 to 44 we find that the best subjects gave less non responses. Besides, analysing the ratio of wrong answers over total answers, the items from 41 to 44 show no differences between best performing subjects and the total sample. This suggests (confirming the hypotesis of Bruni, 1966) that subjects that give random

[2] There aren't specific methods to fix a cut-off for the person parameters; from the item map in figure 2 we can observe a group of subjects with θ more than 3, so we choose them as the subjects with best ability.

answers reach better results than people that prefer to leap over some items without giving any answer.

Concerning the model fit indexes, the RUMM program uses the parameter estimates to examine the difference between the expected values predicted from the model and the observed values. Furthermore, the program provides item and subject fit statistics and two global tests-of-fit: the *Item-Subject Interaction* and the *Item-Trait Interaction*.

Table 1. *Fraction of wrong answers for the last 10 items*

Item	Best subjects	All subjects
35	0.1	0.34
36	0.02	0.18
37	0.04	0.23
38	0.22	0.3
39	0.16	0.28
40	0.08	0.24
41	0.29	0.34
42	0.23	0.24
43	0.23	0.26
44	0.29	0.24

Table 2. *Fraction of no responses for the last 10 items*

Item	Best subjects	All subjects
35	0.02	0.37
36	0.02	0.33
37	0	0.34
38	0.06	0.47
39	0.08	0.47
40	0.06	0.51
41	0.02	0.44
42	0.06	0.49
43	0.12	0.53
44	0.16	0.57

5.3 *Analysis of residuals*

The item-subject test-of-fit examines the response patterns of subject across items and item across subjects. It takes into account the residuals between the expected estimate and the actual values for each subject-item, summed over all items for each subject and over all subjects for each item. The fit statistics for the item-subject interaction approximate a distribution with zero mean and unitary standard deviation, when the data fit the measurement model.

Let x_{ij} and π_{ij} be respectively the observed and the expected values; the standardized residuals are then $z_{ij} = (x_{ij}- \pi_{ij})/(\pi_{ij}\ (1-\pi_{ij}))^{1/2}$.

We consider the sums of squares:

for item $U_j= \sum_{i=1}^{n} z_{ij}{}^2 / n \qquad j=1,\ldots.,k;$

for subject $W_i= \sum_{j=1}^{K} z_{ij}{}^2 / k \qquad i=1,\ldots.,n.$

These are the so-called fit mean-squares, taking values from 0 to $+\infty$.

Andrich (1988) proposes a transformation of $\sum_j z_{ij}{}^2$ for each subject *i* (and then for each item *j*) to obtain the following standardized residual:

$$Y_i = \left[\sum_j z_{ij}{}^2 - (k-1) \right] / \sqrt{Var\left(\sum_j z_{ij}{}^2 \right)} \ .$$

Then a large negative value of Y_i implies that the model is overfitted, while a positive large value implies a misfitting pattern; a value close to zero implies a typical pattern.

Table 3 shows the residual printout by RUMM, for items and for subjects.

Table 3. *Summary of global fit statistics*

ITEM-PERSON INTERACTION				
	ITEMS		PERSONS	
	Location	Residual	Location	Residual
Mean	0	-0.319	1.229	-0.345
SD	1.947	1.947	1.07	0.953

If mean and SD of subjects' intelligence are overlap mean and SD of the items' difficulty, the targeting of the scale is good. Subjects' average intelligence (1.229) is greater than item mean difficulty (0) and item SD (1.947) is greater than subject SD (1.07). So, the targeting of the scale doesn't seem very good.

When data perfectly "fit" the model the residuals are expected to have zero mean and SD close to 1. In our case the residual means are quite good, -0.345 for subjects and -0.319 for items; the subjects SD is good (0.953), while the item SD is a little too big (1.947).

Appendix A shows for each item the estimated parameter and the residual (Individual item fit); Appendix B contains for each subject the estimated parameter and the residual (Individual person fit), only when residuals are larger than 2 in absolute value.

The fit of individual items of the measurement model is not definitely good: 13 item residuals are larger than 2 in absolute value (bold typed in Appendix A).

As shown in Appendix B the fit of individual persons of the measurement model is not good at all, but only 47 person residuals have absolute value larger than 2.

We analyse the residuals to check their distribution, the tails and the symmetry[3]. In figure 6 and 7 it can be noticed that the distribution of standardized residuals differs normal distribution[4] but less than 5% values exceed the limit –2 and +2, so the tails are fine like in a normal distribution. The subjects 409 and 208 have the biggest positive residuals and subject 262 has the biggest negative residual. Furthermore, we can also observe that the distribution of residuals presents higher variability for subject with very low level of intelligence.

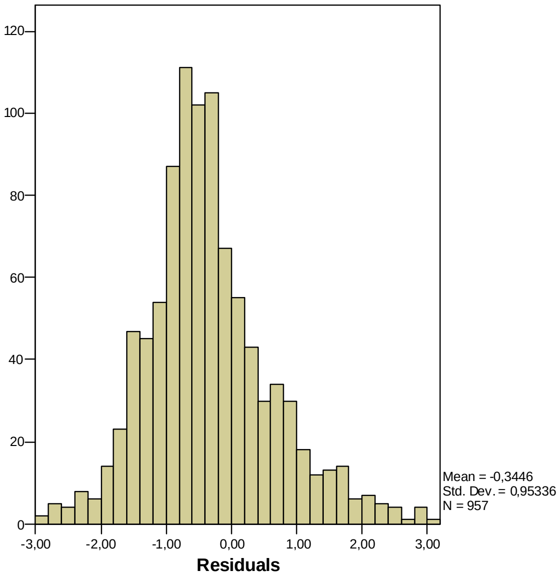


Figure 6. *Histogram of residuals*

[3] This is a descriptive analysis. Rasch model doesn't presuppose distributional assumptions.

[4] The Kolmogorov-Smirnov statistic is 0.081 with p-value 0.000, so the residual distribution is not normal; this depends on the bias –0.345 and symmetry index 0.610.

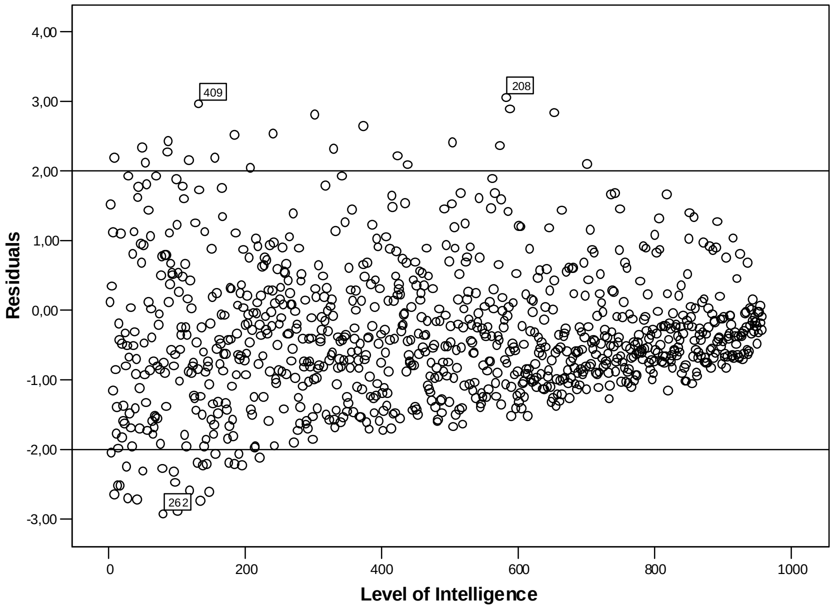


Figure 7. *Person residuals, sorted by Level of Intelligence*

We have identified the 18 subjects that show the worst fitting with the model, all of them have a standardized residual greater in absolute value than 2.5. These subjects are bold typed in Appendix B.

The 8 subjects with residual values bigger than 2.5 employ a profitable answer strategy. They give random answer instead of BLANK answer. This is observable in two factors: correct answers in items very difficult and wrong answers in easier items. In general intuitive items are correct, on the contrary items that require sequential argument approach are often non correct.

The 10 subjects with residual values lower than - 2.5 employ a not profitable answer strategy. They proceed in a sequential way and so they give BLANK in the final items; may be they need more time. The mean score for these subjects is low and correct answers are in the first items. Bruni (1966) underlined the time problem in intelligence tests. The time limit may be rewards fast and not accurate subjects and penalizes slow and accurate subjects. Analysing their results we find that they made many mistakes during the initial part of questionnaire, for instance in item 2 all the errors of respondents are due to these subjects. The mean score of this sample is 22.50 and the standard deviation is 5.29, while the mean score of the total sample is 28.77 with a standard deviation of 6.06. Nevertheless, in the more difficult items their performance appears better than the global average. Item 14 represents a turning point for the total sample; wrong responses are 2% on the 13 and 43% on the 14. In item 13 the majority of misfitting subjects gave the right answer. They reach better results on the items that can be done through an intuitional basis instead of a reasoning basis (e.g. items 20 and 21). The subjects of this group did not use the non response strategy from item 24 to item 44, in fact they gave less non response than the total sample. We can say that these subjects are more fast than precise, probably they are not too engaged in doing the test or they are not too much concentrated in the proof; a support of this hypothesis comes from the bad performance on items that require a logical process with precise steps (e.g. 27 and 28).

5.4 *Item-Trait Interaction*

The item-trait test-of-fit examines the consistency of every item parameters across the subject measures: data are combined across all items to give an overall test-of-fit. This shows the overall agreement for all items across different subjects. Rasch "misfit" values indicate those items which do not share the same construct with the other ones (items with higher misfit should be removed).

The observed answer distribution is compared with the expected answer distribution, calculated with the logistic function, by means of the Chi-squared criterion. The following steps may be performed.

i) Examine the χ^2 probability (p-value) for the whole item set; there is not a well-defined lower limit defining a good fit (minimum acceptability level); a reference level may be 5%. The null hypothesis is that there is no interaction between responses to the items and locations of the subjects along the trait. In our case (see table 4) Total Item ChiSq = 826.577 and Total ChiSq Prob = 0.000, so the null hypothesis is strongly rejected.

Table 4. *Summary of global statistics*

ITEM-TRAIT INTERACTION		*RELIABILITY INDICES*	
Total Item Chi Sq	826.577	Separation Index	0.842
Total Deg of Freedom	396	Cronbach Alpha	0.838
Total Chi Sq Prob	0.000		

ii) If the overall χ^2 probability is less than 5%, examine the χ^2 for each item to identify anomalous statements (see appendix A, where ChiSq is the Item-trait interaction chi-square statistic for each item and Probability is the probability of its occurrence for the degrees of freedom listed).

iii) Analyse each misfitting item to understand the misfit causes.

The subjects are splitted into "intelligence level" classes, with constant width; in every class the observed answers' proportions are compared with the model's estimated probabilities, for each answers category, and the χ^2

value is worked out. The overall χ^2 is the sum of the single group χ^2. The contribution's amount to the sum highlights the misfit seriousness in the respective class: the highest the χ^2 value in the single class, the most serious the damage by the gap between data and model. This is the so called "Differential Item Functioning" (DIF) and the term indicates the instability of the hierarchy of item difficulty levels (the same scale may not be suitable for measuring exactly the same variable across groups).

DIF can be observed for some items. The DIF is not an all-or-none phenomenon. It is always in the background, so that detection is only a matter of power of measurement. The greater the sample, the more any DIF may become statistically significant. It must be reminded that the DIF in itself does not bias the cumulative expected scores across groups. The Rasch model assigns an overall measure, an expected score to each subject, whatever his/her group assignment. If a given group gets a score higher than expected in one item, it gets a score lower than expected in at least one of the other items. Similarly, an item easier than expected for some groups, may result more difficult than expected for another.

DIF challenges the nature of the measure, not the amount. In this case the ruler is not independent of the person measured. Being very restrictive on what would be expected, the Rasch analysis is a powerful tool to detect any DIF.

DIF indicates that the instrument tackles qualitatively different constructs across the groups, whatever the ability measures of the subjects. A decreasing DIF in subsequent replications flags the right direction during the scale construction.

The items with the biggest value of Chi-Square are evidenced in italics in the Appendix A. A more detailed search for "systematic" misfits, DIF across subjects' subgroups (classes), can be conducted in a graphical way (see Tesio et al. 2002).

For example, Figure 8 and Figure 9 present the so-called Item Characteristic Curve (ICC) of the item 26 and the item 18 respectively. The ICC reflects the probability of getting the maximum score of 1. The ordinate gives the score ideally expected by the model, ranging from 0 to 1. The abscissa gives the intelligence of the subjects in logit units. For dichotomous items (1-threshold items) the curve follows the S-shaped (logistic) function given by the core equation of the Rasch model (1). The two curves in figure 8 and 9 share the same slope for each dichotomous item, but the average location along the abscissa changes according to the item average difficulty.

Moreover, the sample was split into 10 equally-sized subgroups, representing different classes of overall ability. For each class, the mean expected score was plotted in dot symbols as a function of the mean ability. This is a basic investigation of DIF. The analysis is conducted in order to understand if subjects of different level of intelligence follow the Rasch model and to measure if a generic item is more or less easy, in itself, in the various classes.

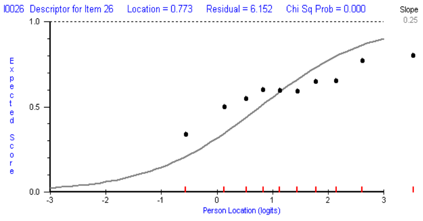


Figure 8. *ICC for Item 26*

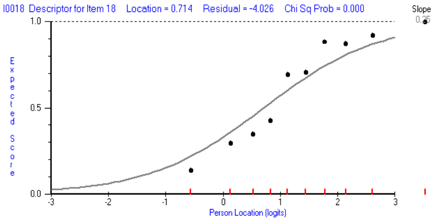


Figure 9. *ICC for Item 18*

The item 26 (Figure 8) is easier than expected for classes of subjects with low level of intelligence and it is more difficult for classes of subjects with high level of intelligence. For the item 18 we found an opposite performance (Figure 9).

In this view, absolute score deviations should be assessed in conjunction with standardized residuals. Whether these residuals on these items are acceptable or not, is a matter of context decision. The robustness of the scale with respect to DIF can only be ascertained from its performance in real applications.

It is interesting to distinguish the analysis by the different groups of subject. For example sex may influence the results in some items (Crescentini et al. 2003), like in item 14. By RUMM is possible to appreciate this influence by performing the DIF analysis for multiple factor with multiple levels. We consider a single factor, sex, with two levels. The ICC of item 14 is reported in Figure 10.

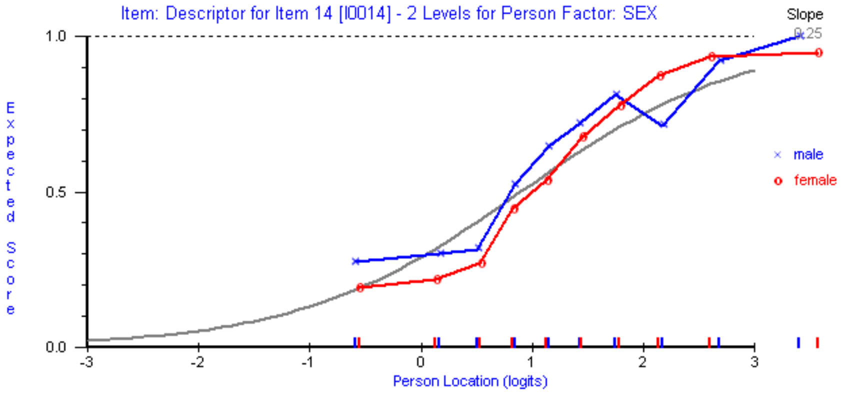


Figure 10. *ICC for Item 14 divided by sex*

From ICC of item 14 (Figure 10), it seems that low level females show more difficult to find the correct answer than the expectation. In this item males are always better than the expectation.

The RUMM program calculates a Person Separation Index, which is the Rasch reliability estimate, computed as the ratio (true/(true+error)) variance whose estimates come from the model. A value of 1 indicates lack of error variance, and thus full reliability. This index is usually very close to the classic Cronbach α coefficient computed on raw scores. In our case (see table 4) the Separation Index is 0,842; this means that the proportion of observed subject variance considered true is 84,2%. The power of test-of-fit, based on the Person Separation Reliability of 0,842, is good.

5.5 *Factor Analysis on Residuals*

The Rasch model assumes that residuals are randomly distributed across items. A high correlation (computed on standardised residuals across pairs of items) would thus suggest inter-item dependency coming from an extraneous shared construct challenging the undimensionality of the measure. We have considered the highest correlations, with values almost greater than |0.5|; in this case there is significant dependence.

A way of detecting important deviations from the fundamental undimensionality requirement of Rasch model is the application of factor analysis techniques to the residual matrix. If the information about person – item interaction modelled by Rasch and extracted from the data matrix leaves a random dispersion of residuals, then the claim is that the solution is accounting for just one dimension.

The factor analysis, as it is shown in the scree plot in Figure 11, confirms all the other results obtained. In fact some residuals are correlated. The presence of factor loadings in the analysis of residuals would suggest the presence of more than one underlying test dimension.

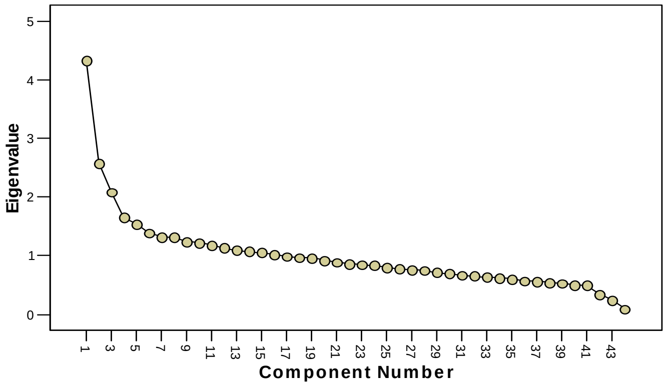


Figure 11. *Scree Plot*

The Component matrix, partially reported in Table 5, displays that the items for which the residuals are strong correlated are items from 14 to 19, in according with comments in Figure 2.

Table 5. *Component Matrix*

item	*Component 1*
14	-0.725
15	-0.756
16	-0.809
17	-0.293
18	-0.609
19	-0.717

5.6 *Critical Themes*

In appendix A it is shown the χ^2 test for each item. 22 items have a high value of χ^2 with significance value less than 0,05. In order to calibrate the questionnaire these items have to be deleted. So the global χ^2 decreases and the χ^2 for each item becomes acceptable.

The deleted items are chosen by considering both residuals and χ^2 values (see appendix A). 13 items have residual value greater than |2|. Three homogeneous groups can be evidenced.

Table 6. *Misfitting and Overfitting Items*

GROUP	ITEMS	DESCRIPTION
1	16, 17, 18, 19	Residuals < -2, only item 17 has χ^2 significance value greater than 0,05. These items are discriminated for the sample (see DIF)
2	22, 23, 24, 26	Residuals > 2, no items has χ^2 significance value greater than 0,05. These items are not discriminated for the sample (see DIF)
3	27, 28, 29, 30, 32	Residuals < -2, only item 27 has χ^2 significance value greater than 0,05. These items are discriminated for the sample (see DIF)

We have reported separately the group 1 and the group 3 because they are formed by items with different logic process and spatial representation.

So 11 items can be deleted: the ones with high residuals and high χ^2 value.

The analysis is iterated step by step in order to obtain a global χ^2 greater or near than 0.05, and also for each item. In particular, after to have deleted 11 items, we have performed a new Rasch analysis to check if the global χ^2 was greater or near than 0.05 and, if not, to individuate other items with χ^2 significance value lower than 0.05. We have removed these items, we have made another analysis, and so on, to obtain finally a global χ^2 greater or near than 0.05, and also for each item. In this iterative procedure the items 4, 2, 40 and 7 are step by step deleted. In the final version only 29 items are maintained[5]. As said above the items of the initial version of the instrument are organized in sequences that follows the same logic process and spatial representation. The differences between sequences help to obtain a better evaluation. This version of 29 items contains items from all the sequences, maintaining the initial structure of the test. This new instrument must be tested by combining time factor and evaluation of errors.

As a preliminary analysis we apply the Rasch Model to only the 29 items. The analysis on the reduced questionnaire evidences 3 items with residual values greater than |2|. Item 36 in particular has residuals value equal to -2.22, item 14 equal to 2.32 and item 15 equal to 2.49. The factor analysis of residuals evidences the presence of one component more relevant than the others, correlated with item 14, 15 and 20.

This result encourage to submit the reduced version of the test on a new sample of subjects.

6. Conclusions

The analysis system has shown that the closed version of the instrument has the same metric characteristics of the original opened one (see Csonka, 1973), but to fill in the whole questionnaire is required less time, in particular 28% of time can be saved. It is also possible to automate the correction phase using an appropriate answer sheet. The instrument requires a minimum level of education, connected with the use of numbers in the closed answers. The time limit question (Bruni, 1966) still remains; a time limit imposes a cut to the slower subjects and facilitates the faster ones. The time limit without a penalization for the wrong answers drives subjects to use a strategy that maximizes the number of given answers and reduces the accuracy. We are evaluating the hypothesis of using a penalty to the wrong answers.

In the "Rash ruler" we found the dimensions of the instrument, which give the chances to make further analysis. The use of θ to select the better subjects makes us find a group of people that follow the same answer

[5] The items in the final version are: 1, 3, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 17, 20, 21, 25, 27, 31, 33, 34, 35, 36, 37, 38, 39, 41, 42, 43, 44.

strategies; the same happens with the analysis of residuals. The analysis of the Item Characteristic Curves with the two levels of sex operating confirms the hypothesis formulated by Crescentini et al. (2003) and gives some new perspectives. We have yet not found an acceptable hypothesis that explains the difference, between male and female, in the answer strategies.

The ability of the instrument defined on the 29 not redundant items, proposed in the previous paragraph, has to be experimentally proved on a sample of subjects.

Acknowledgement

The authors are grateful to Prof. B. V. Frosini and Prof. G. Boari for their helpful suggestions.

References

Andrich D. (1988). *Rasch models for measurement*, Sage, Beverly Hills.

Andrich D., Sheridan B., Lyne A. and Luo G. (2000). *RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models*. Perth, Australia: Murdoch University.

Anstey E. (1955). *Test de Dominò D-48*, Centre de Psychologie Appliquée, Paris.

Binet A., Simon T. (1905). Méthodes nouvelles pour le diagnostique du niveau intellectuel des anormaux. *L'Année Psychologique*, **11**, 191-245.

Boncori L. (1987). Secondo anno di sperimentazione di un servizio di orientamento universitario. *Orientamento Scolastico e Professionale*, **27**, 269-301.

Boncori L. (1993). *Teoria e tecniche dei test*, Bollati Boringhieri, Torino.

Bond T.G., Fox C.M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, Lawrence Erlbaum Associates, Mahwah, New Jersey.

Brown P. J. (1993). *Measurement, Regression and Calibration*, Oxford University Press, Oxford.

Bruni P. (1966). Uno studio sulle possibilità del D-48 nella scuola media. *Bollettino di Psicologia Applicata*, Agosto-Ottobre-Dicembre, **77-78**, 157-170.

Cattell J. McK. (1890). Mental tests and measurements. *Mind*, **15**, 373-380.

Cattell R. B. (1940). A culture free intelligence test. *Journal of Educational Psychology*, **31**, 161-169.

Crescentini A., Bonelli E., Giossi L. (2003). Revisione di una versione a scelta multipla del test D48. *III Incontro nazionale degli psicologi del lavoro e delle organizzazioni*, 30-31 maggio 2003, Roma, raccolta degli abstract, 27-28.

Cristante F., Borgatti P. (1975). Contributo alla taratura del test D-48. *Bollettino di Psicologia Applicata*, Febbraio, Aprile, Giugno, **127-128-129**, 107-116.

Csonka L. (1973). Norme per il test D 48 in base alla riuscita degli studenti genovesi. *Bollettino di Psicologia Applicata*, Agosto-Ottobre-Dicembre, **118-119-120**, 131-146.

De Battisti F., Nicolini G., Salini S. (2005). The Rasch model to measure service quality. *The ICFAI Journal of Services Marketing*, **III**(**3**), 58-80.

Domino G. (2001). The D-48—Application in Mexican American children of a culture fair test. *School-Psychology-International*, **22**(**3**), 253-257.

Les Editions du Centre de Psychologie Appliquées (2000). *Manuel d'application Test D 2000*. Les Editions du Centre de Psychologie Appliquées, Paris.

Galimberti G. (2000). *Dizionario di Psicologia*. UTET, Torino.

Gould S. J. (1996). *The Mismeasure of Man*. W.W. Norton & Company, New York.

Pichot P. (1949). *Les test mentaux en Psychiatrie: tome premier; Instruments et Méthodes*. Paris, Presses Universitaries de France.

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

Raven J. C. (1940). *Progressive Matrices.* H. K. Lewis, London.

Salini S. (2003). Taratura Statistica Multivariata. *Doctoral Thesis in Statistics – XV ciclo*, Università degli Studi di Milano – Bicocca.

Spearman C. (1904). General Intelligence, objectively determined and measured. *The American Journal of Psychology,* **15**, 201 - 293.

Spearman C. (1927). *The abilities of man*. Macmillan, New York.

Stevens S. (1951). *Handbook of Experimental Psychology*. Willey, New York.

Sundberg R. (1999). Multivariate calibration – direct and indirect regression methodology (with discussion). *Scandinavian Journal of Statistics*, **26**, 161-207.

Tesio L., Valsecchi M.R., Sala M., Guzzon P., Battaglia M.A. (2002). Level of Activity in Profound/Severe Mental Retardation (LAPMER): a Rasch-derived Scale of Disability. *Journal of Applied Measurement*, **3**, 50-84.

Test D 48 (1954). *Manuale di applicazione del reattivo.* Organizzazioni Speciali, Firenze.

Thurstone L.L. (1938). *Primary mental abilities*. University of Chicago Press, Chicago.

Vernon M. D. (1947). Different Types of Perceptual Abilities. *British Journal of Psychology*, **38**, 79-89.

Wright B.D., Linacre J.M. (1989). Observations are always ordinal; measurement, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, **70**, 857-860.

Wright B.D., Masters G.N. (1982). *Rating scale analysis*. MESA Press, Chicago.

Wright B.D., Stone M.H. (1979). *Best test design*. MESA Press, Chicago.

Appendix A

Table 1a. *Individual item fit*

Item	*Location*	*SE*	*Residual*	*ChiSq*	*Probability*
1	-3.233	0.24	-0.069	10,563	0.307
2	-4.073	0.36	0.069	7,226	0.614
3	-2.772	0.2	0.446	11,631	0.235
4	-2.273	0.16	1.946	*33,933*	*0.000*
5	-2.329	0.16	-0.732	12,903	0.167
6	-3.199	0.24	-0.01	4,425	0.881
7	-1.495	0.12	0.956	*22,405*	*0.008*
8	-2.166	0.15	0.226	13,857	0.128
9	-1.326	0.11	-1.139	*16,974*	*0.049*
10	-1.877	0.14	0.172	7,157	0.621
11	-2.184	0.15	-0.634	*18,037*	*0.035*
12	-2.260	0.16	-0.217	13,309	0.149
13	-2.833	0.2	-0.347	11,744	0.228
14	0.897	0.07	-0.826	*23,016*	*0.006*
15	0.936	0.07	-1.617	14,995	0.091
16	0.721	0.07	**-2.659**	*24,872*	*0.003*
17	0.689	0.07	**-2.379**	15,439	0.080
18	0.714	0.07	**-4.026**	*38,980*	*0.000*
19	0.619	0.07	**-4.023**	*33,005*	*0.000*
20	-0.952	0.1	-1.525	*20,586*	*0.015*
21	-1.270	0.11	-0.969	9,745	0.372
22	-0.094	0.08	**2.662**	*26,743*	*0.002*

23	0.953	0.07	**3.965**	*35,916*	*0.000*
24	2.021	0.08	**2.427**	*19,535*	*0.021*
25	2.031	0.08	1.256	7,677	0.567
26	0.773	0.07	**6.152**	*65,082*	*0.000*
27	-0.422	0.09	**-2.252**	15,320	0.083
28	-0.81	0.1	**-3.000**	*29,236*	*0.001*
29	-0.437	0.09	**-2.573**	*26,878*	*0.001*
30	-1.095	0.1	**-2.381**	*28,740*	*0.001*
31	0.401	0.07	-0.458	14,108	0.119
32	1.405	0.07	**-2.240**	*17,291*	*0.044*
33	0.54	0.07	-1.243	8,977	0.439
34	1.635	0.07	0.125	2,956	0.966
35	2.278	0.08	-0.329	*21,387*	*0.011*
36	1.219	0.07	-1.078	*17,694*	*0.039*
37	1.489	0.07	0.165	14,956	0.092
38	2.530	0.08	0.74	8,358	0.498
39	2.473	0.08	0.339	*20,173*	*0.017*
40	2.519	0.08	-1.693	10,683	0.298
41	2.527	0.08	0.908	*20,297*	*0.016*
42	2.290	0.08	0.725	15,602	0.076
43	2.655	0.08	0.122	15,371	0.081
44	2.785	0.09	0.973	*18,793*	*0.027*

Appendix B

Table 1b. *Individual person fit (for subjects with residual more than +/- 2).*

ID	*Locn*	*SE*	*Residual*
13	1.375	0.4	2.371
19	**0.266**	**0.4**	**2.529**
42	-1.27	0.43	-2.042
66	0.105	0.4	-2.063
68	-0.058	0.4	2.165
72	-0.222	0.41	2.428
76	**-0.908**	**0.42**	**-2.513**
93	-0.39	0.41	2.129
94	**-0.908**	**0.42**	**-2.513**
128	0.105	0.4	-2.200
143	0.105	0.4	2.197
184	1.875	0.42	2.100
186	-0.058	0.4	-2.466
208	**1.538**	**0.41**	**3.060**

212	-0.39	0.41	-2.308
218	**1.704**	**0.41**	**2.845**
254	-0.732	0.42	-2.244
257	0.266	0.4	-2.202
262	**-0.222**	**0.41**	**-2.924**
277	**-0.732**	**0.42**	**-2.701**
295	-0.222	0.41	2.280
352	0.425	0.4	2.046
355	**1.538**	**0.41**	**2.894**
409	**0.105**	**0.4**	**2.967**
422	-1.27	0.43	2.195
431	**0.583**	**0.4**	**2.537**
444	**0.105**	**0.4**	**-2.604**
465	**-0.559**	**0.41**	**-2.720**
550	**-0.058**	**0.4**	**-2.885**
566	0.425	0.4	-2.112
642	0.266	0.4	-2.062
651	0.74	0.4	2.327
668	1.214	0.4	2.413
676	1.055	0.4	2.227
685	0.266	0.4	-2.222
748	**0.105**	**0.4**	**-2.734**
768	-0.39	0.41	2.343
791	0.105	0.4	-2.224
835	**0.898**	**0.4**	**2.646**
841	**-1.087**	**0.42**	**-2.640**
857	**-0.058**	**0.4**	**-2.585**
874	0.105	0.4	-2.192
897	-0.222	0.41	-2.311
901	0.266	0.4	-2.190
916	**0.74**	**0.4**	**2.816**
933	1.055	0.4	2.095
949	-0.222	0.41	-2.267