Rapporto n. 13 2004
dmsia unibg.it

V. Caviezel

# Estimation for the Rasch Model under a linkage structure: a case study

*Serie Ricerca*

Dipartimento
di Matematica, Statistica,
Informatica e Applicazioni
"Lorenzo Mascheroni"

UNIVERSITÀ DEGLI STUDI DI BERGAMO

# ESTIMATON FOR THE RASCH MODEL UNDER A LINKAGE STRUCTURE: A CASE STUDY

Valeria Caviezel

*Abstract:* The purpose of this study is to measure the student's ability and the course's difficulty of a sample of students of the Faculty of Economics, University of Bergamo, using a Rasch measurement model. The problems of the linkage structure and of the choice of optimal categorization are discussed too.

# Introduction

In many disciplines, as psychology, medicine, sociology, sport and education sciences, it is very important to have un instrument for measuring some individual's characteristics: performance, ability, attitude, opinion, or problems and difficulties to make some exercises. It is important to asses and evaluate the tests and items proposed to subjects.

More specifically (Zhu, Timm, and Ainswoth, 2001), after developing a number of items with the predetermined response categories (e.g., Likert scale), a set of items or exercises is administered to the target sample. Based on subjects' responses to the items, item statistics (e.g., means and standard deviations) and personal measures (e.g., total score) were computed, and some sort of psychometric analysis was conducted to further evaluate the psychometric quality of the instrument. Several known psychometric problems, however, are related to this commonly used practice. Among others:

◦ The calibrations under the conventional procedure are often sample-dependent and item-dependent. Sample-dependent, in the context of the measurement, means that characteristics of an item, or instrument, are determined and based on the sample used in the study. By the same token, the characteristics of subjects are also determined by the type and the number of items included in a particular instrument.

◦ Items and subjects are calibrated on different scales. While the former are usually summarized based on means and standard deviations of the responses to individual items, the letter are often represented by total scores. As a result, it is difficult to judge whether a subject with a certain score will have a problem on a particular item.

◦ It is often incorrectly assumed in these studies that items with Likert scale are already set on a interval scale and that item responses are additive. Generally in these cases the items are based on the ordinal scale.

◦ When instrument developers choose a response category, they often assume that the category selected was already the most appropriate one. The number of categories and type of anchors, are known to have an effect on the categorization of a scale.

◦ It should be pointed out that the description of the attribute or trait being measured, as well as the characteristics of items, in previous studies assessment have been somewhat confusing.

The problems above described can be solved using the Rasch calibration. The Rasch calibration belongs to the response-centered calibration method, in which both examinees and testing items are located on a common continuum based on the amount of the trait possessed by each other. Theoretically, the Rasch calibration lies on the foundation of the item response theory, an advanced testing theory developed during the past five decades.

Rasch models are probabilistic mathematical models. Under Rasch models expectations (Conrad, and Smith, 2004), a person with higher ability always has a higher probability of endorsement or success on any item than a person with lower ability. Likewise, a more difficult item always has a lower probability of endorsement or success than a less difficult item, regardless of person ability.

◦ Rasch models require unidimensionality and result in additivity. Unidimensionality means that a single constuct is being measured. If the assessment contains multiple subscale, unidimensionality refers to the set of items for each subscale. Additivity refers to the properties of the measurement units, which are the same size (i.e., interval) over the entire continuum if the data fit the model. These units are called logits (logarithm of odds units) and are a linear function of the probability of obtaining a certain score or rating for a person of a given ability. These interval measures may be used in subsequent parametric statistical analysis that assume an interval level scale.

◦ The placement of items according to their difficulty or endorsability and persons according to their ability on the common logit scale is displayed in figure 1.

*Figure 1*: Example of persons' ability
and items' difficulty on the same axis.

In this figure the students' ability and the exams' difficulty are represented on the same axis. The logit scale (from "- 4" to "+ 4") is on the middle of the figure, on the left there are the students, classified in frequencies, and on the right the exams. At the bottom in figure 1 there are the students with less ability and the less difficult exams, on the top the students with more ability and the more difficult exams. It is possible to see that the less difficult exam is Company Administration (Az) and the more difficult exams are Computing Science (Inf) and Accounting and Auditing (Rag), at the same level there are Political Economics I (Pol I), Financial Mathematics (Fin), Mathematical Methods (Mat) and Statistics (Stat).

◦ The use of Rasch models enables predictions of how persons at each level of ability are expected to do on each item. This capability of having estimates for the item hierarchy and person ability levels enables us to detect anomalies, such as someone failing to endorse the 5 least severe (or easiest) items while endorsing the 5 most severe (hardest) items.

◦ To deal with these issues of unusual patterns or "misfitting" cases, once the parameters of the Rasch models are estimated, they are used to compute expected (predicted) response patterns for each person on each item. "Fit statistics" are then derived from a comparison of the expected patterns and the observed patterns. These "fit statistics" are used as a measure of the validity of the data-model fit.

◦ "Person fit" statistics measure the extent to which a person's pattern of responses to the items corresponds to that predicted by the model. A valid response requires that a person of a given ability have a greater probability of providing a higher rating on easier items than on more difficult items. Depending on the degree to which misfitting persons degrade the measurement system, one may elect to remove the misfitting from the calibration process, edit the misfitting response string, or choose to leave the misfitting persons in the data set.

◦ "Item fit" statistics are used to identify items that may not be contributing to a unitary scale or whose response depends on response to other items (i.e., a violation of local independence). The model require that an item have a greater probability of yielding a higher rating for persons with higher ability than for persons with lower ability. Those items identified as not fitting the Rasch model need to be examined and revised, eliminated, or possibly calibrated with other misfitting items to determine if a second coherent dimension may exist. There are many potential reasons an item may misfit. For example, an item may not be related to the rest of the scale or may simply be statistically redundant with the information provided by other items.

In summary, some advantages of Rasch models include the characteristic to equate responses from different sets of items intended to measure the same construct; to develop equal interval units of measurement if the data fit the model; to incorporate missing data by using estimation methods which rely on sufficient statistics and estimation methods that simply summarize the non-missing observations that are relevant to each parameter and compare them with their expectations; conducting validity and reliability assessments in one analysis for both item calibration and person measures; estimate person ability freed from the sampling distribution of the items attempted; estimating item difficulty freed from the sampling distribution of the sample employed; to express item calibrations and person measures on a common linear scale.

What every scientist and layman means by a "measure" (Wright, and Linacre, 1989; Kornetti et al., 2004) is a number with which arithmetic (and linear statistic) can be done, a number which can be added and subtracted, even multiplied and divided, and yet with results that maintain their numerical meaning. The original observations in any science are not yet measured in this sense. They cannot be measured because a measure implies the previous construction and maintenance of a calibrated measuring system with a well-defined origin and unit which has been shown to work well enough to be useful. The linear scales are an essential prerequisite to unequivocal statistical analysis. Something must be done with counts of observed events to build them into measures. A measurement system must be constructed from a related set of relevant counts and its coherence and utility established.

The valuation obtained by a judge is by construction an ordinal scale. The mark of an university exam is measured by a number (to "18" from "30 e lode" = "31"), but this is a judge, and then it is not measured by an interval or ratio scale, but by an ordinal scale. The mark is not obtained by a count or a measure's instrument. Although ordinal scales are often used for statistical analysis, equal interval data are fundamental for even basic mathematical operations; it is impossible to assert that the university marks are equal interval data.

In this study I considered a sample of students at the end of the first academy year and the judge (mark) obtained in eight exams. The latent variable I want to study, the student ability, cannot be measured by the marks obtained in the exams, because these are not determined on an interval scale. In particularly this scale is not additive; therefore the procedure leading to a total score as a sum of the partial scores (a final mark obtained as a sum of the individual marks, like in the “laurea” mark) is not a good procedure.

As described earlier, the Rasch analysis is a set of techniques and models for measuring a latent variable on an interval scale and to place on the same axe the subject’s (student) ability and the item’s (exam) difficulty (Waugh, 2003).

Therefore the most important purpose of this study is to obtain a meaningful valuation of the measure of the students ability and the exams difficulty by the Rasch analysis.

In this context it is important that the data fit the chosen model. For this purpose it is important to recategorize the marks in a little number of categories and determine the optimal categorization.

Categorization has always considered an important element in constructing an ordered-response scale (Zhu, Updyke, and Lewandowski, 1997). Ordered-response scales include scales having ordinal response categories. Categorization of an ordered-response scale has two very important characteristics. First, while all categories of a scale should measure a common trait or property (e.g., attitude, opinion, or ability), each of them must also have its own well-defined boundaries, and the elements in a category should also share certain exclusively specific properties. Second, categories must be in an order and numerical values generated from the categories which must reflect the degrees or magnitudes of the trait. An optimal categorization is the one that best exhibits these characteristics.

Moreover, once the optimal categorization was determined, it is possible to compare the studied situation with some similar situations, with those of later years (e.g, one or two years after) or with those in different towns or regions. In this way it is possible to observe if the optimal categorization is the same or not.

## Methods

Georg Rasch (Rasch, 1961) developed a mathematical model for constructing measures based on probabilistic relation between any item's difficulty and any person's ability. Rasch argued that the difference between these two measures should govern the probability of any person being successful on any particular item. The basic logic is simple: all persons have a higher probability of correctly answering easier items (e.g., to endorse the easier exam) and a lower probability of correctly answering more difficult items (e.g., to endorse the more difficult exam).

The simplest Rasch model, the dichotomous model, predicts the conditional probability of a binary outcome, given the person's ability and the item's difficulty. If correct answers are coded as 1 and the incorrect answers are coded as 0, the model expresses the probability of obtaining a correct answer as a function of the size of the difference between the ability of the subject $S_v$ *(v = 1, 2, …, n)* and the difficulty of item $I_i$ *(i = 1, 2, …, k)*.

This probability is given by:

$$P(X_{vi} = 1 | S_v, I_i) = \frac{\exp\{\theta_v - \beta_i\}}{1 + \exp\{\theta_v - \beta_i\}} \tag{1.a}$$

and then

$$P(X_{vi} = 0 | S_v, I_i) = \frac{1}{1 + \exp\{\theta_v - \beta_i\}} \tag{1.b}$$

where $\theta_v$ *(v = 1, 2, …, n)* is an uni-dimensional person parameter (person ability), and $\beta_i$ *(i = 1, 2, …, k)* is an uni-dimensional item parameter (item difficulty).

The odds of making response 1 instead of response 0 is:

$$ODDS = \frac{P(X_{vi} = 1 | S_v, I_i)}{P(X_{vi} = 0 | S_v, I_i)} = \exp\{\theta_v - \beta_i\}$$

and then its natural logarithm has the simple linear form:

$$Logit = Ln(ODDS) = \theta_v - \beta_i\,.$$

The characteristics of the Rasch model to compare persons and items directly means that we have created person-free measures and item item-free calibration; abstract measures that transcend specific persons' responses to specific item at a specific time. This characteristic is called parameter separation. Thus, Rasch measures represent a person's ability as independent of the specific test item, and item difficulty as independent of specific sample.

Let us consider, now, the responses of $n$ persons, $S_1, S_2, \ldots, S_n$ to a sequence of k items, $I_1, I_2, \ldots, I_k$, in which each subject may respond to item $I_i$ in $m_i+1$ ($m_i \geq 1$) ordered categories, $C_0, C_1, \ldots, C_{mi}$; for each item, the subject chooses one and only one of the $m_i+1$ categories. The categories' number can be different in the items.

The probability function is given by, following the partial credit model (PCM) (Master, 1982):

$$\pi_{vih} = P(X_{vih} = 1 | S_v, I_i) = \frac{\exp\{\theta_v h - \beta_{ih}\}}{\sum_{z=0}^{m_i} \exp\{\theta_v z - \beta_{iz}\}} \qquad h = 0, 1, \ldots, m_i \tag{2}$$

where $\theta_v$ $(v = 1, 2, \ldots, n)$ is an uni-dimensional person parameter, and $\beta_{ih}$ $(i = 1, 2, \ldots, k$ and $h = 0, 1, \ldots, m_i)$ is an uni-dimensional item parameter.

Formula (2) gives the probability - for a subject $S_v$, with person parameter $\theta_v$ - of scoring $h$ on item $I_i$ . By considering the couple of adjacent categories $C_{h-1}$ and $C_h$, the logit becomes:

$$Logit = \log \frac{\dfrac{\pi_{vih}}{\pi_{vih} + \pi_{vih-1}}}{1 - \dfrac{\pi_{vih}}{\pi_{vih} + \pi_{vih-1}}} = \theta_v - \delta_{ih}$$

$v = 1, 2, \ldots, n;\quad i = 1, 2, \ldots, k;\quad h = 1, \ldots, m_i$

where $\frac{\pi_{\text{vih}}}{\pi_{\text{vih}} + \pi_{\text{vih-1}}}$ is the probability that the subject $S_v$ for the item $I_i$ chooses the category $C_h$ rather than $C_{h-1}$, given that the response is only one between $C_h$ and $C_{h-1}$, and where $\delta_{ih} = \beta_{ih} - \beta_{ih-1}$ , $h = 1, ..., m_i$ and $\delta_{i0} = 0$.

To make the model identifiable, the constraints

$$\beta_{i0} = 0 \quad i = 1, 2, ..., k \quad \text{and} \quad \sum_{i=1}^{k} \sum_{h=0}^{m_i} \beta_{ih} = 0 \tag{3}$$

may be adopted.

By virtue (3), the formula (2) becomes

$$P(X_{vih} = 1 | S_v, I_i) = \frac{\exp\{\theta_v h - \beta_{ih}\}}{1 + \sum_{z=1}^{m_i} \exp\{\theta_v z - \beta_{iz}\}} \quad h = 0, 1, ..., m_i. \tag{2.a}$$

An equivalent expression for (2.a) is:

$$P(X_{vih} = 1 | S_v, I_i) = \frac{\exp\left\{\theta_v h + \sum_{l=0}^{h} \delta_{il}\right\}}{1 + \sum_{z=1}^{m_i} \exp\left\{\theta_v z + \sum_{l=0}^{h} \delta_{iz}\right\}} \quad h = 0, 1, ..., m_i;$$

where $\delta_{il}$ is referred to as uncentralized threshold parameter (Andrich's thresholds), and represents the magnitude of the supplementary difficulty from category $C_{h-1}$ to category $C_h$ for item $i$.

In both the dichotomous and the polytomous models the data matrix is a matrix with $n$ (n subjects) raws and $k$ (k items) columns. The raw score totals are ordinal-level, yet they are both necessary and sufficient for estimating person ability and item difficulty.

It is worth noting that to estimate the parameters with the maximum likelihood method the data matrix must not to be ill-conditioned. The data matrix is said to be ill-conditioned (Bertoli-Barsotti; Fischer, 1981) if there exists a partition (that may not be unique) of the set of the respondents into two non-empty subsets $G_1$ and $G_2$ such that if a subject belongs to $G_2$, his response score on $I_i$ $(i = 1, 2, \ldots, k)$ is not better than the response score on $I_i$ of any other subject in $G_1$.

As described earlier, the Rasch analysis was not originally developed for determining the optimal categorization, but rather as a measurement model. Only recently (Zhu, Updyke, and Lewandowski, 1997; Zhu, 2003) this model was proposed for identifying optimal categorization; information provided by the Rasch rating scale analysis, especially those on categories by the Rasch rating scale model, make it very useful for such a purpose.

Conceptually, the Rasch analysis belongs to a post-hoc approach in which the categories in the collected data can be recombined and the optimal categorization is determined and based upon a set of statistics provided by the Rasch analysis. Technically, the Rasch analysis starts by combining adjacent categories in a "collapsing" process, in which new categories are constructed. By comparing related statistical indexes, the optimal categorization can be determined. Three sets of statistics or parameter estimates are provided by the Rasch analysis, including model-data fit statistics, category statistics and parameter estimates and separation statistics. An optimal categorization, according to the Rasch analysis, should be the one that fits the Rasch model, has ordered categories (numerical values generated from the categories must reflect the increasing or decreasing trait to be measured), and leads to a greater discrimination among items and subjects (Zhu, Updyke, and Lewandowski, 1997; Linacre, 2003).

The procedure has been demonstrated as a useful means in determining the optimal categorization of an ordered-response scale (Zhu, Updyke, and Lewandowski, 1997).

The identified categorization based on the procedure, however, is merely the result of a post-hoc analysis. It is unknown if a modified categorization based on a Rasch post-hoc analysis could maintain its psychometric characteristics in the later measurement practice (Zhu, Updyke, and Lewandowski, 1997). More specifically, if, based on the categorization information provided by the Rasch analysis, a scale's optimal categorization was identified, could the revised scale maintain the psychometric characteristics of the original optimal categorization when it is applied to the same population?

The model-data fit statistics included two indexes: Infit and Outfit. The Infit statistic denotes the information-weighted mean-squares residual difference between observed and expected responses. The Outfit statistic, which is more sensitive to outliers and is used as an additional reference, denotes the usual unweighted mean-squares residual. Infit and Outfit, with a value of 1, are considered satisfactory model-data fit, and a greater value (e.g., > 1.3) or a smaller value (e.g. < 0.7) are considered a misfit. A greater value often indicates inconsistent performance, while a smaller value reflects too little variation.

The category statistics also included two indexes: average measure and Andrich's threshold. The average measures estimate approximatively the average ability of the respondents observed in a particular category, average across all occurrences of the category. The threshold, as described earlier, is the location parameter of the boundary on the continuum between category *k* and category *k-1* of a scale. A categorization, according to the categories statistics and parameter estimates, should be ordered, the basic property of the categorization in any ordered-response scale. If the thresholds are ordered, the categories used by survey participants were congruent with the intention of the scale designer (Piquero, MacIntosh, and Hickman, 2001).

The separation statistics, again, included two indexes: item and person separation (Zhu, Updyke, and Lewandowski, 1997; Zhu, Timm, and Ainsworth, 2001).

The item separation ($G_I$) is a measure used to describe how well the scale separates testing items:

$$G_I = \frac{SA_I}{SE_I}$$

where $SA_I$ is the item standard deviation and $SE_I$ is the root mean square calibration error of item.

The person separation ($G_P$), on the other hand, is a measure used to describe how well the scale identifies individual differences:

$$G_P = \frac{SA_P}{SE_P}$$

where $SA_P$ is the respondent standard deviation and $SE_P$ is the root mean square calibration error of respondents. The greater separation, the better the categorization, since the items will be better separated and the respondent's differences will be better distinguished.

Among commonly used conventional statistics, it is important to remember the coefficient Cronbach's Alpha. That is, perhaps, the most popular one at the scale level. Cronbach's Alpha is a measure of the internal consistency of a scale, and is a direct function of both the number of items in the scale and their magnitude of intercorrelation.

Therefore, either increasing the number of items or raising their intercorrelation can increase Cronbach's Alpha. Further, it is generally believed that increasing the number of categories will increase Cronbach's Alpha, but maximum gains will be reached with five or seven scale-points, after which Cronbach's Alpha values will level off.

Another commonly used conventional statistical index is the item point-biserial correlation coefficient (Zhu, Updyke, and Lewandowski, 1997), which reflects the correlation between responses and respondents' total scores. The point serial correlation coefficient is a discrimination index at the item level. Generally, the higher the point-biserial coefficient, the better the discrimination of an item, and a negative value often reveals a problematic item. While both coefficient Alpha and point-biserial coefficient may used to examine the quality of a scale or an item, neither provides any information on the quality of the categories.

Finally, the Rasch analysis, technically, starts by combining adjacent categories in a "collapsing" process, in which new categories are constructed.

Utilizing the collapsing process, parameter estimates and above mentioned goodness of fit, a new and useful post-hoc procedure based upon the Rasch analysis can be proposed to determine the optimal categorization empirically:

- ◦ Combine adjacent categories in a "collapsing" process, in which new categorizations are constructed;
- ◦ Select an appropriate Rasch model, applying the Rasch calibration, and examine the model-data fit;
- ◦ If the model-data fit is satisfactory, identifying the "candidates" of the optimal categorization whose categories are ordered;
- ◦ Determine the optimal categorization by selecting it from the "candidate" categorization exhibiting the greatest separation.

The purpose of this study was to find the optimal categorization for the marks (from "18" to "31") of a group of eight university exams.

# Data

*The data matrix*

For this study I considered the students of the Faculty of Economics, University of Bergamo, (enrolled in the academy year 2003/04) at the end of first academy year. This first year concerns the passing of 10 exams: Company Administration, Computing Science, Political Economics I, Political Economics II, Statistics, Financial Mathematics, Mathematical Methods, Accounting and Auditing, Private Law and Business English.

These two last exams have not been considered for a lot of reasons:

◦ Less of 100 students have not passed the exam of Private Law and so this exam couldn't be considered as a meaningful item.

◦ The Business English exam too, passed by few students, didn't give marks higher than "28". A null category (here "29", "30" and "31") poses some problems to the estimation of the parameters (see Bertoli-Barsotti, Fischer, 1981).

However, these two exams can be considered less important than others for an Economics University.

Afterwards I have chosen the 300 students who, at the moment of the analysis (October 2004), had passed at least four exams, the half of those concerned. The data matrix was formed by $n$ = 300 students and $k$ = 8 items. The frequencies for both each mark and exam are reported in table 1 (with "0" I have outlined the not passed exam).

Observing this table it is possible to note that Company Administration and Accounting and Auditing not have the maximum mark (31), Company Administration has overall high marks ("28", "29" and "30") and Accounting and Auditing the low marks. Several exams have not many students with mark "29" or "19", "20" and "21".

To have a clearer representation of the frequencies' distributions it is interesting to see the distribution functions, reported together in figure 2.

| | Company Administration | Computing Science | Political Economics II | Statistics | Political Economics I | Financial Mathematics | Mathematical Methods | Accounting and Auditing |
|---|---|---|---|---|---|---|---|---|
| **0** | 3 | 26 | 33 | 41 | 57 | 65 | 68 | 73 |
| **18** | 10 | 6 | 7 | 24 | 13 | 12 | 14 | 46 |
| **19** | 4 | 9 | 1 | 19 | 8 | 6 | 10 | 6 |
| **20** | 8 | 22 | 3 | 16 | 14 | 18 | 13 | 18 |
| **21** | 8 | 14 | 3 | 5 | 12 | 17 | 7 | 7 |
| **22** | 7 | 27 | 16 | 16 | 16 | 18 | 9 | 19 |
| **23** | 15 | 32 | 8 | 24 | 15 | 14 | 12 | 9 |
| **24** | 10 | 36 | 30 | 25 | 15 | 23 | 17 | 16 |
| **25** | 14 | 28 | 23 | 14 | 18 | 13 | 19 | 13 |
| **26** | 17 | 24 | 37 | 22 | 22 | 18 | 24 | 8 |
| **27** | 28 | 31 | 41 | 30 | 17 | 12 | 18 | 30 |
| **28** | 41 | 18 | 13 | 25 | 25 | 35 | 26 | 4 |
| **29** | 45 | 7 | 19 | 8 | 21 | 2 | 6 | 5 |
| **30** | 90 | 16 | 51 | 21 | 35 | 36 | 38 | 46 |
| **31** | 0 | 4 | 15 | 10 | 12 | 11 | 19 | 0 |

Table 1: Frequency distribution of marks for each exam.

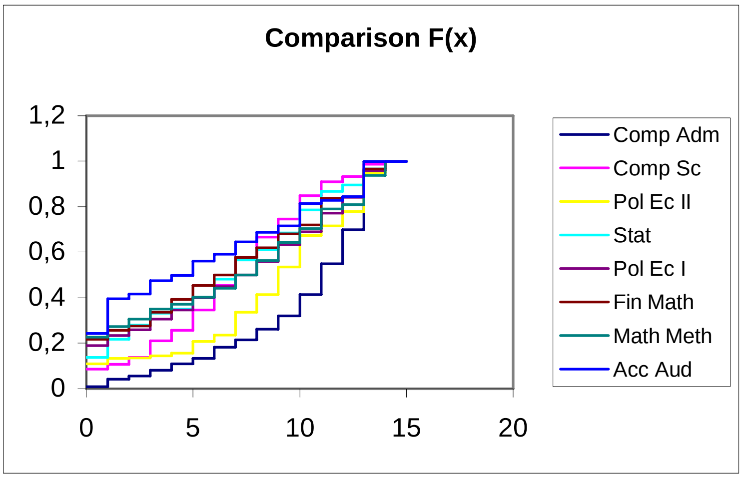



Figure 2: Comparison the exams' distribution function.

*The missing data*

In some data matrix it is possible to have the problem of missing data, because some cells of the matrix can be empty.

Generally there are multiple reasons for a non-response to an item. The non-responses can arise from a priori decision to not administer certain items or when respondents are directed to answer only relevant items represent conditions in which the missingness process may be ignored for purpose of estimating the person's location on the latent continuum of interest. In contrast, non-responses for "not-reached" items occur because a respondent has insufficient time to even consider responding to the items. Another source of missing data occurs because respondents have the capability of choosing not to respond to certain items. These intentionally omitted responses represent non ignorable missing data. This latter condition is referred to as missing not at random.

Different strategies have been developed for handling missing data (De Ayala, 2003) and were investigated for their capability to mitigate against the effect of omitted responses on person location estimation: ignoring the omitted response, selecting the “midpoint” response category, hot-decking, and a likelihood- based approach.

◦ Ignoring the omitted response had effect of reducing the number of items used for estimating the person’s location and thereby affecting the respondent’s sufficient statistics for location estimation. This strategy assumes that the omissions do not contain any useful information for estimating the respondent’s location.

◦ Replacing the omitted response with the “midpoint” response category (in effect, assuming the response is neutral like) does not reduce the number of items used in calculating the sufficient statistics. However, to the extent that this “neutral” response is not reflective of the respondent’s true response, so this approach may introduce additional measurement error.

◦ The hot-decking strategy selects a respondent (B) who is most similar to the respondent with missing response (A) in terms of the respondent’s string, but who has also answered the item that respondent A did not response to. Respondent B’s response to the item in question is used for respondent A’s omitted response to the item. If there are multiple matching candidates, then an individual was selected at random from the multiple matching candidates.

◦ In the likelihood approach the various possible responses are substituted for each omitted response and the likelihood of that response pattern is calculated conditional on the location estimate, $\hat{\vartheta}$, corresponding to the response vector’s sufficient statistic. For instance, let us say that the respondent has omitted one item and there are four possible response options (1, 2, 3, 4). In this approach the omitted response would be replaced a response of 1 and the likelihood based the corresponding sufficient statistic’s $\hat{\vartheta}$ calculated. Then the omitted response would be replaced by a response of 2 and likelihood recalculated and so forth for responses of 3 and 4. The $\vartheta$ associated with the largest of the four likelihoods was taken as the $\hat{\vartheta}$. Obviously, as the number of omissions increases the number of combinations of potential responses  also increases.

Now it is clear that it is not informative to compare the responses on two items A and B if these items have been administered to different groups (Van Buuren, and Hopman-Rock, 2001). Differences in the score distribution of A and B may be due to either differences between studies or to differences between items, or to a combination of both. However, if a third item C, that assesses the same trait, is measured in both studies, then the distribution of A and B can be compared through this common item.

Therefore, (Lee, 2003) to solve this problem there are two possible linkage structure: in figure 3.a is represented the linkage structure in which only some subjects responded to all items (horizontal linkage) and in figure 3.b is represented the linkage structure in which some items are administrated to all subjects (vertical linkage).

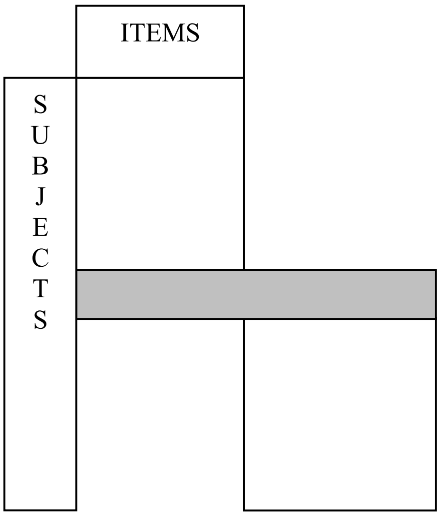



Figure 3.a: Horizontal linkage

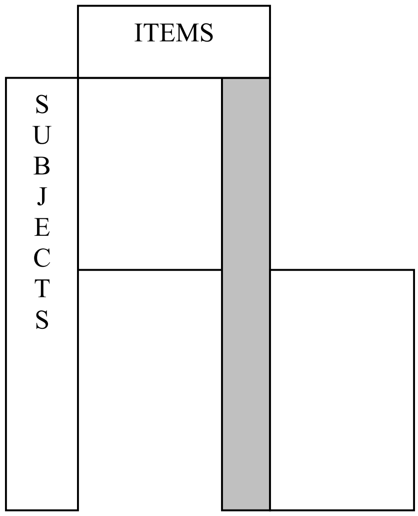



Figure 3.b: Vertical linkage

In this study there is the problem of “missing” data, because not all the 300 students passed the 8 exams, but only 118 students. For a number of reasons a student didn’t passed an exams: exam tried but failed, or unshown student, or others more. In any case in this study the missing response may be considered as a “wrong” response and then the respondent’s response vector doesn’t contain responses to each item.

In this data matrix there is not an exam passed by all students (see frequencies distribution in which all the exams have some “non passed”, minimum 3 cases in Company

Administration), but there are 118 students who passed all the eight exams; therefore it is possible to use the horizontal linkage.

*The categorization*

As the numerous individual categories (14 marks + not passed exam), I couldn't follow the "collapsing" process of adjacent categories, described by (Zhu, Updyke, and Lewandowski, 1997; Zhu, 2003), and so I tried to highlight some basic characteristics analysing the above table to determine the optimal categorization.

The non passed exam doesn't have to be considered like a missing data, because this data is not a very "missing", but not yet available information, due to a student's choice or an item too difficult for this student. In this context a not passed exam is a penalty, like a minimum mark, therefore the first category, coded by 0. This category can not be "collapsed" with the adjacent categories (marks "18", "19", …).

The marks "30" and "31" indicate greatest student's performance (ability), a perfect test, and then these categories together couldn't "collapse" with others indicating imperfect test. I think it is very important and meaningful to isolate the maximum marks.

To indicate the "collapsing" process of adjacent categories, I used this formalization (Zhu, Updyke, and Lewandowski, 1997; Zhu, 2003). For example, if the data analysis starts by recombining the original five adjacent categories (1, 2, 3, 4, 5) into three new categories, it is possible to obtain six "collapsing": 11123; 11233; 11223; 12223; 12233; and 12333. The expression "11123" means that the original category "1" was retained as "1", but the original categories "2" and "3" were collapsed into category "1", category "4" into category "2" and category "5" into category "3".

In this study the original categories are:

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

With 15 original categories, the number of new categories is explained in table 2, where:

k = New categories – 1,

r = 15 – New categories,

$\binom{k+r}{r}$ = number of possible combinations.

| New categories | k | r | $\binom{k+r}{r}$ |
|---|---|---|---|
| 2 | 1 | 13 | 14 |
| 3 | 2 | 12 | 91 |
| 4 | 3 | 11 | 364 |
| 5 | 4 | 10 | 1001 |
| 6 | 5 | 9 | 2002 |
| and so on | … | … | … |

Table 2: Number of possible combinations.

As described earlier, it is not feasible considering all the possible categorizations, and then, on base of the common experience, I chosen to analyze the following ones:

(1): 011111122233344 (2): 011111222333344 (3): 011111222233344

(4): 011112222333344 (5): 011112222233344 (6): 011122223334455

(7): 011222233344455 (8): 011111112222233 (9): 011111111222233

in which all the eight exams are classified by the same categorization; and these others:

(10): 011111222333344 and Accounting and Auditing: 011111112222233

(11): 011112222333344 and Company Administration: 011111112222233

(12): 011112222333344 and Company Administration: 011111111122233

in which an exam has a different categorization's number compared than others.

In these analyses I couldn't consider all the 300 students but 298, because two of them reported the maximum mark ("30" or "31") in the all exams and therefore the data matrix was ill-conditioned.

In the matrix the data are reported with growing column total and decreasing raw total, such that if the generic element of the matrix is $a_{vi}$ *($v = 1, 2, \ldots, n; i = 1, 2, \ldots, k$),*

$$r_v = \sum_{i=1}^{k} a_{vi} \qquad c_i = \sum_{v=1}^{n} a_{vi}$$

and then

$$r_1 \geq r_2 \geq \ldots \geq r_n \qquad \text{and} \qquad c_1 \leq c_2 \leq \ldots \leq c_k \ .$$

To analyze the data and determine the optimal categorization I used RUMM (Rasch Unidimensional Measurement Models) 2020.

RUMM 2020 is an interactive Rasch software package, which uses a variety of graphical and tabular displays to provide an immediate, rapid, overall appraisal of a analysis. This software is entirely interactive, from data entry to the various analysis, permitting rerunning analysis based on diagnosis of previous analysis, for example, rescoring items, eliminating items, carrying out test equating in both raw score and latent metrics.

RUMM 2020 handles 5000 or more items with the number of persons limited by available memory. It allows up to 9 distractor responses for multiple-choice items, a maximum of 64 thresholds per polytomous item. The software employs a range of special Template files for allowing the user customize analysis adding convenience and speed repeated, related and future analyses.

RUMM 2020 implements the Rasch models for dichotomous and polytomous data using a conditional estimation procedure that generalizes the equation for one pair of items in which the person parameter is eliminated to all pairs of items taken simultaneously. This procedure is conditional estimation in the sense that the person parameters are eliminated while the item parameters are estimated. The procedure generalizes naturally to handing missing data.

To estimate parameters RUMM 2020 uses a procedure based on the successive iterations until the convergence. The iteration is said to converge when the maximum difference in item and person value during successive iterations meets a preset convergence value.

## RESULTS OF THE OPTIMAL CATEGORIZATION

Implementing RUMM 2020 on the described categorizations (from 1 to 12) in all the cases, excepted one, the thresholds are disordered, for example as in figure 4.

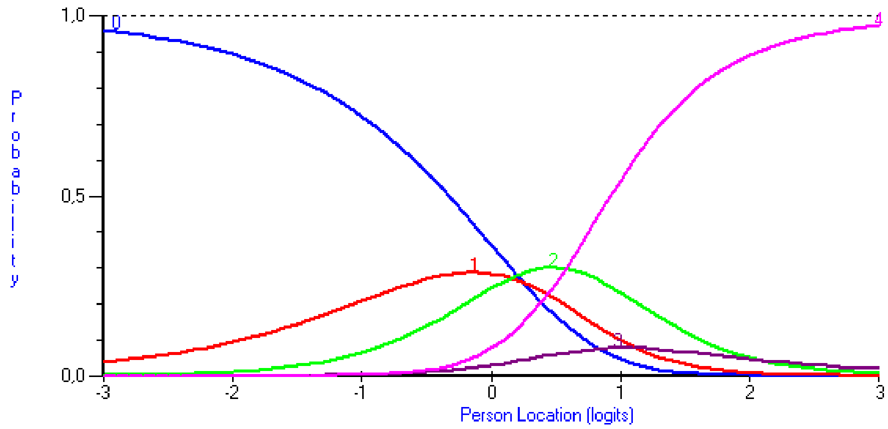



Figure 4: An example of disordered thresholds.

in which the categories 1 and 3 are never more probable than the categories 0, 2 and 4. Therefore, as written earlier, these categorizations are not optimal.

The categorization chosen as "optimal categorization" is: 011111112222233, in which:

| Category | Marks |
|---|---|
| 0 | Non passed exam |
| 1 | 18, 19, 20, 21, 22, 23, 24 |
| 2 | 25, 26, 27, 28, 29 |
| 3 | 30, 31 |

After this codification, the frequencies' distribution for each category is displayed in table 3.

Implementing RUMM 2020 with these data, all 24 parameters converged after 26 iterations.

For the chosen categorization all the thresholds (uncentralised) are ordered, see table 4.

| Item | Cat. 0 | Cat. 1 | Cat. 2 | Cat. 3 |
|---|---|---|---|---|
| Accounting and Auditing | 73 | 121 | 60 | 44 |
| Financial Mathematics | 65 | 108 | 80 | 45 |
| Statistics | 41 | 129 | 99 | 29 |
| Computing Science | 26 | 146 | 108 | 18 |
| Mathematical Methods | 68 | 82 | 93 | 55 |
| Political Economics I | 57 | 93 | 103 | 45 |
| Political Economics II | 33 | 68 | 133 | 64 |
| Company Administration | 3 | 62 | 145 | 88 |

Table 3: Frequency distribution for the optimal categorization.

| Item | Location=Mean | Threshold 1 | Threshold 2 | Threshold 3 |
|---|---|---|---|---|
| Accounting and Auditing (A A) | 0,356 | - 0,863 | 0,865 | 1,067 |
| Financial Mathematics (F A) | 0,278 | - 0,977 | 0,357 | 1,454 |
| Statistics (St) | 0,310 | - 1,616 | 0,307 | 2,240 |
| Computing Science (C S) | 0,290 | - 2,121 | 0,369 | 2,621 |
| Mathematical Methods (M M) | 0,131 | - 0,611 | - 0,210 | 1,213 |
| Political Economics I (P E I) | 0,172 | - 0,874 | - 0,207 | 1,597 |
| Political Economics II (P E II) | - 0,270 | - 1,134 | - 0,852 | 1,175 |
| Company Administration (C A) | - 1,267 | -3,443 | - 1,187 | 0,829 |

Table 4: Location parameters and thresholds for each exam.

and the Category Probability Curves are displayed for each item in figure 5. These eight figures show that all categories are more probable to emerge at different ability level. For example in Statistics it is possible to observe that for a logit lower than – 1,616 receiving 0 is more probable then receiving any other category; this indicates that students of low ability will have the greatest probably of not passing the exam. If the logit is between – 1,616 and 0,307, receiving a 1 is more probable than receiving any other category, and between 0,307 and 2,24 receiving a 2 is more probable. Only for a logit greater than 2,24 a student has a greatest probability to have a maximum mark.

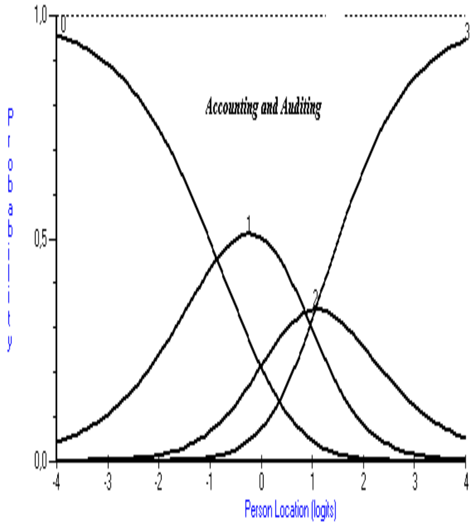
In figure 6 it is possible to observe the item map with uncentralised thresholds for each item: on the left there are the students frequencies for each class of logit and on the right there are the thresholds for each item.

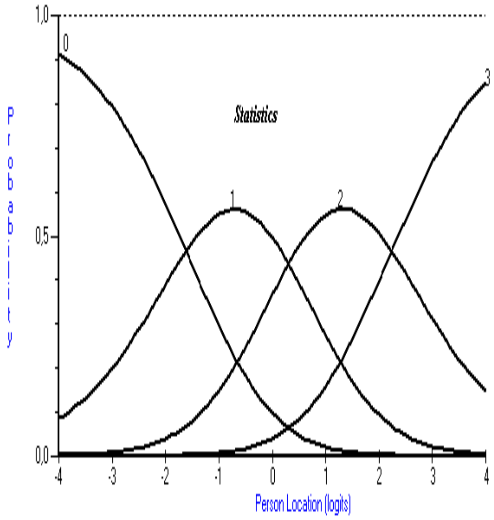
In the output of RUMM 2020 it is possible reading that:

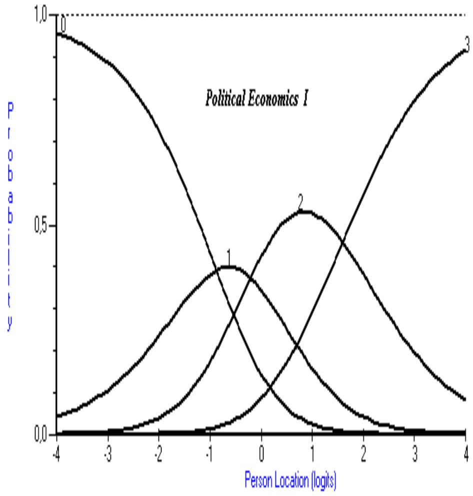
Cronbach Alpha = 0,784

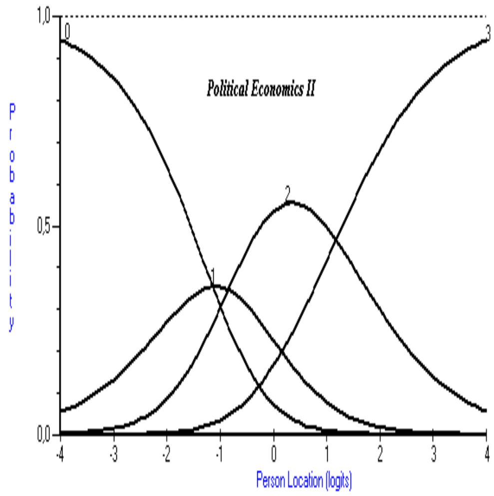
Person Separation Index = 0,770

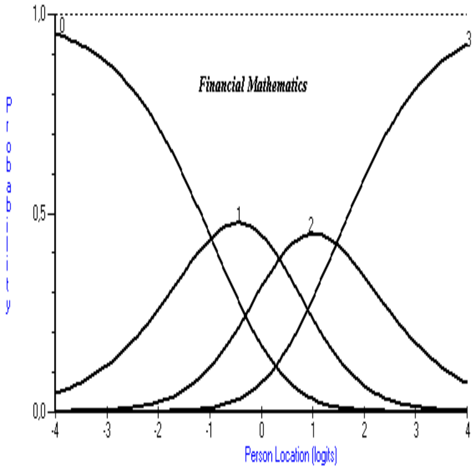
and the power of test-of-fit = GOOD.

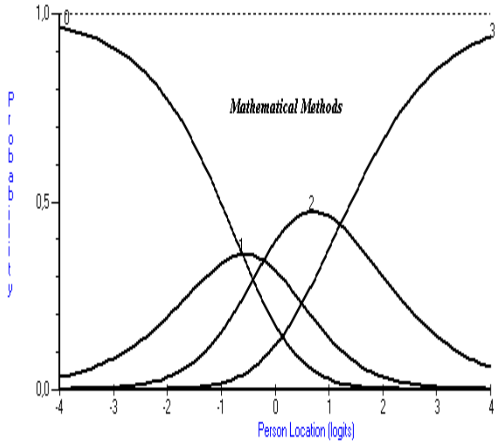


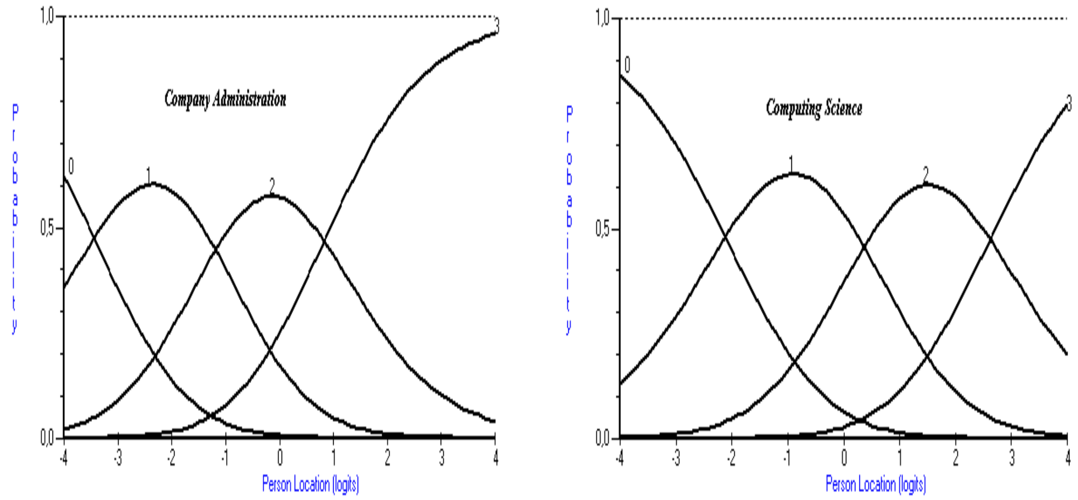



Figure 5: Category Probability Curves of the eight exams.

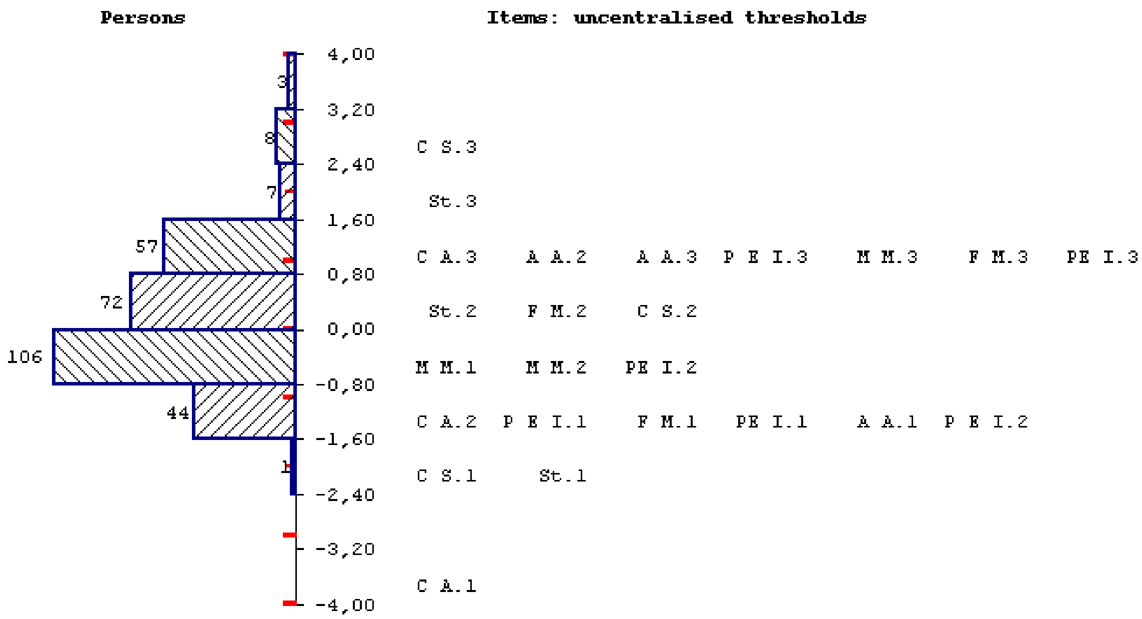



Figure 6: The item map.

# Conclusions

In many disciplines, as psychology, medicine, sociology, sport and education sciences, it is very important to have un instrument for measuring some individual's characteristics: performance, ability, attitude, opinion, or problems and difficulties to make some exercises, in any case a latent variable. It is important to asses and evaluate the tests and items proposed to subjects.

The Rasch analysis is a set of  techniques and models for measuring a latent variable on an interval scale and to place on the same axe the subject's ability and the item's difficulty. Under Rasch models expectations, a person with higher ability always has a higher probability of endorsement or success on any item than a person with lower ability. Likewise, a more difficult item always has a lower probability of endorsement or success than a less difficult item, regardless of person ability.

In this study I wanted to obtain a meaningful valuation of the measure of the students' ability and of the exams' difficulty by the Rasch analysis.

I considered the 300 students who, at the end of the first academy year, had passed at least four exams among Company Administration, Computing Science, Accounting and Auditing, Political Economics I, Political Economics II, Financial Mathematics, Mathematical Methods and Statistics.

Given that not all the 300 students passed the eight exams, I discussed the problem of the data matrix with some "missing " data; in this study these "missing" responses maid be considered as a "wrong" response, because a not passed exam is a penalty, a "minimum" mark for the student.

To analyse these data, I had to reduce the number of categories (the marks from "18" to "30 e lode"). By a "collapsing" process, the analysis of the thresholds and the calculation of some statistical index  permitted to obtain the optimal categorization, in which only four categories are considered.

## References


Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrica*, 43, 561-573.

Bertoli-Barsotti, L. On the existence and uniqueness of JML estimates for the partila credit model. *Psichometrika* (To appear).

Conrad, K. J., and Smith, E. V. (2004). Application of Rasch Analysis in health care. *Medical Care*, 42 (1 suppl.), I1-I6.

De Ayala, R. J. (2003). The effect of missing data on estimation a respondent's location using ratings data. *Journal of applied measurement*, 4 (1), 1-9.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psichometrika*, 46 (1), 59-77.

Kornetti, D. L., Fritz, S. L., Chiu, Y., Light, K. E., and Velozo, C. A. (2004). Rating scale analysis of the Berg balance scale. *Archives of Physical Medicine and Rehabilitation*, 85, 1128-1135.

Lee, O. K. (2003). Rasch simultaneous vertical equating for measuring reading growth. *Journal of Applied Measurement*, 4 (1), 10-23.

Linacre, J. M. (2003). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3, 85-106.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Piquero, A. R., MacIntosh, R., and Hickman, M. (2001). Applying Rasch modelling to the validity of a control balance scale. *Journal of Criminal Justice*, 29, 493-505.

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

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Theory of Probability* (Vol. IV, pp. 321-333). Berkeley: University of California Press.

Van Buuren, S., Hopman-Rock, M. (2001). Revision of the ICIDH severity of disabilities scale by data linking and item response theory. *Statistics in Medicine*, 20, 1061-1076.

Waugh, R. F. (2003). Evaluation of quality of student experiences at a university using a Rasch measurement model. *Studies in Educational Evaluation*, 29, 145-168.

Wright, B. D., and Linacre, J. M. (1989). Observations are always ordinal; measurement, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, 70 (12), 857-860.

Zhu, W. (2003). A confirmatory study of Rasch-Based optimal categorization of a rating-scale. *Journal of Applied Measurement*, 3, 1-15.

Zhu, W., Timm, G., and Ainsworth, B. (2001). Rasch calibration and optimal categorization of an instrument measuring women's exercise perseverance and barriers. *Research Quarterly for Exercise and Sport*, 72, 104-116.

Zhu, W., Updyke, W. F., and Lewandowski, C. (1997). Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. *Journal of Outcome Measurement*, 1, 286-304.