



RASCH ANALYSIS OF SOME MMPI-2 SCALES IN A SAMPLE OF UNIVERSITY FRESHMEN

Enrico Gori

University of Udine, Italy

Raffaella Fiormaria Marin

Non-cognitive skills have been recognized important predictors of student's outcomes. The aim of this paper is to examine the psychometric properties of the Mt, DEP and ANX scales from Minnesota Multiphasic Personality Inventory, through the modern test theory approach represented by Rasch analysis; to evaluate the correlation between these traits, and between students and parents, using the data collected on a sample of Italian university freshmen and their families.

Keywords: Rasch models, Mt, DEP and ANX scales, MMPI-2, Validity.

Introduction

The prediction of student's success and the identification of factors that can affect such phenomenon is a main topic in school effectiveness and human capital studies (Gori, 2004). Beside cognitive skills, such as "verbal, reading, and writing abilities as well as those in mathematics, science, music, and art" (Farkas, 2003, p. 543), also non-cognitive skills are recognized important factors that can explain student's careers, at all levels (Heckman & Rubinstein, 2001). Non-cognitive skills are more vaguely specified as motivational and personality traits such as hard work, conscientiousness, self-discipline, determination, and the way individuals think and feel about themselves in terms of self-concept, self-esteem and self-efficacy (Borghans *et al.*, 2008). Many studies (Carneiro & Heckman (2003), Cunha *et al.* (2002) and Heckman & Masterov (2004)) moreover found that parents play an important role in developing both the cognitive and non-cognitive skills of their children. These skills typically belong to the area of psychometrics and the interest in this field is growing in professional psychology. It is for such reason that well established psychometric test such as Minnesota Multiphasic Personality Inventory (MMPI-2) are widely used, especially at college level, to identify students in need of intervention and counseling, and results of such tests are employed to verify their prediction validity regarding student outcome and career. The aim of this paper is to consider three scales belonging to MMPI-2, which have been recently recognized as important predictors of student's outcomes (Lauterbach *et al.*, 2002; Bethune, 2011). In particular, the Mt Scale, with 41 items, originally designed to identify college students classified as maladjusted, the DEP Scale, with 33 items, designed to measure depression, and the ANX scale, with 23 items, designed to measure anxiety. They have many items in common. High point scores on the Mt Scale were significantly associated with lower grade point average, arriving late to classes more often, post-traumatic stress disorder symptoms of avoidance and arousal, and a history of treatment (Lauterbach *et al.*, 2002). Other work (Bethune, 2011) found that some scales of the MMPI-2, such as depression, had

incremental and independent effects on graduate grade point average and internship evaluation ratings, after controlling for traditional admissions criteria based on cognitive factors. These results imply that personality characteristics are relevant to academic performance indicating that objective measures of personality can be useful predictors of academic outcomes. As Poropat (2009), who conducted the most comprehensive meta-analytic investigation of personality-academic performance relationships concluded, “Personality should take a more prominent place in future theories of academic performance and not merely as an adjunct to intelligence” (p. 333). The aim of this paper is: (a) to examine the psychometric properties of the Mt, DEP and ANX scales in a sample of university freshmen and their families, through modern test theory approach represented by Rasch analysis, and (b) to analyze the correlation of these traits between students and parents. The use of Rasch analysis presents many advantage with respect to the Classical Test theory on which the MMPI has been originally developed. Rasch analysis is strongly founded from the point of view of measurement theory and represents a powerful construct validation tool, allows calculating standard errors and can easily deal with missing data. Moreover, the standard errors of measurements estimates allows to correct correlations and coefficients when such estimates are used as explanatory variables in classical regression and multilevel models (Battauz *et al.*, 2011).

Methods

The MMPI-2

The Minnesota Multiphasic Personality Inventory (MMPI) is the most widely used and researched standardized psychometric test of adult personality and psychopathology (Camara *et al.*, 2000). Psychologists and other mental health professionals use various versions of the MMPI to develop treatment plans; assist with differential diagnosis; help answer legal questions (forensic psychology); screen job candidates during the personnel selection process; or as part of a therapeutic assessment procedure (Butcher & Williams, 2009). The original MMPI, first published by the University of Minnesota Press in 1943, was replaced by an updated version, the MMPI-2, in 1989. A version for adolescents, the MMPI-A, was published in 1992. An alternative version of the test, the MMPI-2 Restructured Form (MMPI-2-RF), published in 2008, retains some aspects of the traditional MMPI assessment strategy, but adopts a different theoretical approach to personality test development. The MMPI-2 was the first major revision of the MMPI, which was standardized on a new national sample of adults in the United States and released in 1989 (Butcher *et al.*, 1989). The new standardization was based on 2,600 individuals from a more representative background than the MMPI (Gregory, 2007). It is appropriate for use with adults 18 and over. Subsequent revisions of certain test elements have been published and a wide variety of subscales was introduced over many years to help clinicians interpret the results of the original clinical scales. The current MMPI-2 has 567 items, and usually takes between one and two hours to complete depending on reading level. It is designed to require a sixth-grade reading level (Gregory, 2007). The original form of the MMPI-2 is the third most frequently utilized test in the field of psychology, behind the most used IQ and achievement tests. The 567 items are used to compose several scales divided in 4 main groups: Basic, Subscales, Supplementary, and Content. The items are classified as 0/1 if FALSE/TRUE or vice versa, depending the scale they belong. The total number of 1 for every scale determine its score S . A transformation called Uniform T-score, defined as $T = 50 + [10 \cdot (S - \mu)] / \sigma$ allows to obtain numbers that are directly comparable with respect to the percentiles of a population taken as standard, with mean μ and standard deviation σ . The standardization is different for male and female. In particular, the T-score may be interpreted with the aid of the table 6. In this work, we have employed the MMPI-2 Italian version; the item belonging to the three scales of interest here are reported in table 6.

The Mt Scale

The College Maladjustment (Mt) Scale is a 41-item Supplementary scale of the MMPI-2, introduced by Kleinmuntz (1960), to discriminate between psychologically adjusted and maladjusted college students. Items were selected from the MMPI item pool by comparing responses of 40 adjusted and 40 maladjusted male and female students. The adjusted students had contacted a university clinic to arrange a routine mental health screening examination as part of teacher certification procedures; none reported a history of psychiatric treatment. The maladjusted students had contacted the same clinic for assistance with emotional disorders and had remained in psychotherapy for three or more sessions. Item analyses yielded 43 items that discriminated between emotionally adjusted and maladjusted students (Kleinmuntz, 1961). The scale was revised for the MMPI-2 (Butcher *et al.*, 1989) and includes 41 of the original 43 items. The scores on the Mt scale are related to general maladjustment as assessed by psychiatric diagnosis, mental health treatment seeking, or scores on other maladjustment scales (Kleinmuntz, 1961; Kuczka & Handal, 1990; Stewart, 1994; Svanum & Ehrmann, 1993). High scores (T-scores > 65): identify students in need of treatment; psychosomatic complaints, poor concentration, lethargy, depression, lack of self-confidence, irritable, overly sensitive, sleep disturbance and anxiety. Low scores (T-scores < 40): identify students motivated, energetic, optimistic, self-confident and with good judgment. This scale correlates .90 with D-O, .90 with D4 Mental Dullness, .89 with Welsh Anxiety, and .89 with D1 Subjective Depression. Very little research has been conducted on the Mt Scale in the 50 years since its inception. Barthlow *et al.* (2004) analysed the construct validity of the Mt Scale using 376 student clients at a university psychological clinic: "A principal components analysis and correlations of Mt scale scores with clients' and therapists' ratings of symptoms and functioning showed that the Mt scale identifies the presence of maladjustment as defined in terms of depressive and anxious symptoms. There is no evidence to show that the scale is specific to college students or that it is sensitive to severe psychological disturbance". Merker & Smith (2001) validate the Mt Scale by comparing it with the Student Adaptation to College Questionnaire (SACQ): "Significant negative correlations existed between the Mt scale and SACQ scores, indicating that the Mt scale measures maladjustment, especially emotional maladjustment, in college students".

The DEP and ANX scales

DEP e ANX belong to Content Scales (Butcher *et al.*, 1989). These scales were developed through a combined rational approach with some empirical refining. All the scales have items, which are obvious in content, and measure what the respondent wishes to communicate. In particular:

ANX Anxiety - (23 items). High: Tension, worry, fears of losing one's mind, lack of confidence, somatic indications of anxiety such as heart pounding, shortness of breath, and disturbed sleep. Correlates .82 Welsh's Anxiety, .82 with Pt and .82 with Wiggins DEP.

DEP Depression - (33 items). High: Severe or major depression, brooding, crying easily, pessimism, suicidal ideation, guilt, remorse, overly sensitive, apathy, feeling worthless, unresolved object loss, and feeling empty. Correlates .90 with Wiggins DEP, .84 Welsh's Anxiety, and .82 with Pt.

The RASCH Models

Rasch Models as Basis for Fundamental Measurement

Campbell (1920) showed that scientific measurement requires an ordering system and the kind of additivity illustrated by physical concatenation (<http://www.rasch.org/rmt/rmt21b.htm>). Campbell called this "fundamental measurement." Thurstone (1927), with his Law of Comparative Judgment produced results that are successful instances of fundamental measurement. The concept of order and additivity

recurs in Guilford's (1936) definition of measurement. The main consequence of additivity is the maintenance of the unit of measurement and hence of the invariance of comparisons of measures across the scale. Guttman (1950) formulated a criterion for judging whether data were good enough to build a scale. The data must demonstrate a joint order shared by items and persons. The Danish mathematician Georg Rasch (1960) found that he could obtain an invariance of test item characteristics over variations in persons only if the function through which persons and items interact has linear form (Rasch, 1960, p. 120). Such property is known as specific objectivity: the comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion. Rasch showed that invariance could be maintained only when data satisfy a probability response model:

$$(1) \quad \text{Dichotomous Rasch model: } \ln \left(\frac{P(X_{ij} = 1)}{P(X_{ij} = 0)} \right) = \alpha_i - \beta_j, \quad X_{ij} \in \{0,1\},$$

where X_{ij} is the response of person i to item j , α_i is the ability" of the person (level of the latent trait), and β_j is the difficulty of the item (expressed on the same scale of the latent trait), that produce a joint order of response probabilities similar to the joint order defined by Guttman (Rasch, 1960, p. 117). Rasch discovered that the minimally sufficient statistics from which to estimate person and item measures were simply the unweighted sums of right answers for persons and for items, which is the score. Later on Luce and Tukey (1964) introduce the concept of "conjoint measurement" and showed that it could produce results that satisfy Campbell's fundamental measurement: in particular, the way to produce such kind of measurement is to gather data (items and persons) such that the 'effects' of different factors are additive (p. 4). The Rasch's models can do exactly this as shown by Perline *et al.* (1979). Subsequently Andersen (1977) showed that the sufficient statistics, which allows the Rasch model for dichotomously scored items to produce fundamental measurement, may be extended to response formats with more than two ordered categories (Andrich, 1978; Wright & Masters, 1982):

$$(2) \quad \text{Rating Scale model: } \ln \left(\frac{P(X_{ij} = k)}{P(X_{ij} = k-1)} \right) = \alpha_i - \beta_j - \tau_k, \quad X_{ij} \in \{0,1,2 \dots K\},$$

where τ_k is a "threshold" that measure the difficulty to reach category k , identical for every item

$$(3) \quad \text{Partial Credit model: } \ln \left(\frac{P(X_{ij} = k)}{P(X_{ij} = k-1)} \right) = \alpha_i - \beta_j - \tau_{jk}, \quad X_{ij} \in \{0,1,2 \dots K\}$$

where τ_{jk} is a "threshold" that measure the difficulty to reach category k for the item j .

Therefore, the Rasch models are the formal measurement model required to construct fundamental measurement from dichotomous or ordinal data. In particular, the matrix of expected response probabilities table derived from any set of Rasch item and person estimates will satisfy fundamental measurement axioms as defined by Guttman scaling (1950). Because Rasch models satisfy fundamental measurement axioms, the key point in any application is if data adhere "sufficiently" to the model. Although the procedures for determining whether the matrix of actual response frequencies adheres sufficiently to the Rasch expected response probabilities are still an open problem (Bond, 2003; Linacre, 2009) the possibility of this comparison allows, in principle, to reject the theory on which data are collected and items constructed: that is the theory relative to the problem considered (construction of an ability scale, depression scale and so on). This allows what Carl Popper (1934) calls "falsifiability" of a

theory, a fundamental requisite to build scientific theories. Under this respect Rasch models are scientific twice: firstly because they are the only one that satisfy fundamental measurement axioms and secondly because they provide a way (comparison between actual frequencies and expected response probabilities) to falsify the theory used to build the measurement. Therefore we may say that other approaches to construct measurement are not scientific: the axioms of fundamental measurement are routinely violated for example in IRT models (Karabatsos, 1999a, b; 2000), and True Score model (Allen & Yen, 2002) is not falsifiable.

Rasch Analysis as a Falsification Process

Thus the process of Rasch analysis mainly relates with testing to see if the data accord to model expectations, satisfy the various assumptions of the model, and other key measurement issues such as local independence, unidimensionality and the absence of differential item functioning. First we may look at the point-measure correlations (Pearson correlation coefficients, between the observations and the measures, estimated from the raw scores), indicating how much the responses to each item within a measure are correlated with the overall measure. These are particularly useful in a preliminary phase to check eventual coding errors. Zero or negative point-measure correlation means that the responses to the item contradict the latent variable defined by the consensus of the other items. In this case, the item needs to be omitted, or rescored in the opposite direction, before to go on in the analysis. We use moreover point measure correlations in conjunction to item fit statistics (see below) to verify that our measures only include items that are measuring the degree to which people endorse the single, underlying concept of interest. A second important thing to check, when we use a Rating Scale or Partial Credit model for observations assuming categories 0, 1, 2..., is that the thresholds, marking off successive categories, need to be ordered to be interpretable. However, in estimating the thresholds from the data, it is possible to discover that the estimates are not properly ordered. This is a sign that the categories are not working as intended and an anomaly in the data that needs to be understood and corrected is disclosed (<http://www.rasch.org/rmt/rmt151j.htm>). Often the problem can be solved reducing the categories. Then we may proceed to verify the assumption of local independence comprising two elements, response dependency and trait dependency (Marais & Andrich 2008). The former is where items are linked in some way, such as a items reflecting the same content. The latter is multidimensionality. Both these are tested by analysis of the residuals where the former is judged to be absent when residual correlations are below 0.3, and the latter to be unidimensional where patterns of items in the residuals (as identified by a Principal Component Analysis - PCA) are shown to give similar person estimates (Smith, 2002), or the residual variance explained by the first component is lower than some critical level (around 2 in most of the cases)) identified by simulation studies on unidimensional Rasch datasets (<http://www.rasch.org/rmt/rmt233f.htm>, <http://www.rasch.org/rmt/rmt191h.htm>, Brentari & Golia, 2007). Moreover, unidimensionality is excluded if comparison of items with large positive and negative factor loadings do not identify cluster with clearly different meaning (<http://www.winsteps.com/winman/principalcomponents.htm>). Response dependency can be accommodated by dropping similar items. Multidimensionality problems can be solved splitting the items in different analysis. A final check is based on fit indices, to verify the assumption of stochastic ordering of items, so testing the probabilistic Guttman pattern. To this end, we calculate an expected response, under the Rasch model, for each person to each item, and produces fit statistics indicating the degree to which people and items are acting in accordance with expectation. The difference between the expected response and the observed response for that person is the residual. The person fit statistic is then just an aggregation of all the residuals from all items for that person. Analogously, the item fit statistic is calculated from an aggregate of the residuals for all people to that item. The fit statistic we use, the mean-squared residual, varies between 0 and infinity: the weighted version is called INFIT, the unweighted version is called OUTFIT (<http://www.rasch.org/rmt/rmt162f.htm>). It can be interpreted as the ratio of the observed variation in the responses to the expected variation. Values greater than 1.0 indicate more than expected variation. We

generally treat values less than 1.40 (40% more variation than expected) as acceptable, but there is always some leeway and room for interpretation (Bond & Fox, 2007, p. 243). Values less than 1.0 indicate less than expected variation (item over fitting). This is not necessarily a bad situation, and is less problematic than the case where the fit value are greater than 1.4 (under fitting). A person with a poor fit statistic is likely someone who has responded randomly. We are less likely to believe a person measure with a large misfit statistic. We can inflate the standard error of the person to reflect this uncertainty. Moreover items fit indices helps determine whether there are items measuring a concept other than the one being assessed by the remaining items in that measure, indicating that we should perhaps reject the presumption of unidimensionality. Indeed PCA is much more sensitive to a large number of misfitting items: a small number of items related to a second dimension inflate slightly the percentage of variance explained by first PCA residual component, while fit indices are more sensitive to such situation (Brentari & Golia, 2007). The hypothesis of misfit (fit indices different from 1) can be moreover tested using normal standard or t-student like statistics (in case of small number of observations): normalized fit indices are expected to be within the range ± 2.5 . By removing under fitting items, we can clearly increase the internal validity of our measures and contribute to satisfy unidimensionality assumption. A final check should be that of invariance. This requires the absence of Differential Item Functioning (DIF) that is no significant difference in the residuals (via ANOVA) across key contextual groups, such as age or gender: to this extent, we may use some test statistics such as Rasch-Welch and Mantel-Haenszel (<http://www.rasch.org/rmt/rmt42f.htm>). A recommendation is the following: when contrast groups differ in measurement level is worth to construct measurement-matched samples of the largest possible size before to test DIF. In this case, Rasch-Welch is more sensitive to DIF than Mantel-Haenszel test and at least as reliable. In the analysis that follows similar results for DIF were observed using matched and not matched (original) samples. A final check for invariance can be done splitting the sample in two parts, one composed of the persons with highest measure, the other composed of persons with the lowest measure: invariance would require that the estimates of items measure using on the two samples should be the same unless of an additive constant. Given that Rasch measurement instantiates interval level, rather than ratio level measurement, invariance of item and person estimate values remains relative. Individual Rasch analyses (by default) adopt the mean of the item difficulty estimates as the zero point on the calibration scale, so it is the differences between item (or person) estimates, rather than those estimates per se which should remain invariant across investigations (Bond, 2003, p. 181). This means that invariance of items estimates across samples could be verified if these estimates lie within opportune confidence bands around a 45° straight line (Bond & Fox, 2007, p. 72), called “Identity line”.

The Main Results of a Rasch Analysis

The main results from a set of item-persons satisfying the Rasch model are the following.

Person, item and threshold measures are linear on a logit scale.

Person measures α_i , generally called “abilities”, and item measures β_j , generally called “difficulties”, are placed on the same linear, unbounded logit scale; the reason and the origin of the “logit” name, is that this measures come from a log-odds transformation of the probabilities (1): the odd is $Odds = \frac{p}{1-p}$, which ranges from 0 to $+\infty$ (bounded at the bottom). Taking the natural logarithm of the odds gives the logit, which is on a scale of $(-\infty, +\infty)$ (completely unbounded). If a Rating (2) or partial Credit model (3) is used, also the thresholds (τ_k or τ_{jk}) are placed on the same scale. Survey item response categories, instead, are not linear. For example, the difference between “strongly disagree” and “disagree” on a 4-point Likert scale is not necessarily the same as the difference between “disagree” and “agree” and so it would be a mistake to treat the category codes as numbers and do arithmetic with them, as true score

measurement models do. Instead, Rasch creates linear measures from the counts of responses: a run of the Winsteps program, for example, provide estimates of the linear parameters α_i , β_j and τ_k or τ_{jk} . The fact that Rasch places both person measures and item difficulties on the same scale allow us to directly make inferences about a person's performance relative to the scale of items. For example, we can say that a person with a measure of 1.0 is expected to agree with the items with measure much lower than 1.0 and disagree with the items much greater than 1.0. This enables us to concretely characterize a person's attitude more meaningfully than just saying, "her measure was 1.0". Therefore, although the logit measures are not meaningful in themselves, being able to state expected responses enables us to concretely describe any measure value.

Person, item and threshold measures have a standard error

The precision of each person, item and threshold measure (standard error), is derived from the estimation process. The person standard error, in particular, depends on how many items a person responds to (more data lead to better precision), and how extreme the measures are (measures that are very low or very high will have the least precision). Calculated on a large number of data, persons who are near the average item difficulty (where item are in generally more numerous) will have low standard errors, meaning that we are more confident that the measure we have calculated for that person is very close to the person's actual ability. The impact of large person standard errors can be reduced in two ways: (a) increase the precision estimates of person measures by adding more items or thresholds for items and/or (b) use these person standard errors to adjust correlations, and regression coefficients in regression analysis, for variation in measurement error. The average squared person standard errors (variance of measurement error) also contribute to the calculation of the reliability of the measure i.e. the ratio between the (measure variance) and (measure variance + variance of measurement errors): the lower the standard errors, the higher the reliability. True score models do not give individual standard errors but only overall standard errors, or average standard errors.

Results of the model can be used to validate the construct

Rasch model provides a way to verify construct validity, "*the degree to which a test measures what it claims, or purports, to be measuring*" (Cronbach & Meehl, 1955) and the six aspects of construct validity in Messick's Unified Theory of Construct Validity (Messick, 1995):

- A. *Consequential* - What are the potential risks if the scores are, in actuality, invalid or inappropriately interpreted? Is the test still worthwhile given the risks?
- B. *Content* - Do test items appear to be measuring the construct of interest?
- C. *Substantive* - Is the theoretical foundation underlying the construct of interest sound?
- D. *Structural* - Do the interrelationships of dimensions measured by the test correlate with the construct of interest and test scores?
- E. *External* - Does the test have convergent, discriminant, and predictive qualities?
- F. *Generalizability* - Does the test generalize across different groups, settings and tasks?

The analysis of fit help us to highlight the reason way some items are retained in the scale and other are eliminated because their content is not coherent with the unidimensional construct defined by the others (criterion A). Once unidimensionality of measures is established (i.e. "not rejected" by the Rasch model falsification process), we are measuring people and items on a single unitary concept on which they exhibit more or less of the construct of interest. Therefore, items manifest differing degrees of that construct, and then all together form a scale. The order of items along their difficulty allows us to ascertain whether the items included in a particular scale match the conceptual difficulty of the items. The alignment of item difficulties with conceptual difficulties indicates that we are measuring what we really want to measure (criterion B, C and D). The analysis of the thresholds may bring some information in the validation process indicating the reason why some items can be observed at ordinal scale while for other a dichotomous scale is more convenient (criterion B). The analysis of DIF help to find the reason way some

items are more easy for some group and more difficult for others, confirming in this way the that the empirical scale adhere to the conceptual theory regarding the construct, and the analysis of invariance help to asses if the test can be generalized across different groups and settings (criterion F). The linearity of the constructed measures with their standard errors help to analyze the correlation with other phenomena and therefore the predictive validity of the test (criterion E).

Data, Sample and Rasch Analysis of the Mt, DEP and ANX Scales

The MMPI-2 was self administered to a sample of 229 persons (Table 1), 134 Female and 95 Male: 139 units were freshman enrolled in the first year statistics course faculty of Economics - Udine University (Italy) and 90 were their parents, involved in the research by their sons. The average age of the sample was 33.1, divided in two subgroups, respectively of 21.3 years (the sons) and 51.5 (the parents). The data were collected using the following likert scale: 0 = FALSE, 1 = More FALSE than TRUE, 2 = More TRUE than FALSE, 3 = TRUE. However, they were initially analyzed in a dichotomous way, as usually done in the MMPI-2 scoring systems. Items scored 1 if TRUE, in MMPI-2, were rescored as follows: 0 if the answer were (0 = FALSE) or (1 = More FALSE than TRUE); 1 if the answer were (2 = More TRUE than FALSE) or (3 = TRUE). Items scored 1 if FALSE, in MMPI-2, were instead rescored in the reversed way. Subsequently, in order to reduce measurement error and increase person reliability (<http://www.winsteps.com/winman/reliability.htm>), we introduced the likert scale originally used in the survey. In the following (see table 2) the labels of the items belonging to DEP and ANX scales (which do not have any item in common) begin with “D” and “A” respectively, while the labels of items belonging to the Mt scale only begin with “M”. We used the Winsteps software (<http://www.winsteps.com/>) to estimates the Rasch models.

Table 1. The sample

Sample composition				AGE: average values			
Sex\Relat	Son	Parent	Total	Sex\Relat	Son	Parent	Total
Female	80	54	134	Female	21.3	50.0	32.9
Male	59	36	95	Male	21.2	53.8	33.6
Total	139	90	229	Total	21.3	51.5	33.1

Table 2. Items considered for the analysis

Description	Mt	DEP	ANX
N. of original items	41 (10 from DEP and 10 from ANX scales)	33	23
Original items Belonging to the scales (*)	M002 D003 D009 M010 A015 M016 M020 M028 A031 D038 M043 D071 M073 M081 D082 D095 M110 D130 M131 A140 M148 M152 D215 M218 A223 M233 M269 A273 A299 M302 M325 D331 A339 M357 A405 A408 D411 M449 M464 A469 M472	D003 D009 D038 D052 D056 D065 D071 D075 D082 D092 D095 D130 D146 D215 D234 D246 D277 D303 D306 D331 D377 D388 D399 D400 D411 D454 D506 D512 D516 D520 D539 D546 D554	A015 A030 A031 A039 A140 A170 A196 A208 A223 A273 A290 A299 A301 A305 A339 A405 A408 A415 A463 A469 A496 A509 A556
N. of items selected from Rasch analysis	31 (8 from DEP scale and 10 from ANX scale)	22	20

Items select ed from Rasch analys is (**)	Likert 0123	D003 M010 A031 D082 D095 M110 M148 D215 A273 M325 A405 A408 M472	D009 D065 D095 D331 D377 D388 D554	A273 A299 A301 A305 A339 A408 A415 A463
	Binary	M002 D009 A015 M016 M028 D038 M073 D130 A140 M152 M218 A223 A299 M302 D331 A339 M464 A469	D038 D052 D056 D071 D130 D146 D215 D234 D277 D303 D411 D454 D520 D539 D546	A015 A030 A031 A039 A140 A170 A196 A405 A469 A496 A509 A556
Items with DIF (SEX)		D009 M016 M028 M152 M325 M464	D009 D038 D052 D146	
Items with DIF (AGE)		M002 M010 A015 D038 M073 A299 M302 M472	D065 D095 D411 D454	A031 A140 A196 A301 A496
(*)		Items eliminated during the analysis Items eliminated because highly correlated with other items		
(**)		For "TRUE" items 0 = FALSE, 1 = More FALSE than TRUE, 2 = More TRUE than FALSE, 3 = TRUE For "FALSE" items 0 = TRUE, 1 = More TRUE than FALSE, 2 = More FALSE than TRUE, 3 = FALSE		

The Mt Scale

Mt Scale: the Rasch Dichotomous Model

The item M357 (*"I am quite often not in on the gossip and talk of the group that I belong to"*), scored 1 if TRUE in the MMPI-2, showed 0 point-measure correlation. After rescoring M357 as 1 if FALSE, and re-running the Winsteps program, we got a point-measure correlation of 0.15 for it: so we decided to keep it in the dataset. We then go on to analyze possible violation of the assumption of unidimensionality. The PCA (Principal Component Analysis) of the standardized residuals showed a 2.6 points of variance explained by the first component: being this greater than 2 (usually accepted as upper limit for a unidimensional construct (<http://www.rasch.org/rmt/rmt191h.htm>), with the sample size and number of items in this case), this means possible violation of the unidimensional hypothesis on which the Rasch model is based. This is confirmed by values lower than 1 for the disattenuated correlations between person measures, calculated using subsets of items representing the opposite poles of the factors. Being however, the first component not very big, we may think that violation of the unidimensional hypothesis may be due to some correlated items and possible misfitting items. The largest residual correlations shows the presence of couples (M073, D411) and (D071, M073) with correlation greater than 0.30: we decide to delete D411 and D071 from the set of items and rerun the program. The variance explained by the first component of PCA decreased to 2.3 but still the (disattenuated) correlations between person measures, based on opposite clusters, show values lower than 0.90. We then looked at the fit indices: item M357 (despite rescoring), M269 and M081 show INFIT and OUTFIT values significantly greater than 1, indicating serious under fitting. We decided to remove them from the analysis and rerun the program. We then looked at the fit indices: 5 items show INFIT and OUTFIT indices significantly greater than 1: M043, M131, M020, M449. We decided to eliminate them from the analysis and rerun the program. The variance explained by the first component of PCA decreased at 2.2 and the (disattenuated) correlations between person measures, based on opposite clusters, show values 1 for almost the components; the correlations between items were all less than 0.30 and fit indices were quite good with the exception of the item M233 that under fit significantly. After deletion of this item, we end up with a set of items and persons satisfying, at least approximately, the model. Point-measure correlations were in the interval (0.23, 0.59); the variance explained by the first component of PCA was 2.07 and the

(disattenuated) correlations between person measures, based on opposite clusters, were all greater than 0.86, but no clear evidence of different dimensions came from the analysis of the content of the items corresponding to opposite loadings. Correlations between item residuals were in the range (-0.23, 0.23); INFIT MNSQ in the range (0.83, 1.22), OUTFIT MNSQ in the range (0.64, 1.80) but all not significantly different from 1 on the basis of the Z-test; some person were significantly underfitting for OUTFIT indices and with negative point-measure correlation. We decided to eventually remove misfitting persons after the analysis of DIF.

Mt Scale: the Analysis of DIF

In this dataset, there are two possible source of DIF: SEX and AGE. Given the DIF analysis results reported in Table W1, we considered for DIF(SEX): D009, M016, M028, M152, M325, M464, and for DIF(AGE): M002, M010, A015, D038, M073, A299, M302, M472. To take explicitly account of DIF we create new items, such as follows: in the case of M016, for example, M016F (for female) and M016M (for male), where M016F contains missing values when the answer corresponds to a male person, vice versa for M016M. The other values are those of the original M016 item. Rasch model, as well known, can handle missing values without problems (Linacre, 2004). For DIF(AGE) we used for example D009S for sons, and D009P for parents. We rerun the program and first we deleted the 13 most underfitting persons. We rerun. Point-measure correlations were in the interval (0.17, 0.62); the variance explained by the first component of PCA was 2.4 and a (disattenuated) correlation between person measures, based on opposite clusters, for the second component were low as 0.76, but no clear evidence of different dimensions come out from the analysis of the content of the items; correlations between items were in the range (-0.29, 0.26); INFIT MNSQ in the range (0.73, 1.16), OUTFIT MNSQ in the range (0.14, 1.42) all non significantly different from 1 of the base of the Z-test; just a few persons were significantly underfitting. The summary statistics and other information regarding this last model are contained in Tab. W2, W3, and W6. Item's reliability was 0.95 and Model RMSE 0.29; person's reliability was 0.85 and their Model RMSE was 0.54. In order to reduce measurement error and to increase reliability, we introduced in the analysis the likert scale originally used to collect the data. This may reduce moreover the gap, that we observe from the item-map (Tab. W6), between the most difficult (the easiest) item and the items which span the interval (-1.5, 1.5) logits around the average.

Table W1. TABLE 30.1 Mt SCALE: DICHOTOMOUS MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 41 ITEM REPORTED: 229 PERSON 31 ITEM 2 CATS																	

- DIF FOR SEX (F = female, M = male)																	

	PERSON	Obs-Exp	DIF	DIF	PERSON	Obs-Exp	DIF	DIF		JOINT	Rasch-Welch	Mantel-Haenszel	Size	Active	ITEM				
	CLASS	Average	MEASURE	S.E.	CLASS	Average	MEASURE	S.E.	CONTRAST Name	S.E.	t	d.f.	Prob.	Chi-squ	Prob.	CUMLOS	Slices	Number	

	1	-.04	3.54	.50	2	.07	1.41	.37	M002	2.13	.62	3.46	221	.0007	5.2274	.0222	1.98	21	1
	1	.03	1.06	.25	2	-.05	2.03	.43	M010	-.98	.49	-1.98	173	.0492	4.0478	.0442	-1.09	21	4
	1	-.06	-.02	.21	2	.09	-.85	.25	A015	.83	.33	2.56	197	.0112	6.1320	.0133	.95	21	5
	1	.07	-1.20	.20	2	-.11	-.19	.26	D038	-1.01	.33	-3.07	195	.0025	8.0834	.0045	-1.04	21	10
	1	.05	.05	.21	2	-.08	1.05	.32	M073	-1.00	.38	-2.62	183	.0096	5.8907	.0152	-1.33	21	13
	1	-.03	1.68	.28	2	.05	.85	.31	A299	.83	.42	2.00	206	.0472	1.6648	.1970	.68	21	29
	1	-.05	.47	.21	2	.08	-.33	.26	M302	.80	.34	2.37	200	.0189	2.7079	.0999	.58	21	30
	1	-.06	1.54	.27	2	.09	.31	.28	M472	1.23	.39	3.19	207	.0017	9.0306	.0027	1.31	21	41

Table W2. TABLE 3.1 Mt SCALE: DICHOTOMOUS MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 45 ITEM REPORTED: 216 PERSON 45 ITEM 2 CATS								
SUMMARY OF 212 MEASURED (NON-EXTREME) PERSON								
	TOTAL			MODEL		INFIT		OUTFIT
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	10.5	30.9	-1.22	.52	.99	.1	.91	.0
P.SD	6.2	.5	1.38	.16	.22	.8	.34	.5
S.SD	6.2	.5	1.38	.16	.22	.8	.34	.5
MAX.	26.0	31.0	1.99	1.21	1.76	2.4	2.23	1.7
MIN.	1.0	26.0	-4.70	.40	.32	-2.3	.04	-1.0
REAL RMSE	.56	TRUE SD	1.26	SEPARATION	2.24	PERSON RELIABILITY	.83	
MODEL RMSE	.54	TRUE SD	1.27	SEPARATION	2.35	PERSON RELIABILITY	.85	
S.E. OF PERSON MEAN = .09								
MINIMUM EXTREME SCORE: 4 PERSON 1.9%								
DELETED: 13 PERSON								
SUMMARY OF 216 MEASURED (EXTREME AND NON-EXTREME) PERSON								
	TOTAL			MODEL		INFIT		OUTFIT
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	10.3	30.9	-1.31	.54				
P.SD	6.3	.5	1.52	.25				
S.SD	6.3	.5	1.52	.25				
MAX.	26.0	31.0	1.99	1.96				
MIN.	.0	26.0	-6.22	.40				
REAL RMSE	.62	TRUE SD	1.39	SEPARATION	2.25	PERSON RELIABILITY	.84	
MODEL RMSE	.60	TRUE SD	1.40	SEPARATION	2.34	PERSON RELIABILITY	.85	
S.E. OF PERSON MEAN = .10								
PERSON RAW SCORE-TO-MEASURE CORRELATION = .95								
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .81 SEM = 2.73								
SUMMARY OF 45 MEASURED (NON-EXTREME) ITEM								
	TOTAL			MODEL		INFIT		OUTFIT
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	49.4	148.2	.00	.26	1.01	.1	.91	-.2
P.SD	40.3	54.8	1.36	.13	.10	.9	.26	.8
S.SD	40.8	55.4	1.38	.14	.10	.9	.26	.8
MAX.	200.0	216.0	4.57	1.01	1.19	1.5	1.42	1.7

```

| MIN.      1.0      84.0      -4.82      .16      .73      -1.7      .14      -1.7 |
|-----|-----|-----|-----|-----|-----|-----|-----|
| REAL RMSE  .30 TRUE SD  1.33 SEPARATION 4.49 ITEM   RELIABILITY .95 |
| MODEL RMSE .29 TRUE SD  1.33 SEPARATION 4.57 ITEM   RELIABILITY .95 |
| S.E. OF ITEM MEAN = .21 |
|-----|-----|-----|-----|-----|-----|-----|-----|

ITEM RAW SCORE-TO-MEASURE CORRELATION = -.83

Global Statistics:
Active PERSON: 216
Active ITEM: 45
Active datapoints: 6669 = 68.6% of Active+Missing datapoints
Missing datapoints: 3051 = 31.4% of Active+Missing datapoints
Standardized residuals N(0,1) mean: -.01 P.SD: .94
Log-likelihood chi-squared: 5768.0403 with approximately 5771 d.f., probability =
.5085
Global Root-Mean-Square Residual: .3769 with expected value: .3758
Capped Binomial Deviance: .1879 for 6669.0 dichotomies with expected value: .1881

```

Table W3. TABLE 13.1 Mt SCALE: DICHOTOMOUS MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 45 ITEM REPORTED: 216 PERSON 45 ITEM 2 CATS												

--												
PERSON: REAL SEP.: 2.25 REL.: .84 ... ITEM: REAL SEP.: 4.49 REL.: .95												
ITEM STATISTICS: MEASURE ORDER												

ENTRY	TOTAL	TOTAL		MODEL	INFIT		OUTFIT	PTMEASUR-AL	EXACT	MATCH		
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%
ITEM												
+-----+-----+-----+-----+-----+												
	1	1	131	4.57	1.01	.94	.3	.14	-1.1	.17	.09	99.2
	7	6	84	1.75	.45	.96	.0	.47	-.6	.31	.26	92.6
	13	6	92	1.72	.44	1.01	.2	.64	-.3	.25	.23	93.5
	31	14	130	1.64	.30	1.06	.3	.77	-.3	.30	.31	88.4
	44	17	130	1.39	.28	1.05	.3	.78	-.4	.33	.34	87.6
	10	19	124	1.25	.27	.82	-1.2	.57	-1.0	.45	.35	88.3
	2	10	85	1.11	.37	.93	-.2	.62	-.5	.36	.32	90.2
	14	31	216	1.11	.21	.94	-.4	.66	-1.0	.40	.35	86.8
	35	23	124	.97	.26	1.01	.1	1.18	.6	.36	.38	84.2
	6	22	124	.94	.26	1.19	1.3	1.29	.9	.26	.39	80.5
	32	12	85	.87	.34	1.10	.5	.95	.1	.30	.34	85.4
	5	13	92	.77	.32	1.10	.6	1.36	.8	.23	.32	85.9
	18	13	85	.75	.33	.73	-1.5	.42	-1.2	.51	.35	86.6
	30	44	216	.59	.19	.98	-.2	.72	-1.1	.43	.40	78.8
A273												

	28	45	216	.55	.19	.99	.0	.87	-.4	.41	.40	81.1	80.6
						M218							
	12	31	124	.48	.24	1.16	1.4	1.42	1.4	.32	.43	75.8	77.7
						M028F							
	29	52	216	.32	.18	1.10	1.2	1.07	.4	.37	.42	73.6	78.4
						A223							
	33	35	131	.30	.22	1.02	.3	1.17	.7	.41	.44	81.5	76.9
						M302S							
	19	54	215	.24	.18	.99	-.1	.87	-.6	.44	.43	77.7	77.9
						D082							
	36	19	92	.23	.28	1.02	.2	.77	-.6	.40	.38	80.4	80.4
						M325M							
	11	21	92	.08	.27	1.03	.3	1.03	.2	.37	.39	80.4	78.9
						M016M							
	45	20	84	.08	.29	1.13	.9	1.16	.5	.35	.42	76.5	78.5
						M472P							
	43	60	214	.06	.17	.89	-1.4	.69	-1.7	.52	.45	78.1	76.4
						A469							
	22	65	216	-.08	.17	.89	-1.5	.75	-1.5	.53	.46	81.6	75.6
						D130							
	8	43	126	-.14	.22	.96	-.4	.90	-.5	.50	.47	76.8	74.0
						A015S							
	20	67	216	-.14	.17	.88	-1.7	.75	-1.5	.54	.46	79.2	75.3
						D095							
	17	44	130	-.15	.21	.89	-1.2	.75	-1.4	.56	.48	79.8	74.4
						M073S							
	39	69	215	-.20	.17	1.01	.1	.94	-.3	.47	.47	71.6	74.9
						A405							
	26	26	92	-.27	.26	1.17	1.4	1.11	.5	.32	.43	70.7	75.8
						M152M							
	42	27	90	-.37	.26	.94	-.5	.77	-.8	.49	.44	74.4	75.0
						M464M							
	4	49	124	-.43	.22	1.13	1.3	1.10	.6	.44	.50	67.5	73.5
						D009F							
	16	28	85	-.53	.27	1.07	.6	1.02	.2	.44	.47	76.8	74.2
						D038P							
	34	29	85	-.60	.27	1.13	1.1	1.15	.6	.42	.48	65.9	73.9
						M302P							
	27	90	216	-.77	.16	.99	-.1	.99	.0	.51	.50	73.1	73.2
						D215							
	23	95	215	-.92	.16	1.10	1.4	1.06	.5	.46	.51	68.7	73.1
						A140							
	40	97	215	-.96	.16	1.04	.6	1.13	1.1	.48	.51	74.4	73.0
						A408							
	25	65	124	-1.18	.22	1.05	.6	1.10	.7	.51	.54	71.7	74.2
						M152F							
	9	38	85	-1.22	.26	1.17	1.5	1.36	1.7	.43	.52	70.7	73.1
						A015P							
	41	68	124	-1.32	.22	.85	-1.5	.82	-1.1	.62	.55	79.2	74.4
						M464F							
	15	73	130	-1.39	.21	.99	-.1	1.01	.1	.52	.52	73.6	73.5
						D038S							
	38	115	215	-1.43	.16	.93	-1.0	.86	-1.2	.57	.53	76.3	73.5
						A339							
	3	119	216	-1.53	.16	1.03	.4	1.00	.0	.52	.53	74.1	73.7
						D003							
	37	120	216	-1.56	.16	.99	.0	.95	-.4	.54	.53	72.6	73.8
						D331							
	24	127	216	-1.75	.17	1.08	1.1	1.10	.8	.49	.53	71.7	74.6
						M148							
	21	200	216	-4.82	.32	1.13	.6	.86	-.1	.40	.44	94.3	94.3
						M110							
	-----+-----												

MEAN	49.4	148.2	.00	.26	1.01	.1	.91	-.2		79.5	79.6
P.SD	40.3	54.8	1.36	.13	.10	.9	.26	.8		7.6	6.8

Table W4. TABLE 3.1 Mt SCALE: GROUPED RATING SCALE MODEL - WINSTEPS 3.90.0

```

INPUT: 229 PERSON 45 ITEM REPORTED: 216 PERSON 45 ITEM 54 CATS
-----
SUMMARY OF 216 MEASURED PERSON
-----
| TOTAL          MODEL          INFIT          OUTFIT          |
| SCORE          COUNT          MEASURE          S.E.          MNSQ          ZSTD          MNSQ          ZSTD          |
|-----|-----|-----|-----|-----|-----|-----|-----|
| MEAN          20.6          30.9          -1.04          .33          1.05          .1          .93          .0          |
| P.SD          9.5          .5          .97          .06          .41          1.2          .31          .8          |
| S.SD          9.5          .5          .97          .06          .41          1.3          .31          .8          |
| MAX.          42.0          31.0          .94          .65          3.33          4.0          2.10          2.9          |
| MIN.          3.0          26.0          -3.75          .29          .44          -2.5          .32          -1.8          |
|-----|-----|-----|-----|-----|-----|-----|-----|
| REAL RMSE          .36 TRUE SD          .90 SEPARATION          2.46 PERSON RELIABILITY          .86          |
| MODEL RMSE          .34 TRUE SD          .91 SEPARATION          2.69 PERSON RELIABILITY          .88          |
| S.E. OF PERSON MEAN = .07          |
|-----|-----|-----|-----|-----|-----|-----|-----|

DELETED: 13 PERSON
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .82 SEM = 4.02

SUMMARY OF 45 MEASURED ITEM
-----
| TOTAL          MODEL          INFIT          OUTFIT          |
| SCORE          COUNT          MEASURE          S.E.          MNSQ          ZSTD          MNSQ          ZSTD          |
|-----|-----|-----|-----|-----|-----|-----|-----|
| MEAN          98.9          148.2          .00          .21          1.00          -.1          .92          -.3          |
| P.SD          111.3          54.8          1.09          .14          .13          1.3          .22          1.4          |
| S.SD          112.5          55.4          1.10          .15          .13          1.4          .22          1.4          |
| MAX.          521.0          216.0          4.35          1.01          1.28          2.8          1.33          2.9          |
| MIN.          1.0          84.0          -3.02          .08          .78          -3.0          .24          -2.9          |
|-----|-----|-----|-----|-----|-----|-----|-----|
| REAL RMSE          .26 TRUE SD          1.06 SEPARATION          4.15 ITEM RELIABILITY          .95          |
| MODEL RMSE          .25 TRUE SD          1.06 SEPARATION          4.19 ITEM RELIABILITY          .95          |
| S.E. OF ITEM MEAN = .16          |
|-----|-----|-----|-----|-----|-----|-----|-----|

ITEM RAW SCORE-TO-MEASURE CORRELATION = -.60

Global Statistics:
Active PERSON: 216
Active ITEM: 45
Active datapoints: 6669 = 68.6% of Active+Missing datapoints
Missing datapoints: 3051 = 31.4% of Active+Missing datapoints
Standardized residuals N(0,1) mean: -.01 P.SD: .97
Log-likelihood chi-squared: 9370.7824 with approximately 9379 d.f., probability =
.5220
Global Root-Mean-Square Residual: .5764 with expected value: .5606
Capped Binomial Deviance: .1978 for 3874.0 dichotomies with expected value: .2142

```

Table W5. TABLE 13.1 Mt SCALE: GROUPED RATING SCALE MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 45 ITEM REPORTED: 216 PERSON 45 ITEM 54 CATS														

PERSON: REAL SEP.: 2.46 REL.: .86 ... ITEM: REAL SEP.: 4.15 REL.: .95														
ITEM STATISTICS: MEASURE ORDER														
ENTRY	TOTAL	TOTAL		MODEL	INFIT	OUTFIT	PTMEASUR-AL	EXACT	MATCH					
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	ITEM	G
1	1	131	4.35	1.01	.96	.3	.24	-.7	.17	.07	99.2	99.2	M002S	1
13	6	92	1.75	.43	.99	.1	.69	-.4	.23	.19	93.5	93.5	M028M	1
31	14	130	1.51	.29	.96	-.1	.67	-.9	.32	.24	89.2	89.2	A299S	1
10	19	124	1.12	.26	.84	-1.0	.59	-1.5	.46	.29	84.7	84.7	M016F	1
2	10	85	1.12	.35	.92	-.3	.63	-.9	.37	.26	88.2	88.2	M002P	1
7	39	84	1.09	.20	1.22	1.3	1.15	.8	.43	.44	60.7	65.5	M010P	3
32	12	85	.89	.33	1.00	.1	.89	-.2	.29	.28	87.1	85.9	A299P	1
5	13	92	.84	.31	1.03	.2	1.15	.5	.22	.26	85.9	85.9	D009M	1
18	13	85	.78	.32	1.78	-1.2	.50	-1.7	.53	.29	85.9	84.8	M073P	1
44	71	130	.68	.13	1.28	1.9	1.11	.6	.41	.49	59.2	60.0	M472S	H
28	45	216	.55	.18	.92	-.9	.75	-1.6	.43	.33	80.1	79.6	M218	1
12	31	124	.43	.22	1.02	.3	1.07	.4	.32	.35	76.6	76.0	M028F	1
35	87	124	.37	.12	1.00	.1	1.21	1.1	.52	.54	56.5	55.7	M325F	E
14	152	216	.34	.10	.99	-.1	.93	-.6	.52	.51	53.7	56.4	A031	5
29	52	216	.33	.17	.99	-.1	.94	-.4	.36	.34	77.8	76.7	A223	1
33	35	131	.29	.21	.93	-.7	.90	-.5	.42	.35	78.6	74.6	M302S	1
11	21	92	.19	.26	.97	-.2	.84	-.6	.37	.32	78.3	77.5	M016M	1
6	114	124	.12	.14	1.21	1.6	1.25	1.9	.32	.52	53.2	55.1	M010S	3
30	175	216	.11	.09	1.01	.2	.92	-.7	.56	.55	52.3	52.3	A273	C
43	60	214	.10	.16	.84	-2.3	.68	-2.7	.54	.36	76.6	73.8	A469	1
22	65	216	-.02	.16	.83	-2.6	.73	-2.5	.54	.37	80.1	72.6	D130	1
19	200	215	-.05	.09	1.09	1.0	1.04	.4	.52	.56	49.3	50.2	D082	6
36	68	92	-.09	.14	1.07	.5	.96	-.1	.52	.53	57.6	52.4	M325M	E
8	43	126	-.10	.20	.88	-1.4	.81	-1.4	.50	.38	76.2	70.3	A015S	1
17	44	130	-.11	.20	.84	-2.1	.75	-2.0	.55	.38	73.8	70.6	M073S	1
45	68	84	-.13	.14	1.27	1.7	1.33	1.6	.42	.57	48.8	52.0	M472P	H
26	26	92	-.13	.25	1.05	.5	.95	-.2	.32	.34	69.6	73.4	M152M	1
42	27	90	-.23	.25	.91	-.9	.80	-1.2	.46	.35	75.6	72.3	M464M	1
39	233	215	-.32	.09	1.14	1.5	1.09	1.0	.52	.58	44.2	47.6	A405	F
20	239	216	-.32	.09	.97	-.3	.97	-.3	.58	.57	53.2	49.8	D095	7
16	28	85	-.36	.25	.94	-.6	.87	-.7	.45	.38	74.1	71.1	D038P	1
4	49	124	-.36	.20	.99	-.1	.96	-.3	.42	.41	68.5	69.0	D009F	1
34	29	85	-.42	.25	.98	-.2	1.14	.9	.39	.38	70.6	70.6	M302P	1
27	272	216	-.63	.08	1.10	1.1	1.08	.8	.58	.62	42.6	44.6	D215	B
23	95	215	-.75	.15	.96	-.6	.93	-.9	.45	.41	68.8	68.3	A140	1
40	292	215	-.80	.08	1.12	1.4	1.16	1.6	.58	.63	42.3	42.8	A408	G
9	38	85	-.95	.24	1.04	.4	1.05	.4	.37	.41	71.8	68.1	A015P	1
25	65	124	-.99	.20	.89	-1.5	.85	-1.5	.53	.43	71.0	68.7	M152F	1
3	330	216	-1.00	.09	1.17	1.9	1.18	1.9	.48	.58	51.9	49.4	D003	2
41	68	124	-1.11	.20	.78	-3.0	.74	-2.8	.63	.43	79.0	69.0	M464F	1
38	115	215	-1.19	.15	.85	-2.8	.80	-2.9	.56	.42	74.9	68.1	A339	1
15	73	130	-1.19	.19	.91	-1.2	.95	-.5	.49	.41	73.1	68.2	D038S	1
37	120	216	-1.29	.15	.89	-2.0	.85	-2.1	.53	.41	73.6	68.4	D331	1
24	372	216	-1.38	.09	1.26	2.8	1.28	2.9	.51	.63	40.7	43.2	M148	L
21	521	216	-3.02	.12	1.21	2.0	1.22	2.1	.34	.50	56.0	62.2	M110	8
MEAN	98.9	148.2	.00	.21	1.00	-.1	.92	-.3			69.0	67.9		
P.SD	111.3	54.8	1.09	.14	.13	1.3	.22	1.4			14.9	13.9		

Table W6. TABLE 12.2 Mt SCALE - WINSTEPS 3.90.0

DICHOTOMOUS MODEL	GROUPED RATING SCALE MODEL
INPUT: 229 PERSON 45 ITEM REPORTED: 216 PERSON	INPUT: 229 PERSON 45 ITEM REPORTED: 216 PERSON
45 ITEM 2 CATS	45 ITEM 54 CATS
MEASURE PERSON - MAP - ITEM	MEASURE PERSON - MAP - ITEM - 50% Cumulative probabilities (Rasch-Thurstone thresholds)
5 <more> <rare> + M002S + + + T + . # A299S M010P M028M T . S M472S # A031 M002P M016F . + M010S M325F # A299P D009M M073P .## A273 .### M028F M218 ##### A223 M302S #### S D082 M325M #### +M A469 D130 M016M M472P .##### A015S A405 D095 M073S ### M152M M464M ##### D009F D038P #### M302P .### D215 -1 ##### + A140 A408 .## M A015P M152F ##### S D038S M464F .##### A339 D003 D331 .### M148 # -2 ##### + .### .## ### .#### S T -3 .##### + . # #### -4 T+ ## M110 -5 ## + <less> <freq> EACH "#" IS 2: EACH "." IS 1	5 <more> <rare> + M002S .1 + + + + T + M472S .3 M010S .3 M028M .1 A299S .1 M325F .3 M010P .2 S M016F .1 # + # T M073P .1 # ##### M472S .2 M010S .2 M218 .1 M302S .1 M325F .2 A273 .2 M016M .1 S+M D082 .2 D130 .1 A015S .1 D095 .2 M148 .3 M152M .1 M472P .2 M464M .1 A405 .2 M472S .1 M302P .1 A140 .1 D215 .2 M325F .1 A408 .2 M010P .1 M325M .1 M110 .3 A339 .1 D003 .2 M472P .1 D331 .1 M148 .2 D082 .1 M010S .1 S+ D095 .1 T ### .## .## M148 .1 T+ # # + M110 .1 + -5 <less> <freq> EACH "#" IS 2: EACH "." IS 1

Table W7. TABLE 3.14 Mt SCALE: GROUPED RATING SCALE MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 45 ITEM REPORTED: 216 PERSON 45 ITEM 54 CATS CATEGORY PROBABILITIES: MODES - Andrich thresholds at intersections for items with likert scale		
D003	M010	A031
D082	D095	M110
M148	D215	A273
M325	A405	A408
M472		

Mt Scale: the Rasch Grouped Rating Scale Model

Therefore, we started with a Partial Credit Model, (i.e. a Rating Scale model different for each item) with 4 categories. We run the program and we analyzed the order of the thresholds (<http://www.rasch.org/rmt/rmt174a.htm>). We decided to estimate a Grouped Rating Scale model where groups were defined by the items: for example, D009F and D009M share a common Rating Scale model. Some items seem to support the likert scale (well ordered thresholds: Tab. W7) but for others it seems much better to maintain the binary coding (see Table 2): we kept recoding as binary the items most underfitting until all the INFIT and OUTFIT MNSQ were less than 1.4, a limit usually accepted as good in Rating Scale models (Bond & Fox, 2007, p. 243). The summary statistics and other information regarding this last model are contained in Tab. W4, W5, and W6. The main results of the introduction of the likert scale for some of the items, were that person Reliability growth to 0.88 (0.85 for the dichotomous model), and the Model RMSE decreased to 0.34 (from 0.54 for the dichotomous model): the gain in this indicator was quite sensible (-37%) and a greater precision in person measure estimates was so achieved. For the items the Model RMSE decreased to 0.25 from 0.29 (dichotomous model). The global fit of this model was Log-likelihood chi-squared: 9370.7824 with approximately 9379 d.f. (Tab. W4); the global fit for the dichotomous model (Tab. W2) was Log-likelihood chi-squared: 5768.0403 with approximately 5771 d.f. (Tab. W2). It can be easily verified that the generalized likelihood test based on the -2 Log-likelihood ratio statistics has a p-value of zero: this means that the use of the likert scale, at least for some items, would improve

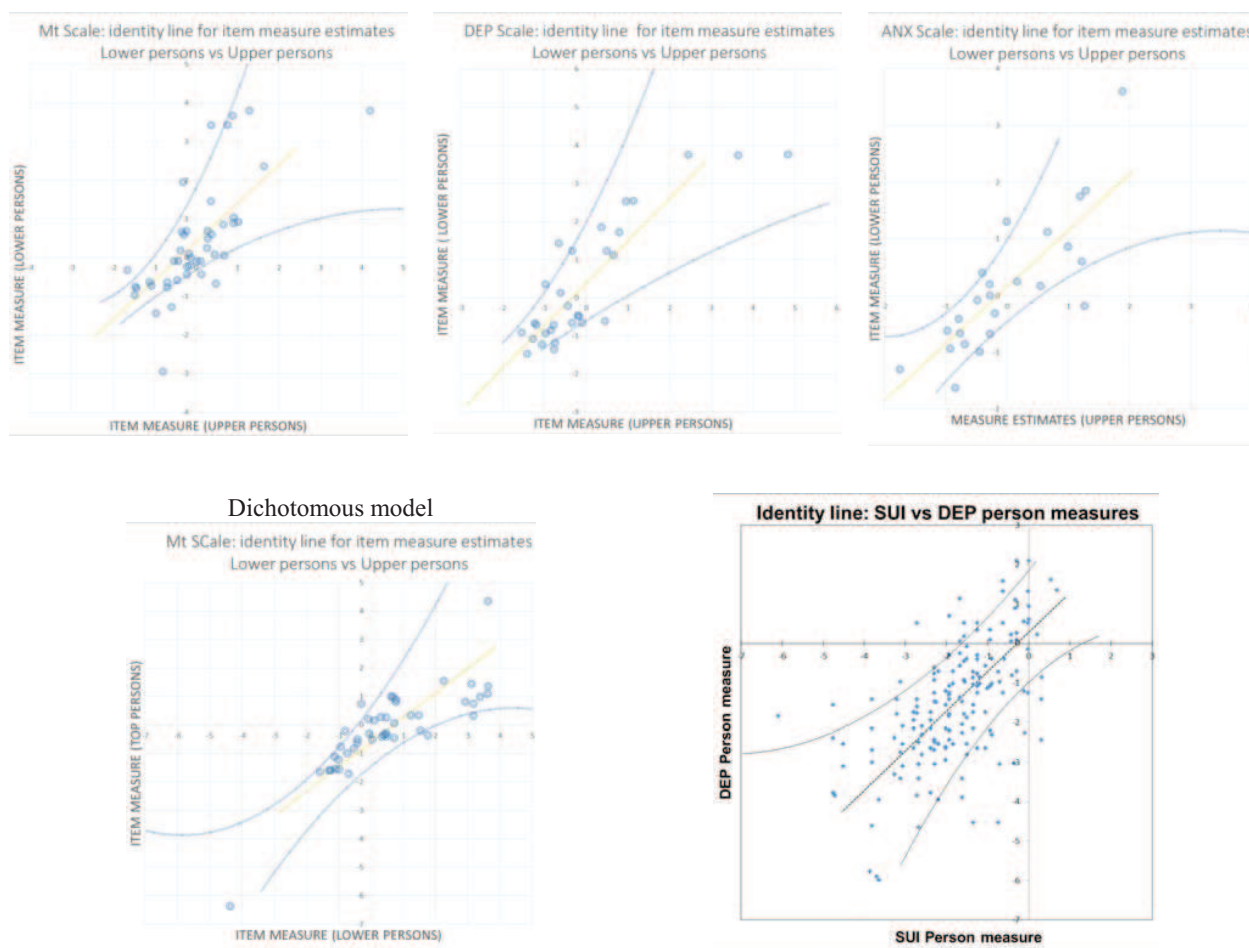


Figure 1. Grouped Rating Scale model Identity lines for item measures estimated using the 50% Top Persons sample vs the 50% Bottom Persons sample

significantly the measurement process. Confronting the item-map (Tab. W6) for the Grouped Rating Scale and the dichotomous models we observed moreover that the introduction of the likert scale for the easiest items M110 and M148 reduced the lack of items in the lower tail of the distribution: the maximum difference between items-categories in the likert scale is 1.5 logits, while in the dichotomous case is 3.5 logits. This is one of the reason for the reduction of the Model RMSE. Finally, regarding the invariance property of the Rasch model, we may take a look at the Figure 1 where are reported the identity lines of the item measures estimated using the Grouped Rating Scale model respectively on the 50% top and on the 50% bottom of the sample: we may observe that most of the items lie inside the identity line bands, but in particular the easiest item falls outside such bands. The same graph for dichotomous model shows that such a problem is avoided and therefore it may be due to the likert scale used. Therefore, we may say that the invariance property is substantially satisfied.

The DEP Scale

DEP Scale: the Rasch Dichotomous Model

The item D092 (*"I don't seem to care what happens to me"*), scored 1 if TRUE in the MMPI-2, showed a negative point-measure correlation. After rescoring D092 as 1 if FALSE, and re-running the Winsteps program, we got a point-measure correlation of 0.20 for it: so we decided to keep it in the dataset. We then go on to analyze possible violation of the assumption of unidimensionality. The PCA of the standardized residuals showed 2.0 points of variance explained by first component, and the lowest disattenuated correlation between person measures, calculated using subsets of items representing the opposite poles of the factors was 0.90. The largest residual correlations showed the presence of couples of items (D506, D516) and (D512, D516) with correlation greater than 0.30: we decided to delete D516 and rerun the program. The first component of the PCA of the standardized residual explained 1.9 points of residual variance. The disattenuated correlations between person measures, calculated using subsets of items representing the opposite poles of the factors, show a couple of item clusters with a level of 0.76, the other being 1 or near 1. The largest residual correlations do not show the presence of dependent items. Looking at the fit indices, we observed that several items have INFIT and OUTFIT indices significantly greater than 1: we kept eliminating in sequence the most underfitting and rerunning the program. We end up with the following list of eliminated items due to underfitting: D003, D082, D092 (despite rescoring), D246, D306, D399, D400, D506, D512. Between them D003 and D082 were valid items of the Mt scale. After this deletion, all point-measure correlations were in the range (0.17, 0.65). The PCA of the standardized residuals showed 1.66 points of variance explained by first component; all disattenuated correlations between person measures, calculated using subsets of items representing the opposite poles of the factors, were at level 1. The largest standardized residual correlations do not show any dependence between items. INFIT MNSQ were in the range (0.78, 1.17), OUTFIT MNSQ (0.51, 1.32): no one significantly underfitting, some significantly overfitting. Some person were misfitting but before to delete them we went trough the analysis of DIF.

DEP Scale: the Analysis of DIF

Given the DIF analysis results reported in Table W8, we considered for DIF(SEX): D009, D038, D052, D146, and for DIF(AGE): D065, D095, D411, D454. To take explicitly account of DIF we create new items with appropriate missing values. We deleted the 21 most underfitting persons and rerun the program. All point-measure correlations were positive in the range (0.13, 0.67). The first component of the PCA of the standardized residual explained 1.8 points of residual variance. The disattenuated correlations between person measures, calculated using subsets of items representing the opposite poles of the factors, were all 1. The largest residual correlations do not show the presence of dependent items. INFIT MNSQ were in the range (0.73, 1.26), OUTFINT MNSQ (0.16, 1.33): no one significantly

underfitting, only one significantly overfitting. The summary statistics and other information regarding this last model are reported in Tab. W9, W10, and W13. Item reliability was 0.96 and Model RMSE 0.33; person reliability was 0.82 and their Model RMSE was 0.67. In order to reduce measurement error and to increase reliability, we introduced in the analysis the likert scale originally used to collect the data. This may help moreover to reduce the lack of items observed in the item-map (Tab. W13), in the lowest tail of the distribution of person measures.

Table W8. TABLE 30.1 DEP SCALE: DICHOTOMOUS MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 33 ITEM REPORTED: 229 PERSON 22 ITEM 2 CATS																	

--																	
DIF FOR SEX (F = female, M = male)																	

Table W9. TABLE 3.1 DEP SCALE: DICHOTOMOUS MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 30 ITEM REPORTED: 208 PERSON 30 ITEM 2 CATS ----- -- SUMMARY OF 186 MEASURED (NON-EXTREME) PERSON -----									
	TOTAL			MODEL	INFIT		OUTFIT		
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	
MEAN	6.8	21.9	-1.37	.65	1.01	.1	.82	.1	
P.SD	4.8	.7	1.56	.17	.20	.7	.35	.4	
S.SD	4.8	.7	1.56	.17	.20	.7	.35	.4	
MAX.	19.0	22.0	2.87	1.07	1.53	2.3	1.86	1.2	
MIN.	1.0	15.0	-3.81	.50	.35	-2.0	.12	-1.1	

REAL RMSE	.69	TRUE SD	1.40	SEPARATION	2.02	PERSON RELIABILITY	.80		
MODEL RMSE	.67	TRUE SD	1.41	SEPARATION	2.10	PERSON RELIABILITY	.82		
S.E. OF PERSON MEAN = .11									

MINIMUM EXTREME SCORE: 22 PERSON 10.6%									
DELETED: 21 PERSON									
SUMMARY OF 208 MEASURED (EXTREME AND NON-EXTREME) PERSON									
	TOTAL			MODEL	INFIT		OUTFIT		
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	
MEAN	6.0	21.9	-1.76	.78					
P.SD	5.0	.6	1.85	.41					
S.SD	5.0	.6	1.86	.41					
MAX.	19.0	22.0	2.87	1.86					
MIN.	.0	15.0	-5.08	.50					
REAL RMSE	.89	TRUE SD	1.62	SEPARATION	1.82	PERSON	RELIABILITY	.77	
MODEL RMSE	.88	TRUE SD	1.63	SEPARATION	1.87	PERSON	RELIABILITY	.78	
S.E. OF PERSON MEAN = .13									
PERSON RAW SCORE-TO-MEASURE CORRELATION = .97									
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .81 SEM = 2.17									
SUMMARY OF 30 MEASURED (NON-EXTREME) ITEM									
	TOTAL			MODEL	INFIT		OUTFIT		
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	
MEAN	41.9	151.8	.00	.29	1.01	.2	.85	-.3	
P.SD	29.5	53.5	1.61	.16	.14	1.1	.30	1.0	
S.SD	30.0	54.4	1.64	.17	.14	1.1	.30	1.0	
MAX.	121.0	208.0	5.12	1.02	1.26	1.7	1.33	1.4	
MIN.	1.0	80.0	-2.33	.18	.73	-2.1	.10	-2.1	
REAL RMSE	.34	TRUE SD	1.57	SEPARATION	4.60	ITEM	RELIABILITY	.95	
MODEL RMSE	.33	TRUE SD	1.58	SEPARATION	4.71	ITEM	RELIABILITY	.96	
S.E. OF ITEM MEAN = .30									
ITEM RAW SCORE-TO-MEASURE CORRELATION = -.71									
Global Statistics:									
Active PERSON: 208									
Active ITEM: 30									
Active datapoints: 4553 = 73.0% of Active+Missing datapoints									
Missing datapoints: 1687 = 27.0% of Active+Missing datapoints									
Standardized residuals N(0,1) mean: -.02 P.SD: .86									
Log-likelihood chi-squared: 3223.5770 with approximately 3273 d.f., probability = .7276									
Global Root-Mean-Square Residual: .3399 with expected value: .3390									
Capped Binomial Deviance: .1542 for 4553.0 dichotomies with expected value: .1562									

Table W10. TABLE 13.1 DEP SCALE: DICHOTOMOUS MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 30 ITEM REPORTED: 208 PERSON 30 ITEM 2 CATS									

PERSON: REAL SEP.: 1.82 REL.: .77 ... ITEM: REAL SEP.: 4.60 REL.: .95									
ITEM STATISTICS: MEASURE ORDER									

ENTRY	TOTAL	TOTAL	MODEL	INFIT		OUTFIT	PTMEASUR-AL	EXACT MATCH	

NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ ITEM	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%
-----+-----+-----+-----+-----+-----+-----												
19	1	208	5.12	1.02	1.00	.3	.16	-1.6	.13	.11	99.5	99.5
					D303							
27	3	203	3.92	.61	.81	-.2	.10	-1.7	.27	.18	98.3	98.4
					D520							
29	9	206	2.67	.37	1.03	.2	.33	-1.1	.33	.29	94.0	95.2
					D546							
28	25	206	1.25	.25	1.03	.3	.68	-.7	.44	.44	87.0	88.3
					D539							
17	28	207	1.07	.24	1.17	1.2	.89	-.2	.41	.45	84.9	87.3
					D234							
24	9	81	.86	.43	.84	-.5	.58	-.6	.54	.47	90.0	89.5
					D411P							
2	11	89	.74	.38	1.17	.8	1.08	.3	.38	.45	87.2	87.9
					D009M							
15	12	89	.60	.37	1.22	1.0	1.03	.2	.38	.46	85.9	87.0
					D146M							
25	28	126	.51	.26	1.04	.4	.85	-.4	.49	.50	80.0	81.3
					D454S							
30	40	206	.46	.21	1.06	.5	.96	.0	.49	.51	84.2	83.3
					D554							
5	26	116	.45	.27	.94	-.4	.75	-.7	.55	.52	82.9	81.9
					D052F							
21	50	208	.04	.20	.93	-.7	.71	-1.4	.58	.54	81.7	80.8
					D377							
11	37	127	-.04	.24	.88	-1.0	.72	-1.2	.60	.54	81.9	78.1
					D095S							
8	38	127	-.10	.24	1.07	.6	.95	-.1	.53	.54	75.0	77.8
					D065S							
23	41	126	-.27	.23	.82	-1.6	.66	-1.7	.63	.55	82.6	76.9
					D411S							
13	64	208	-.48	.19	.81	-2.1	.67	-2.1	.65	.57	81.7	77.9
					D130							
1	43	119	-.54	.24	1.14	1.2	1.23	1.1	.52	.58	73.1	76.2
					D009F							
6	24	88	-.66	.29	1.26	1.6	1.33	1.2	.44	.55	74.0	78.6
					D052M							
18	70	208	-.69	.18	.93	-.7	.86	-.8	.61	.58	77.4	76.9
					D277							
26	21	80	-.70	.32	.98	-.1	.89	-.3	.58	.56	79.7	79.2
					D454P							
10	76	207	-.89	.18	.85	-1.8	.79	-1.5	.65	.59	81.1	75.9
					D071							
9	24	81	-.96	.30	.73	-1.9	.68	-1.2	.67	.57	84.3	77.5
					D065P							
12	24	81	-.96	.30	.85	-1.0	.71	-1.1	.64	.57	81.4	77.5
					D095P							
16	87	208	-1.24	.18	1.14	1.7	1.19	1.3	.54	.60	73.1	75.0
					D215							
7	91	207	-1.38	.18	1.14	1.7	1.21	1.4	.55	.61	71.9	74.6
					D056							
22	92	207	-1.41	.18	1.11	1.3	1.12	.8	.57	.61	70.8	74.6
					D388							
3	59	119	-1.42	.23	1.12	1.1	1.09	.5	.57	.61	69.4	74.6
					D038F							
14	63	119	-1.64	.24	1.07	.6	1.15	.7	.59	.62	71.3	75.4
					D146F							
4	41	88	-1.98	.27	1.19	1.5	1.12	.6	.53	.60	68.8	72.9
					D038M							

	20	121	208	-2.33	.18 1.06	.8 .97	-.1 .60	.62 72.6	76.1
					D331				

	MEAN	41.9	151.8	.00	.29 1.01	.2 .85	-.3	80.9	81.2
	P.SD	29.5	53.5	1.61	.16 .14	1.1 .30	1.0	7.9	7.1

Table W11. TABLE 3.1 DEP SCALE: GROUPED RATING SCALE MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 30 ITEM REPORTED: 208 PERSON 30 ITEM 32 CATS									

SUMMARY OF 204 MEASURED (NON-EXTREME) PERSON									

	TOTAL			MODEL	INFIT		OUTFIT		
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	

	MEAN	11.6	21.9	-1.48	.48	1.05	.1	.85	.1
	P.SD	7.0	.6	1.38	.12	.47	1.2	.47	.5
	S.SD	7.0	.6	1.38	.12	.47	1.2	.47	.5
	MAX.	30.0	22.0	2.10	1.04	2.66	3.5	3.36	1.7
	MIN.	1.0	15.0	-4.65	.39	.27	-2.5	.17	-1.3

	REAL RMSE	.54	TRUE SD	1.27	SEPARATION	2.36	PERSON RELIABILITY	.85	
	MODEL RMSE	.49	TRUE SD	1.29	SEPARATION	2.61	PERSON RELIABILITY	.87	
	S.E. OF PERSON MEAN = .10								

MINIMUM EXTREME SCORE: 4 PERSON 1.9%									
DELETED: 21 PERSON									
SUMMARY OF 208 MEASURED (EXTREME AND NON-EXTREME) PERSON									

	TOTAL			MODEL	INFIT		OUTFIT		
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	

	MEAN	11.4	21.9	-1.57	.51				
	P.SD	7.1	.6	1.50	.22				
	S.SD	7.2	.6	1.50	.22				
	MAX.	30.0	22.0	2.10	1.85				
	MIN.	.0	15.0	-5.99	.39				

	REAL RMSE	.59	TRUE SD	1.37	SEPARATION	2.32	PERSON RELIABILITY	.84	
	MODEL RMSE	.55	TRUE SD	1.39	SEPARATION	2.52	PERSON RELIABILITY	.86	
	S.E. OF PERSON MEAN = .10								

PERSON RAW SCORE-TO-MEASURE CORRELATION = .97									
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .80 SEM = 3.18									
SUMMARY OF 30 MEASURED (NON-EXTREME) ITEM									

	TOTAL			MODEL	INFIT		OUTFIT		
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	

	MEAN	79.1	151.8	.00	.25	1.00	.0	.89	-.1
	P.SD	76.9	53.5	1.45	.18	.14	1.3	.30	1.3
	S.SD	78.2	54.4	1.48	.18	.14	1.3	.31	1.3
	MAX.	328.0	208.0	4.78	1.01	1.33	3.2	1.30	2.8
	MIN.	1.0	80.0	-1.74	.10	.78	-2.8	.12	-2.5

	REAL RMSE	.31	TRUE SD	1.42	SEPARATION	4.66	ITEM RELIABILITY	.96	
	MODEL RMSE	.30	TRUE SD	1.42	SEPARATION	4.70	ITEM RELIABILITY	.96	
	S.E. OF ITEM MEAN = .27								

```

ITEM RAW SCORE-TO-MEASURE CORRELATION = -.48

Global Statistics:
Active PERSON: 208
Active ITEM: 30
Active datapoints: 4553 = 73.0% of Active+Missing datapoints
Missing datapoints: 1687 = 27.0% of Active+Missing datapoints
Standardized residuals N(0,1) mean: -.01 P.SD: .91
Log-likelihood chi-squared: 4945.9214 with approximately 4949 d.f., probability =
.5097
Global Root-Mean-Square Residual: .4780 with expected value: .4663
Capped Binomial Deviance: .1522 for 3100.0 dichotomies with expected value: .1622

```

Active PERSON: 208

Active ITEM: 30

Active datapoints: 4553 = 73.0% of Active+Missing datapoints

Missing datapoints: 1687 = 27.0% of Active+Missing datapoints

Standardized residuals N(0,1) mean: -.01 P.S.D: .91

Log-likelihood chi-squared: 4945.9214 with approximately 4949 d.f., probability = .5097

Global Root-Mean-Square Residual: .4780 with expected value: .4663

Capped Binomial Deviance: .1522 for 3100.0 dichotomies with expected value: .1622

Table W12. TABLE 13.1 DEP SCALE: GROUPED RATING SCALE MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 30 ITEM REPORTED: 208 PERSON 30 ITEM 32 CATS														
--														
PERSON: REAL SEP.: 2.32 REL.: .84 ... ITEM: REAL SEP.: 4.66 REL.: .96														
ITEM STATISTICS: MEASURE ORDER														

ENTRY	TOTAL	TOTAL		MODEL	INFIT	OUTFIT	PTMEASUR-AL	EXACT	MATCH					
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	ITEM	G

19	1	208	4.78	1.01	.97	.3	.13	-1.6	.15	.09	99.5	99.5	D303	1
27	3	203	3.60	.60	.85	-.1	.12	-1.6	.27	.15	98.5	98.5	D520	1
29	9	206	2.42	.36	.93	-.2	.29	-1.4	.34	.25	95.5	95.6	D546	1
28	25	206	1.11	.24	.98	-.1	.66	-.9	.41	.38	86.6	88.5	D539	1
17	28	207	.95	.23	1.08	.6	.79	-.5	.38	.39	83.7	87.4	D234	1
24	9	81	.81	.40	.81	-.6	.93	.1	.48	.40	92.3	90.0	D411P	1
2	76	89	.62	.20	1.14	1.0	1.22	1.5	.48	.61	70.1	62.4	D009M	2
15	12	89	.55	.35	1.08	.4	.92	.0	.35	.39	88.5	87.3	D146M	1
25	28	126	.42	.24	.98	-.1	.78	-.7	.46	.43	80.8	80.8	D454S	1
5	26	116	.37	.26	.91	-.6	.70	-1.0	.52	.45	81.6	81.2	D052F	3
30	142	206	.23	.11	1.12	1.1	1.11	.8	.61	.65	55.9	61.2	D554	B
1	140	119	.16	.17	1.23	1.8	1.21	1.6	.55	.62	60.7	63.5	D009F	2
23	41	126	-.27	.22	.78	-2.2	.63	-1.9	.62	.48	83.2	75.9	D411S	1
11	139	127	-.32	.14	.84	-1.4	.79	-1.8	.69	.66	65.9	56.2	D095S	5
21	186	208	-.37	.10	1.00	.0	1.02	.2	.68	.68	52.9	56.4	D377	8
13	64	208	-.44	.18	.78	-2.8	.63	-2.5	.62	.49	83.3	77.0	D130	1
8	138	127	-.48	.13	1.07	.6	1.09	.7	.63	.68	52.4	52.6	D065S	4
6	24	88	-.57	.28	1.11	.8	1.09	.4	.41	.47	77.9	78.1	D052M	3
26	21	80	-.59	.30	.98	-.1	1.16	.7	.47	.48	80.5	78.8	D454P	1
18	70	208	-.62	.17	.86	-1.7	.82	-1.2	.57	.50	78.4	75.7	D277	1
10	76	207	-.80	.17	.80	-2.8	.75	-1.9	.62	.51	81.8	74.6	D071	1
12	85	81	-.89	.18	1.28	1.7	1.24	1.5	.63	.67	56.4	55.0	D095P	5
9	79	81	-.96	.16	.96	-.2	.93	-.4	.73	.68	53.8	51.8	D065P	4
16	87	208	-1.10	.17	1.00	.0	1.10	.8	.51	.51	75.0	73.3	D215	1
7	91	207	-1.22	.17	1.05	.7	1.12	1.0	.48	.52	71.4	72.9	D056	1
3	59	119	-1.27	.22	1.02	.3	.99	.0	.50	.51	69.2	72.2	D038F	1
22	283	207	-1.29	.10	1.33	3.2	1.30	2.8	.63	.71	53.7	50.2	D388	9
14	63	119	-1.45	.22	.97	-.3	.98	-.1	.53	.51	71.8	72.5	D146F	1
20	328	208	-1.63	.10	1.16	1.6	1.19	1.9	.65	.70	51.5	52.3	D331	7
4	41	88	-1.74	.25	.98	-.1	.90	-.5	.53	.51	70.9	72.0	D038M	1

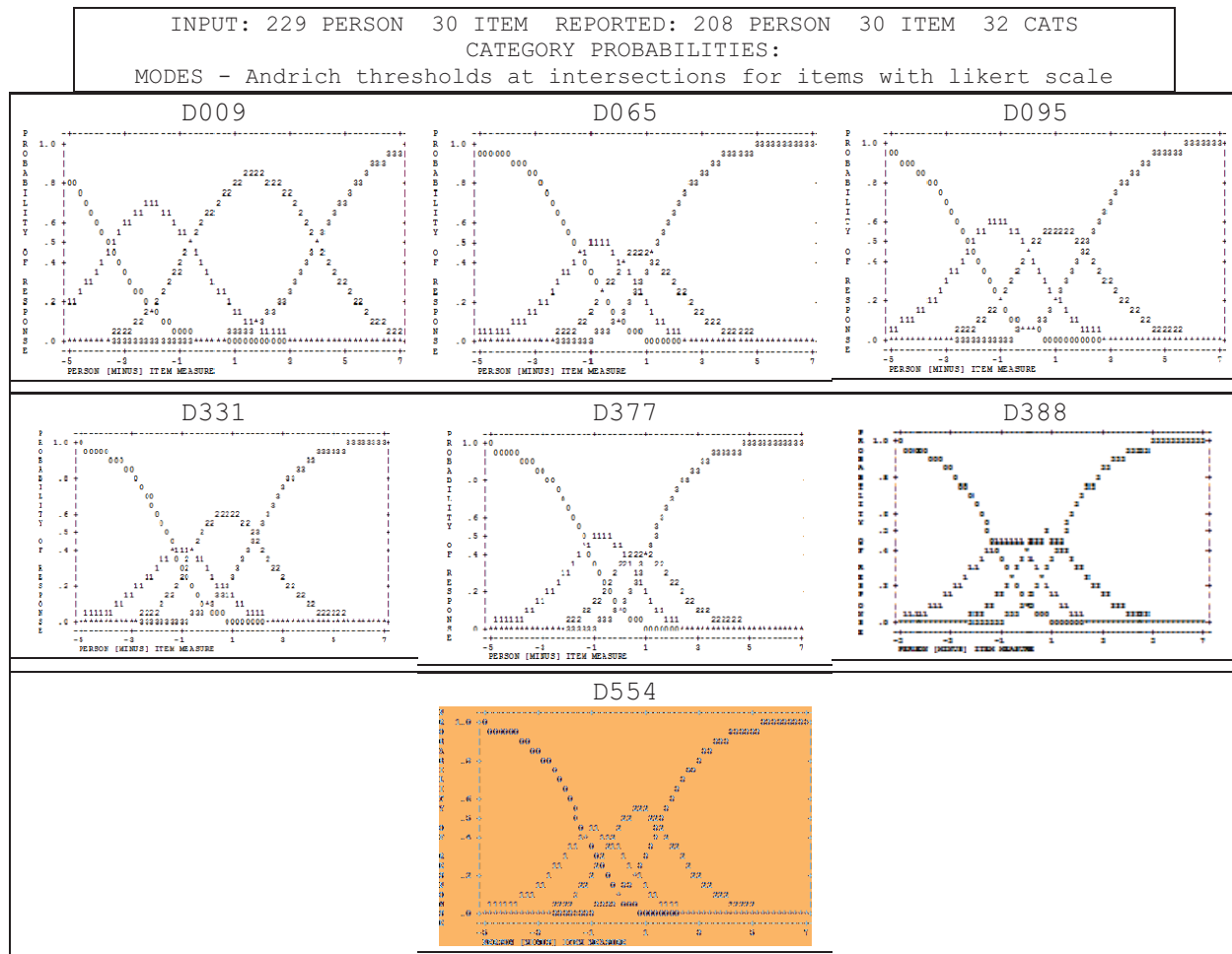
MEAN	79.1	151.8	.00	.25	1.00	.0	.89	-.1			74.1	73.1		
P.SD	76.9	53.5	1.45	.18	.14	1.3	.30	1.3			14.3	14.2		

Table W13. TABLE 12.2 DEP SCALE - WINSTEPS 3.90.0

DICHOTOMOUS MODEL						GROUPED RATING SCALE MODEL					
INPUT: 229 PERSON 30 ITEM						INPUT: 229 PERSON 30 ITEM					
REPORTED: 208 PERSON 30 ITEM 2 CATS						REPORTED: 208 PERSON 30 ITEM 32 CATS					
MEASURE PERSON - MAP - ITEM						MEASURE PERSON - MAP - ITEM - 50% Cumulative probabilities (Rasch-Thurstone thresholds)					
<more> <rare>						<more> <rare>					
6	+					5	+				
								D303	.1		
										D009M	.3
		D303								D009F	.3
5	+					4	+				
								D520	.1		
4	+	D520				3	+				
							T				
								D546	.1		
3	T						#				
	+					2	+				
#		D546								D554	.3
							.				
							S				
	.						.# T			D095P	.3
							#	D539	.1		
2	+					1	# +	D234	.1		D377 .3
##								D411P	.1		
T S							#				
.							###	D454S	.1		D065P .3
.##		D539					###	D052F	.1		D331 .3
.							.##			D009M .2	D388 .3
.#	+	D234				0	##	+M		D554	.2
.		D411P					. S			D377	.2
##		D009M D146M				.#####				D095S	.2
.##		D052F D454S D554						D411S	.1		
##							.##	D130	.1		
.# S							###	D454P	.1		
0	# +M	D095S D377				.#####				D095P	.2
#####		D065S						D071	.1		
#		D411S				-1	#####	+			
.#		D009F D130					.####		D215	.1	
####		D052M D277 D454P					#		D038F	.1	D388 .2
####		D071							D554	.1	
-1	.###	+ D065P D095P				.#####	M S	D146F	.1		
.##		D215					###		D038M	.1	
.##### M	D056	D388					####		D377	.1	
###		D038F				-2	.#	+	D065S	.1	D331 .2
####	S	D146F					.#####				
#							#####				
-2	#####	+ D038M				.#####		D095S	.1		
							.#####				
.###		D331					.# S T	D388	.1		
.#####						-3	####	+			
.							.#		D331	.1	
####							.##		D009F	.1	
-3	##### S+										
	T										
							.###				
.						-4	#	+			
.#####								T			
.###											
-4	#####	+					#				
	<less> <freq>						#				
EACH "##" IS 2: EACH "." IS 1							#				
							##	+			
							<less> <freq>				
							EACH "##" IS 2: EACH "." IS 1				

DEP Scale: the Rasch Grouped Rating Scale Model

We specified a Partial Credit Model with 4 categories. We run the program and we analyzed the order of the thresholds. We decided to estimate a Grouped Rating Scale model where groups were defined by the items. Some items seem to support the likert scale (well ordered thresholds: Tab. W14) but for the others it seems much better to maintain the binary coding (see Table 2): we kept recoding as binary the items

Table W14. TABLE 3.14 DEP SCALE: GROUPED RATING SCALE MODEL – WINSTEPS 3.90.0

most underfitting until the INFIT and OUTFIT MNSQ were less than 1.4. The summary statistics and other information are reported in Tab. W11, W12, and W13. The main results from the introduction of the likert scale for some of the items, were that person Reliability growth to 0.87 (from 0.82 for the dichotomous model), and the Model RMSE decreased to 0.49 (from 0.67 for the dichotomous model): the reduction was quite sensible (-27%) and a greater precision in the persons measure estimates was so achieved. For items the Model RMSE decreased to 0.30 from 0.33 (dichotomous model). The Log-likelihood chi-squared for the Grouped Rating Scale model was 4945.9214 with approximately 4949 d.f. (Tab. W11); the Log-likelihood chi-squared for the dichotomous model was 3223.5770 with approximately 3273 d.f. (Tab. W9). The p-value of the generalized likelihood test was zero, so rejecting the dichotomous model hypothesis in favor of the Grouped Rating Scale model: this means that the use of the likert scale, at least for some items, would improve significantly the measurement process. Confronting the item-map (Tab. W13) for the Grouped Rating Scale and the dichotomous models, we observed moreover that the introduction of the likert scale for some of the easiest and hardest items reduced the lack of items especially in the lower tail of the distribution: this is one of the reason for the reduction of the Model RMSE. Finally, from Figure 1 we may observe that most of the items lie inside the identity line bands, therefore the invariance property is substantially satisfied.

The ANX Scale

ANX Scale: the Rasch Dichotomous Model

To begin with, point-measure correlations were all positive in the range (0.26, 0.62). The first component of the PCA of the standardized residual explained 1.96 points of residual variance. The disattenuated correlations between person measures, calculated using subsets of items representing the opposite poles of the factors, show a couple of item cluster with a level of 0.87, the other being 1. The largest residual correlations shows the presence of a couple of items (A290, A556) highly dependent: A290 "*I worry over money and business*" and A556 "*I worry a great deal over money*". We decided to eliminate A556 because slightly more misfitting and rerun the program. The PCA of the standardized residuals showed 1.92 points of variance explained by first component; all disattenuated correlations between person measures, calculated using subsets of items representing the opposite poles of the factors, were at level 1. The largest standardized residual correlations do not show any dependence between items. Two items (A223, A208) have a large INFIT and OUFIT MNSQ, significantly different from 1. We decided to eliminate them from the set of items for the analysis and rerun the program. The PCA of the standardized residuals showed 1.87 points of variance explained by first component, again all disattenuated correlations between person measures, calculated using subsets of items representing the opposite poles of the factors, were all at level 1. INFIT MNSQ were in the range (0.86, 1.19), OUTFIT MNSQ (0.77, 1.41): no one significantly underfitting, some significantly overfitting. Some person were misfitting but before to delete them we went through the analysis of DIF.

ANX Scale: the Analysis of DIF

No SEX DIF was observed, while some item presented DIF for AGE (Tab. W15). We considered for DIF(AGE): A031, A140, A196, A301, A496. To take explicitly account of DIF we create new items, with appropriate missing values and rerun the program. We deleted the 23 most underfitting persons and rerun the program. All point-measure correlations were all positive in the range (0.32, 0.67). The first component of the PCA of the standardized residual explained 1.8 points of residual variance. The disattenuated correlations between person measures, calculated using subsets of items representing the opposite poles of the factors, were all 1. The largest residual correlations do not show the presence of dependent items. INFIT MNSQ were in the range (0.82, 1.28), OUTFIT MNSQ (0.52, 1.44): no one significantly underfitting, nor significantly overfitting. The summary statistics and other information regarding this last model are reported in Tab. W16, W17, and W20. Item's reliability was 0.96 and Model RMSE 0.23; person's reliability was 0.81 and their Model RMSE was 0.63. In order to reduce measurement error and to increase reliability, we introduced in the analysis the likert scale originally used to collect the data. This may help again to reduce the lack of items observed in the item-map (Tab. W20), in the lower tail of the distribution of person measures.

Table W15. TABLE 30.1 ANX SCALE: DICHOTOMOUS MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 23 ITEM REPORTED: 229 PERSON 19 ITEM 2 CATS																			

--																			
DIF FOR AGE (1 = son, 2 = parent)																			

	PERSON	Obs-Exp	DIF	DIF	PERSON	Obs-Exp	DIF	DIF	DIF	JOINT	Rasch-Welch	Mantel-Haenszel	Size Active	ITEM					
	CLASS	Average	MEASURE	S.E.	CLASS	Average	MEASURE	S.E.	CONTRAST	S.E.	t	d.f.	Prob.	Chi-squ	Prob.	CUMLOR Slices			
									Name							Number			

	1	.03	1.09	.25	2	-.06	2.00	.39	---	-.91	.46	-1.97	170	.0509	3.5277	.0604	-1.10	17	3
										A031									

1	.04	-.91	.20	2		-.07	-.26	.26	-.64	.33	-1.93	184	.0552	2.9031	.0884	-.68	17	5		
1	.05	-1.51	.21	2		-.08	-.84	.26	A140	-.67	.33	-2.02	185	.0451	1.6281	.2020	-.56	17	7	
1	-.06	-.29	.21	2		.10	-1.14	.26	A196	.85	.33	2.56	186	.0112	5.1831	.0228	1.02	17	13	
1	.06	-1.09	.21	2		-.10	-.22	.27	A301	-.87	.34	-2.58	182	.0106	4.6958	.0302	-.83	17	21	
										A496										

Table W16. TABLE 3.1 ANX SCALE: DICHOTOMOUS MODEL - WINSTEPS 3.90.0

SUMMARY OF 193 MEASURED (NON-EXTREME) PERSON								
	TOTAL			MODEL	INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	7.6	19.9	-.75	.61	1.00	.1	.92	.1
P.SD	4.4	.5	1.44	.15	.23	.9	.38	.8
S.SD	4.4	.6	1.44	.15	.23	.9	.38	.8
MAX.	18.0	20.0	2.73	1.05	1.57	2.1	2.05	2.3
MIN.	1.0	13.0	-3.52	.51	.51	-2.1	.22	-1.7
REAL RMSE	.66	TRUE SD	1.28	SEPARATION	1.95	PERSON RELIABILITY	.79	
MODEL RMSE	.63	TRUE SD	1.29	SEPARATION	2.05	PERSON RELIABILITY	.81	
S.E. OF PERSON MEAN = .10								
MINIMUM EXTREME SCORE: 13 PERSON 6.3%								
DELETED: 23 PERSON								
SUMMARY OF 206 MEASURED (EXTREME AND NON-EXTREME) PERSON								
	TOTAL			MODEL	INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	7.2	19.9	-.99	.69				
P.SD	4.6	.5	1.69	.33				
S.SD	4.7	.5	1.69	.33				
MAX.	18.0	20.0	2.73	1.85				
MIN.	.0	13.0	-4.80	.51				
REAL RMSE	.79	TRUE SD	1.50	SEPARATION	1.90	PERSON RELIABILITY	.78	
MODEL RMSE	.77	TRUE SD	1.51	SEPARATION	1.96	PERSON RELIABILITY	.79	
S.E. OF PERSON MEAN = .12								
PERSON RAW SCORE-TO-MEASURE CORRELATION = .97								
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .80 SEM = 2.09								
SUMMARY OF 25 MEASURED (NON-EXTREME) ITEM								
	TOTAL			MODEL	INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	59.0	164.1	.00	.23	1.00	.0	.91	-.2
P.SD	33.1	51.9	1.23	.06	.11	1.2	.22	1.1
S.SD	33.8	53.0	1.25	.06	.11	1.2	.22	1.2
MAX.	138.0	206.0	2.48	.46	1.28	3.5	1.44	3.1
MIN.	6.0	81.0	-2.07	.17	.82	-1.9	.52	-1.9
REAL RMSE	.24	TRUE SD	1.20	SEPARATION	5.04	ITEM RELIABILITY	.96	
MODEL RMSE	.23	TRUE SD	1.20	SEPARATION	5.13	ITEM RELIABILITY	.96	
S.E. OF ITEM MEAN = .25								
ITEM RAW SCORE-TO-MEASURE CORRELATION = -.78								
Global Statistics:								
Active PERSON: 206								
Active ITEM: 25								
Active datapoints: 4102 = 79.7% of Active+Missing datapoints								
Missing datapoints: 1048 = 20.3% of Active+Missing datapoints								
Standardized residuals N(0,1) mean: -.01 P.SD: .93								

Log-likelihood chi-squared: 3505.4267 with approximately 3540 d.f., probability = .6570
Global Root-Mean-Square Residual: .3750 with expected value: .3748
Capped Binomial Deviance: .1857 for 4102.0 dichotomies with expected value: .1873

Table W17. TABLE 13.1 ANX SCALE: DICHOTOMOUS MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 25 ITEM REPORTED: 206 PERSON 25 ITEM 2 CATS													

PERSON: REAL SEP.: 1.90 REL.: .78 ... ITEM: REAL SEP.: 5.04 REL.: .96													
ITEM STATISTICS: MEASURE ORDER													

ENTRY	TOTAL	TOTAL		MODEL	INFIT		OUTFIT		PTMEASUR-AL		EXACT MATCH		
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
-----+-----+-----+-----+-----													

	4	6	81	2.48	.46	1.02	.2	.68	-.1	.31	.30	91.7	91.6
					A031P								
	20	19	204	2.17	.26	.94	-.3	.52	-1.0	.38	.32	90.1	90.3
					A463								
	12	20	205	2.10	.26	1.03	.2	.72	-.5	.33	.33	90.6	89.9
					A299								
	2	33	206	1.39	.21	1.00	.1	.84	-.4	.41	.40	85.0	84.6
					A030								
	3	20	125	1.39	.27	1.11	.7	.87	-.2	.36	.40	81.8	84.9
					A031S								
	8	42	206	1.01	.20	1.16	1.5	1.18	.7	.36	.44	78.8	81.5
					A170								
	11	42	206	1.01	.20	.98	-.2	.74	-1.0	.47	.44	78.8	81.5
					A273								
	5	47	204	.81	.19	.94	-.6	.85	-.6	.49	.46	82.7	79.9
					A039								
	21	62	205	.30	.18	.91	-1.1	.74	-1.5	.55	.50	76.0	76.1
					A469								
	17	67	205	.14	.18	1.04	.5	.99	.0	.50	.51	72.9	75.1
					A405								
	7	29	81	-.09	.28	.90	-.8	.77	-.9	.59	.54	76.4	73.5
					A140P								
	23	29	81	-.09	.28	.99	-.1	1.25	1.0	.53	.54	79.2	73.5
					A496P								
	1	79	202	-.26	.17	1.01	.1	1.05	.4	.53	.54	72.0	73.5
					A015								
	13	52	125	-.39	.22	.91	-1.0	.82	-1.1	.59	.53	74.4	73.9
					A301S								
	15	90	206	-.54	.17	.91	-1.2	.81	-1.5	.60	.55	76.7	73.1
					A305								
	18	95	205	-.69	.17	.91	-1.2	.89	-.9	.60	.56	75.5	73.1
					A408								
	10	37	81	-.72	.28	.95	-.4	.95	-.2	.61	.59	72.2	72.2
					A196P								
	26	96	206	-.72	.17	1.28	3.5	1.44	3.1	.42	.56	62.2	73.1
					A290								
	6	62	124	-.89	.22	1.19	1.9	1.26	1.5	.45	.55	67.5	73.7
					A140S								
	24	105	205	-.99	.17	1.09	1.2	1.18	1.4	.52	.57	72.9	73.8
					A509								

	22	68	124	-1.15	.22	1.04	.4	1.04	.3	.54	.56	71.7	74.1
						A496S							
	14	43	81	-1.19	.28	.91	-.7	.74	-1.1	.66	.61	75.0	74.7
						A301P							
	16	113	205	-1.23	.17	.86	-1.9	.76	-1.9	.65	.58	80.2	74.6
						A339							
	9	81	125	-1.78	.23	.82	-1.7	.68	-1.6	.65	.56	81.8	76.8
						A196S							
	19	138	204	-2.07	.19	1.15	1.5	1.09	.5	.53	.59	77.5	79.2
						A415							
	-----+-----+-----+-----+-----+-----+-----												
	MEAN	59.0	164.1	.00	.23	1.00	.0	.91	-.2			77.7	77.9
	P.SD	33.1	51.9	1.23	.06	.11	1.2	.22	1.1			6.8	5.9
	-----+-----+-----+-----+-----+-----+-----												

Table W18. TABLE 3.1 ANX SCALE: GROUPED RATING SCALE MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 25 ITEM REPORTED: 206 PERSON 25 ITEM 34 CATS									

SUMMARY OF 202 MEASURED (NON-EXTREME) PERSON									
		TOTAL			MODEL		INFIT		OUTFIT
		SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD

	MEAN	13.6	19.9	-.74	.43	1.01	.0	.94	.0
	P.SD	7.0	.5	1.18	.11	.43	1.2	.40	1.1
	S.SD	7.0	.5	1.18	.11	.43	1.2	.40	1.1
	MAX.	30.0	20.0	1.81	1.03	2.86	4.1	2.25	3.2
	MIN.	1.0	13.0	-3.99	.37	.30	-3.0	.25	-3.1

	REAL RMSE	.47	TRUE SD	1.08	SEPARATION	2.27	PERSON RELIABILITY	.84	
	MODEL RMSE	.44	TRUE SD	1.09	SEPARATION	2.47	PERSON RELIABILITY	.86	
	S.E. OF PERSON MEAN = .08								

MINIMUM EXTREME SCORE: 4 PERSON 1.9%									
DELETED: 23 PERSON									
SUMMARY OF 206 MEASURED (EXTREME AND NON-EXTREME) PERSON									
		TOTAL			MODEL		INFIT		OUTFIT
		SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD

	MEAN	13.3	19.9	-.82	.46				
	P.SD	7.2	.5	1.32	.22				
	S.SD	7.2	.5	1.32	.22				
	MAX.	30.0	20.0	1.81	1.84				
	MIN.	.0	13.0	-5.23	.37				

	REAL RMSE	.53	TRUE SD	1.20	SEPARATION	2.25	PERSON RELIABILITY	.84	
	MODEL RMSE	.51	TRUE SD	1.22	SEPARATION	2.40	PERSON RELIABILITY	.85	
	S.E. OF PERSON MEAN = .09								

PERSON RAW SCORE-TO-MEASURE CORRELATION = .96									
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .81 SEM = 3.14									
SUMMARY OF 25 MEASURED (NON-EXTREME) ITEM									
		TOTAL			MODEL		INFIT		OUTFIT
		SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD

	MEAN	109.9	164.1	.00	.18	1.00	.0	.93	-.2
	P.SD	98.0	51.9	.94	.08	.10	1.1	.16	.9
	S.SD	100.0	53.0	.96	.08	.11	1.1	.16	1.0

Table W20.TABLE 12.2 ANX SCALE - WINSTEPS 3.90.0

DICHOTOMOUS MODEL		GROUPED RATING SCALE MODEL	
INPUT: 229 PERSON 25 ITEM		INPUT: 229 PERSON 25 ITEM	
REPORTED: 206 PERSON 25 ITEM 2 CATS		REPORTED: 206 PERSON 25 ITEM 2 CATS	
MEASURE	PERSON - MAP - ITEM	MEASURE	PERSON - MAP - ITEM - 50% Cumulative probabilities (Rasch-Thurstone thresholds)
3	<more> <rare>	3	<more> <rare>
	+		+
	X		
	T A031P		A299 .3
	XX		A463 .3
	X		A031P .1
2	T A299 A463	2	+ T A273 .3
	+		
	X		.
	XXXX		# T
			.
	XXXXX		# A031S .1
	XXX A030 A031S		# A030 .1 A463 .2
	X S	1	# +S A170 .1
	XXXXX		#
1	+ A170 A273		.## A039 .1
	XXX		#
	XXXXXX S A039		A273 .2 A408 .3
	XX		A301P .3
	XXXXX		A415 .3
	XXXXX		.#####
	XXXXXXXXXX A469		.## A469 .1
	XXXXXX A405	0	#### +M A299 .1
0	XXXXXXXXXXXX +M		.## A496P .1
	XXXX A140P A496P		.#### A015 .1
	A015		####
	XXXXXXXXXXXXX A301S		##### A290 .1 A305 .2
	A305		.## M A140S .1
	XXXXXXXXXXXXX		#### A273 .1 A301P .2
	XXXXXXXXXXXXX M A196P A290 A408		A509 .1
	XXXXX A140S	-1	### +S A339 .2
	XX + A509		A496S .1
	XXXXXXXXXXXXX A496S		.#####
	XXXX S A301P A339		.#
	X		.###
	XXXXXXXXXXXXX		.## A196S .1
	XXXXX		.# A301S .1 A415 .2
	XXXXXXXXXXXXX		## A408 .1
	XXXXX		.#### S T
	XXXXXXXXXXXXX	-2	+
	XXXXX		.#### A339 .1
	XXXXXXXXXXXXX A196S		
	X		.###
-2	XXXXXXX +		
	S A415		.##
	XXXXX		
	XX T	-3	+
			T A415 .1
	XXXXXXXXXXXXX		###
	+		
	XXXXXXX		
	XXXXXXX	-4	.### +
	T		<less> <freq>
			EACH "#" IS 2: EACH "." IS 1
-4	XXXXXXXXXXXXX +		
	<less> <freq>		

ANX Scale: the Rasch Grouped Rating Scale Model

We estimated initially a Partial Credit Model with 4 categories. We run the program and we analyzed the order of the thresholds. We decided to estimate a Grouped Rating Scale model where groups were defined by the items. Some items seem to support the likert scale (well ordered thresholds: Tab. W21) but for others it seems much better to maintain the binary coding (see Table 2): we kept recoding as binary the items most underfitting until the INFIT and OUTFIT MNSQ were less than 1.4. The results are reported in Tab. W18, W19, and W20. The main results from the introduction of the likert scale for some of the items, were that person Reliability growth to 0.86 (from 0.80 for the dichotomous model), and the Model RMSE decreased to 0.44 (from 0.63 for the dichotomous model): the gain in this indicator was quite sensible (-30%) and a greater precision in the persons measure estimates was so achieved. For item the Model RMSE decreased to 0.19 from 0.23 (dichotomous model). The Log-likelihood chi-squared for the Grouped Rating Scale model was 5507.5628 with approximately 5527 d.f. (Tab. W18); the Log-likelihood chi-squared for the dichotomous model was 3505.4267 with approximately 3540 d.f. (Tab. W16). The generalized likelihood test based on the -2 Log-likelihood ratio statistics has a p-value of zero: this means that the use of the likert scale, at least for some items, would improve significantly the measurement process. Confronting the item-map (Tab. W20) for the Grouped Rating Scale and the dichotomous models, we observed moreover that the introduction of the likert scale for some of the easiest and hardest items reduced the lack of items especially in the lower tail of the distribution: this is one of the reason for the reduction of the Model RMSE. Finally, from Figure 1 we may observe that most of the items lie inside the identity line bands, therefore the invariance property is substantially satisfied also in this case.

Comment of the Main Results

Analysis of the Excluded Items

The item excluded from the Rasch analysis are reported in table 2. For what it concerns the Mt scale we observe the following. D071 and D411, belonging also to DEP scale, have been excluded because highly correlated with other items of the Mt scale. For what it concerns the others, we can make the following observations. M20 "*I am very seldom troubled by constipation*": is a generic item whose causes could be the most various, not necessarily linked to the construct we wish to measure. This item, moreover, also belonging to the basic scale (2) D that also measure depression, once included in the DEP scale, under fits significantly. M43 "*My judgment is better than it ever was*": 1 if FALSE, doesn't exclude that the judgment capacity was already good in the past; it also belongs to the basic scale (2) D, once included in the DEP scale, under fits significantly. M081 "*I think most people would lie to get ahead*": a generic item whose answer may represent a social stereotype. M131 "*When I was child, I belonged to a group of friends that tried to be loyal through all kinds of troubles*": the item is ambiguous, and is not clear if the answer TRUE or FALSE is referred to the first part ("*I belong to a group...*") or to the second part of the question ("*...that tried to be loyal through all kinds of troubles*"). M233 "*I have difficulty in starting to do things*": the item belongs also to the (2) D scale and once included in the DEP scale, shows good fit indices and could be reasonably included with the other item of the DEP scale. M269 "*If several people find themselves in trouble, the best thing for them to do is to agree upon a story and stick to it*": a generic item whose answer may represent a social stereotype. M357 and M449: is not clear why a response TRUE to these items should represent maladjustment. For what it concerns, the DEP scale, D516 has been excluded because highly correlated with other items included in the scale (D506 and D512). For what it concerns the other items, excluded because under fitting, we may observe that their content belongs to the suicide area and may represent a construct related but not identical to the DEP scale. In order to verify this hypothesis we Rasch analyzed all these item and we found that, excluding the item D092 (actually out of the suicide domain) because misfitting, the others satisfy very well a Grouped

Rating Scale Model with the following characteristics: items D003, D075, D082, D306, D399, D400 were expressed with a common likert scale 0,1,2,3, while D246, D506, D512 were binary; items D082 shows DIF(SEX), while D003 and D512 show DIF(AGE). After the exclusion of the 31 most underfitting persons, the items show a point – measure correlation in the range (0.25, 0.67); the PCA shows 1.6 points of residual variance for the first component and the disattenuated correlations between person measures, calculated using subsets of items representing the opposite poles of the factors, show value 1 or near 1. The largest standardized residual correlations do not show any dependence between items. INFIT MNSQ were in the range (0.78, 1.17), OUTFIT MNSQ (0.33, 1.23): no one significantly underfitting, nor significantly overfitting. From Tab. W22 we may see that persons have a reliability of 0.70, a model RMSE of 0.62, a separation of 1.52 and a mean of -1.58. Items have a reliability of 0.97, a model RMSE of 0.24, a separation of 5.99, a mean set to 0 by default. Being this construct interpretable as “Attitude to Suicide” (SUI), it is clear that a person mean of -1.58 means that this population has very low attitude, with respect to the mean of the items, as we may expect from this sample with no clinical cases. The item measures are reported in Tab. W23 and the item map in Tab. W24: as we may observe to the hardest part of the items does not correspond any person. The hardest item is D506 “*I have recently considered killing myself*” (1 if TRUE), and this seem quite obvious, the hardest item after this is D246 “*I believe my sin unpardonable*” (1 if TRUE); the easiest item for sons is D003 “*I wake up fresh and rested most mornings*” (1 if FALSE), while the easiest item for the overall sample is D399 “*The future is too uncertain for a person to make serious plan*” (1 if TRUE). An attenuated correlation of 0.76 between DEP person measures and SUI person measures, and the identity line represented in fig.1, show that SUI scale represents a dimension different from the DEP scale, and therefore there are highly depressed persons that do not think necessarily to suicide.

For what it concerns, finally, the ANX scale, the item A556 has been eliminated because correlated with other item of the scale. The other two items eliminated from the analysis are A208 and A223: actually, we do not see any reason they should be related to anxiety.

Table W21. TABLE 3.14 ANX SCALE: GROUPED RATING SCALE MODEL – WINSTEPS 3.90.0

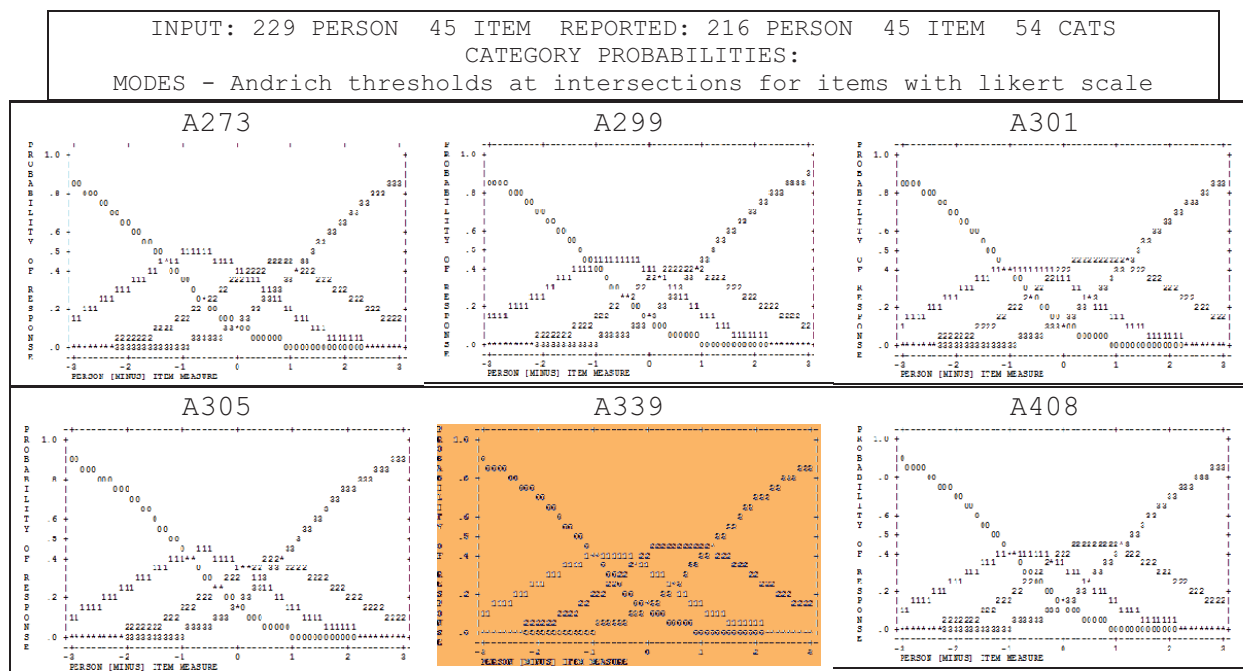


Table W23. TABLE 13.1 SUI SCALE: GROUPED RATING SCALE MODEL - WINSTEPS 3.90.0

INPUT: 229 PERSON 13 ITEM REPORTED: 198 PERSON 12 ITEM 6 CATS WINSTEPS														
3.90.0														

PERSON: REAL SEP.: 1.37 REL.: .65 ... ITEM: REAL SEP.: 5.96 REL.: .97														
ITEM STATISTICS: MEASURE ORDER														

ENTRY	TOTAL	TOTAL		MODEL	INFIT	OUTFIT	PTMEASUR-AL	EXACT MATCH						
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	ITEM	G

11	5	195	2.55	.46	.92	-.1	.33	-1.2	.25	.14	97.4	97.4	D506	1
7	7	198	2.24	.39	.92	-.1	.39	-1.2	.27	.17	96.4	96.4	D246	1
3	52	198	1.55	.16	.96	-.2	.82	-.7	.41	.41	74.0	77.2	D075	2
12	14	124	.93	.30	1.00	.1	.98	.1	.28	.28	88.5	88.5	D512S	1
13	13	71	.18	.33	.99	.0	.89	-.2	.36	.35	83.1	81.9	D512P	1
10	150	197	-.03	.11	1.12	1.2	1.11	1.0	.62	.57	56.9	55.7	D400	2
4	101	115	-.17	.14	1.11	.9	1.23	1.5	.49	.59	50.0	54.3	D082F	2
5	81	82	-.70	.16	.84	-1.1	.79	-1.3	.66	.59	53.1	49.9	D082M	2
2	101	73	-1.39	.16	1.03	.3	1.06	.5	.55	.62	56.2	49.5	D003P	2
8	299	198	-1.58	.10	.93	-.8	.94	-.6	.62	.64	52.0	49.8	D306	2
9	308	197	-1.69	.10	1.17	1.7	1.14	1.5	.67	.64	46.2	49.8	D399	2
1	210	125	-1.89	.13	.78	-2.0	.83	-1.4	.62	.66	53.7	51.6	D003S	2

MEAN	111.7	147.8	.00	.21	.98	.0	.88	-.2			67.3	66.8		
P.SD	104.5	52.1	1.48	.12	.11	1.0	.27	1.0			18.5	18.9		

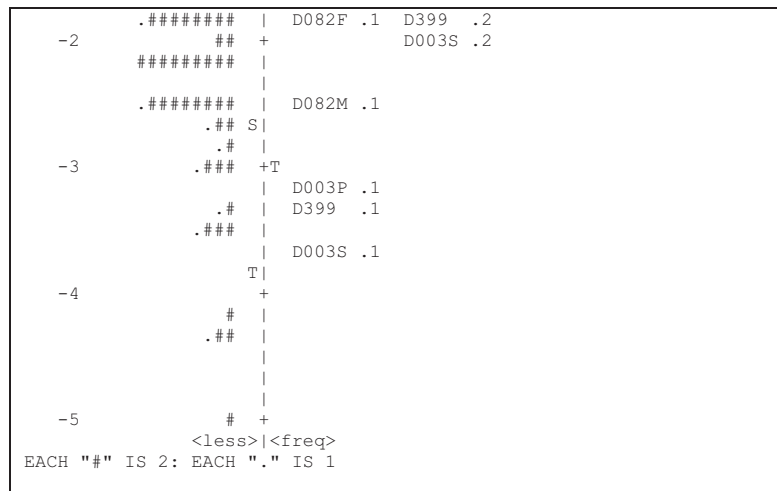
Table W24. TABLE 12.2 SUI SCALE: GROUPED RATING SCALE MODEL - WINSTEPS 3.90.0

```

INPUT: 229 PERSON 13 ITEM REPORTED: 198 PERSON 12 ITEM 6 CATS
MEASURE PERSON - MAP - ITEM
50% Cumulative probabilities (Rasch-Thurstone thresholds)

    <more>|<rare>
      +
      |
      |
      |
      | D075 .3
      |
3     +T
      |
      | D506 .1
      |
      | D246 .1
      |
2     +
      |
      | D400 .3
      | D082F .3
      | S
      |
      | D075 .2
      | D082M .3
      |
1     + D512S .1
      |
      | T
      | #
      | #
      | #
      | ##### D512P .1
      | #
      | +M
0     .##### D075 .1 D400 .2
      | .##### D082F .2
      | .##### S
      |
      | ##### D082M .2
      |
-1    .#
      | #####
      | .###
      | ##### S
      | M
      |
      | D003P .2
      | D306 .2
      |
      | D400 .1

```



Analysis of the Included Items

The item included from the Rasch analysis are reported in table 2.

The final number of item from the Rasch analysis for Mt scale is 31 over 41: 13 of this are expressed with a likert scale, while the other 18 are binary. This choice, although justified by the fit indices of the thresholds, is justified also by the content of the items: for example the item D003 “*I wake up fresh and rested most mornings*” allows very well the possibility of an answer at likert scale, so does for example the item M325 “*I have more trouble concentrating than others seem to have*”. While the answer to items like M002 “*I have good appetite*” or A469 “*I sometimes feel that I’m about to pieces*” is better expressed on a FALSE/TRUE basis. 6 items present differential item functioning with respect to SEX, while 8 with respect to AGE. For example the item M028 “*I am bothered by an upset stomach several times a week*” (1 if TRUE) have a difficulty of 1.75 (the second most hardest item) for Male persons but only 0.43 (an average difficulty) for Female persons: evidently to have stomach problems is more easier for Female person while the same symptoms in case of a Male person means that the he has a high level of maladjustment. In the case of M016 “*Once in a while I think of things too bad to talk about*”, we observe the opposite: the difficulty for Female persons is 1.12, while only 0.19 for Men. In the case of DIF(AGE) we may observe that M002 “*I have a good appetite*” (1 if FALSE) is the most difficult item for Sons with a difficulty of 4.35, while for Parents the difficulty is only 1.12: this means that for a young person not having good appetite is signal of strong maladjustment. Instead the situation is reversed in the case of the item M073 “*I am certainly lacking in self-confidence*” (1 if TRUE), where a difficulty of 0.78 for Parents and -0.11 for Sons means that this fact is symptom of greater maladjustment for an old than young person. The most hardest item for all is M218 “*I have period of such great restlessness that I cannot sit long in a chair*” (1 if TRUE) that clearly could create some problems for studying, while the easiest item is M110 “*Most people will use somewhat unfair means to gain profit or an advantage rather than to lose it*” with an (average) difficulty of -3.02. The final number of items from the Rasch analysis for DEP scale is 22 over 33. As we already observed the greater part of items eliminated contribute to a separate dimension called “Attitude to Suicide” (SUI). 7 items are expressed as likert scale, while the other 15 are binary. In the first group D009 “*My daily life is full of things that keep me interested*” (1 if FALSE) has the most clearly defined thresholds indicating that this item is better expressed at an ordered rather than binary level. So does M095 “*I’m happy most of the time*” (1 if FALSE). Instead items like D052 “*I have not lived the right kind of life*” (1 if TRUE) or D146 “*I cry easily*” (1 if TRUE) are better answered on a FALSE/TRUE basis. 4 items present DIF(SEX) and 4 DIF(AGE). In the first group, for example, D146 “*I cry easily*” has a difficulty of 0.55 for Male persons and -1.45 for Female persons, indicating that crying for a Female person is not a strong symptom of maladjustment as is instead for a Men; the reversed situation is observed for the item D052 “*I have not lived the right kind of life*” with 0.37 for Female and -

0.57 for Male. For what it concerns DIF(AGE) the item D411 “*At times I think I am no good at all*”, with a difficulty of 0.81 for Parents and -0.27 for Sons clearly evidence that this is symptoms of greater maladjustment for older people. The opposite it happens in the case of the item D454 “*The future seems hopeless to me*” (1 if TRUE) with a difficulty of 0.42 for Sons and -0.59 for Parents: clearly this fact in a young person is index of more gravity than it is for a older person. The hardest item for all is D303 “*Most of the time I wish I were dead*” which however over fits greatly. The hardest items not clearly overfitting are D539 “*Lately I have lost my desire to work out my problems*”, with difficulty 1.11 and D234 “*I believe I am a condemned person*” with a difficulty of 0.95. The easiest item are instead D331 “*I am inclined to take things hard*” with difficulty of -1.63, and D388 “*I very seldom have spells of the blues*”. The final number of items from the Rasch analysis for ANX scale is 20 over 23. 8 items are expressed as likert scale, while the other 12 are binary. In the first group, A415 “*I worry quite a bit over possible misfortunes*” has the most clearly defined thresholds, together with item A301 “*I feel anxiety about something or someone almost all the time*”. In the second group items such as A030 “*I have nightmares every few nights*” or A170 “*I am afraid of losing my mind*” are better answered on a FALSE/TRUE basis. 5 items present DIF(AGE). For example, A031 “*I find hard to keep my mind on a task or job*” is the most hard item both for Sons and Parents, but for the first the difficulty is 1.35, while for the second 2.17; A301 “*I feel anxiety about something or someone almost of the time*” has difficulty -0.33 for Sons, but only -0.78 for Parents. The hardest item for all is A030 “*I have nightmare every few nights*” with a difficulty of 1.29, the easiest is A415 “*I worry quite a bit over possible misfortune*”.

Analysis of Person Measures

For what it concerns the person measures, from the Table 3 we may observe that Male persons are less maladjusted, less depressed and less anxious than Female persons. Sons are more maladjusted, more depressed and more anxious than Parents are. Daughters are the more maladjusted, depressed and anxious persons, at the opposite fathers are the less maladjusted, depressed and anxious persons. Looking at the correlations inside the family (Table 4) we observe that son’s Mt measures are positively correlated with that of their mothers (0.29) and negatively correlated with their fathers (-0.34); the same is observed for DEP measures, but with very low level of correlation; for what it concerns ANX, measures of sons are positively correlated both with mothers (0.33) and fathers (0.26). With reference to the correlation between constructs (Figure 2) we observe the higher correlation between Mt and ANX (0.96) which reduces to 0.86 if we exclude from the ANX scale the items included in Mt. Also the correlation between Mt and DEP is high (0.93) and it reduces to 0.87 if we exclude from the DEP scale the items included in Mt. The correlation between DEP and ANX is 0.83.

Table 3. Average values of measures in the sample and their std errors

Mt: average values				DEP: average values				ANX: average values			
Sex\Relat	Son	Parent	Total	Sex\Relat	Son	Parent	Total	Sex\Relat	Son	Parent	Total
Female	-0.79	-1.05	-0.89	Female	-1.02	-1.74	-1.30	Female	-0.62	-0.85	-0.71
Male	-1.09	-1.45	-1.23	Male	-1.68	-2.32	-1.93	Male	-0.96	-1.23	-1.06
Total	-0.92	-1.22	-1.04	Total	-1.30	-1.99	-1.57	Total	-0.76	-0.99	-0.85

Mt: SE of Ave values				DEP: SE of Ave values				ANX: SE of Ave values			
Sex\Relat	Son	Parent	Total	Sex\Relat	Son	Parent	Total	Sex\Relat	Son	Parent	Total
Female	0.04	0.05	0.03	Female	0.04	0.06	0.03	Female	0.06	0.06	0.04
Male	0.05	0.07	0.04	Male	0.06	0.07	0.05	Male	0.06	0.08	0.05
Total	0.03	0.04	0.02	Total	0.03	0.05	0.03	Total	0.04	0.05	0.03

Table 4. Correlations between constructs inside the family

Mt: correlations				DEP: correlations				ANX: correlations			
Relation	Son	Mother	Father	Relation	Son	Mother	Father	Relation	Son	Mother	Father
Son	1.00			Son	1.00			Son	1.00		
Mother	0.26	1.00		Mother	0.09	1.00		Mother	0.28	1.00	
Father	-0.28	0.27	1.00	Father	-0.15	0.00	1.00	Father	0.21	0.35	1.00

Mt: disattenuated correlations				DEP: disattenuated correlations				ANX: disattenuated correlations			
Relation	Son	Mother	Father	Relation	Son	Mother	Father	Relation	Son	Mother	Father
Son	1.00			Son	1.00			Son	1.00		
Mother	0.29	1.00		Mother	0.11	1.00		Mother	0.33	1.00	
Father	-0.34	0.32	1.00	Father	-0.18	0.00	1.00	Father	0.26	0.43	1.00

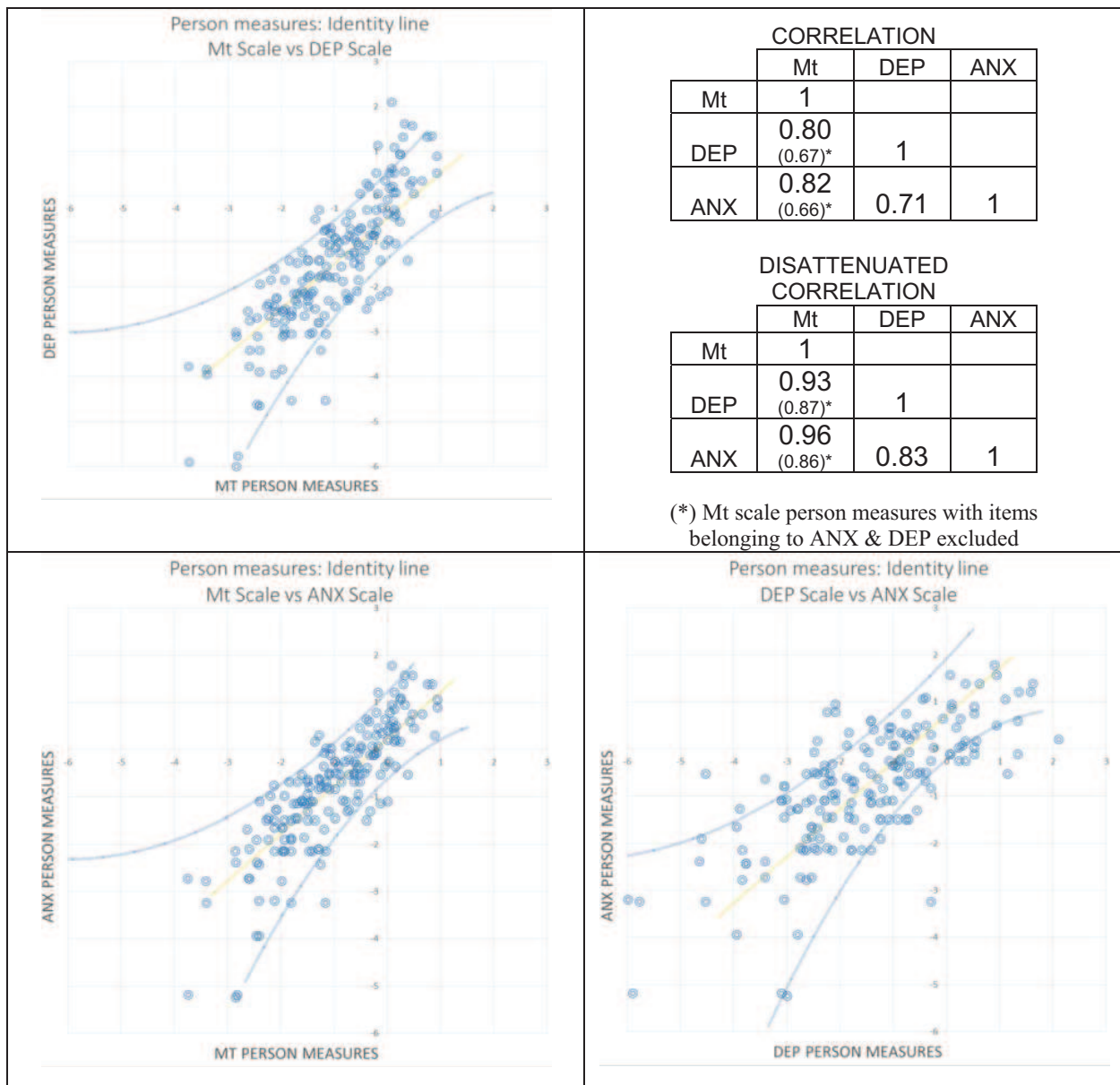


Figure 2. Identity lines and correlations between constructs in the sample

Table 5. Items from MMPI-2 belonging to one of the Mt, DEP or ANX scales

N.	Domanda	T/F
2	Ho un buon appetito	F
3	Mi sveglio fresco e riposato quasi tutte le mattine	F
9	La mia vita di ogni giorno, è piena di cose che mi interessano	F
10	Sono capace di lavorare come lo sono sempre stato	F
15	Lavoro sotto una forte tensione nervosa	T
16	A volte penso a cose così cattive da non poterne parlare	T
20	Molto raramente soffro di stitichezza	F
28	Ho disturbi di stomaco diverse volte alla settimana	T
30	Spesso la notte ho degli incubi	T
31	Mi riesce difficile concentrarmi su un compito o su un lavoro	T
38	Ho avuto periodi di giorni, settimane o mesi in cui mi era difficile combinare qualcosa, perché non riuscivo a cominciare	T
39	Ho il sonno agitato e disturbato	T
43	La mia capacità di giudizio è migliore di quanto sia mai stata	F
52	Non ho vissuto la vita in maniera giusta	T
56	Vorrei essere felice così come sembrano esserlo gli altri	T
65	Di solito sono malinconico	T
71	In questo periodo trovo difficile mantenere la speranza di raggiungere qualcosa nella vita	T
73	Mi manca proprio la fiducia in me stesso	T
75	Generalmente sento che la vita merita di essere vissuta	F
81	Penso che la maggior parte delle persone mentirebbe per farsi strada nella vita	T
82	Faccio molte cose di cui dopo mi pento (più di quanto sembra lo facciano gli altri)	T
92	Non mi sembra di preoccuparmi per quello che mi succede	T
95	Sono quasi sempre contento	F
110	La maggior parte delle persone userebbe mezzi sleali per avere un profitto o un vantaggio piuttosto che perderlo.	T
130	A volte mi sento proprio inutile.	T
131	Quando ero bambino facevo parte di un gruppo di amici in cui si cercava di rimanere uniti di fronte a qualsiasi tipo di avversità.	F
140	La maggior parte delle notti mi addormento senza che pensieri o idee particolari mi disturbino.	F
146	Piango facilmente.	T
148	Mai mi sono sentito bene come adesso nella mia vita.	F
152	Non mi stanco facilmente.	F
170	Ho paura d'impazzire.	T
196	Sono spesso preoccupato per qualcosa.	T
208	Raramente (o mai) mi sento battere forte il cuore o mancare il respiro.	F
215	Passo molto tempo a rimuginare	T
218	Ho periodi di tale irrequietezza da non poter stare a lungo seduto.	T
223	Non credo di essere più nervoso degli altri.	F
233	Trovo difficoltà nel cominciare a fare le cose.	T
234	Credo di essere perseguitato dal destino.	T
246	Credo che i miei peccati siano imperdonabili.	T
269	Se parecchie persone si trovano nei pasticci, la miglior cosa che possono fare è di mettersi d'accordo su una versione dei fatti e mantenerla in ogni caso.	T
273	La vita è quasi sempre una fatica per me.	T
277	Mi sento spesso solo, anche quando sono in mezzo alla gente.	T
290	Sono preoccupato per questioni di affari o di denaro.	T
299	Mi è impossibile mantenere l'attenzione su una data cosa.	T
301	Quasi sempre sto in ansia per qualcuno o per qualcosa.	T
302	Perdo facilmente la pazienza con le persone.	T
303	Quasi sempre vorrei essere morto.	T
305	Ho avuto certamente più motivi di preoccupazione di quanti me ne spettassero.	T
306	A nessuno importa molto di ciò che accade al prossimo.	T
325	Ho più difficoltà degli altri a concentrarmi.	T
331	Ho la tendenza a vedere le cose più difficili di quello che sono.	T
339	Qualche volta mi è sembrato che le difficoltà si accumulassero talmente, da non poterle superare.	T
357	Molto spesso mi tengo fuori dai pettegolezzi e dalle chiacchiere del gruppo a cui appartengo.	T
377	Non sono contento di me stesso, così come sono.	T
388	Molto raramente ho momenti di malinconia o tristezza.	F
399	Il futuro è troppo incerto perché una persona possa fare seri progetti per l'avvenire.	T
400	Spesso, anche se tutto mi va bene, sento che non mi importa di niente.	T
405	Di solito sono calmo e non mi agito facilmente.	F
408	Le delusioni mi colpiscono così intensamente da non potermele togliere dalla mente.	T
411	Talvolta penso di essere un buono a nulla.	T
415	Mi preoccupo abbastanza per le disgrazie che potrebbero capitare.	T
449	Qualcuno della mia famiglia ha un carattere irascibile.	T
454	Il futuro mi sembra senza speranza.	T
463	Parecchie volte alla settimana sento come se stessee per accadere qualche cosa di spaventoso.	T
464	Assai spesso mi sento stanco.	T
469	A volte mi sento come se stessi per andare in pezzi.	T
472	Sono molto preoccupato perché dimentico dove metto le cose.	T
496	Non mi sento sotto pressione o stress in questi giorni.	F
506	Recentemente ho pensato di suicidarmi.	T
509	Dover prendere decisioni importanti mi rende nervoso.	T
512	Nella mia vita ho avuto una perdita tragica e so che non riuscirò mai a superarla.	T
516	La mia vita è vuota e senza significato.	T
520	Ultimamente ho pensato molto al suicidio.	T
539	Ultimamente ho perso la voglia di risolvere i miei problemi.	T

Table 6. T-score of the MMPI-2 and their general meaning

Level	Uniform T-Score	Percentile Equivalent
Extremely High	85-90	>99.8->99.9
Very High	75-80	98->99
High	65-70	92-96
Moderately High	55-60	73-85
Average	45-50	34-55
Moderately Low	35-40	4-15
Very Low	30	<1

Conclusions

The aim of this paper was to examine the psychometric properties of the Mt, DEP and ANX scales in a sample of Italian university freshmen and their families, through modern test theory approach represented by Rasch analysis, and to analyze the correlation of these traits between students and parents. The analysis to these tests helped us to identify subsets of item for each scale that satisfy the Rasch model and therefore the fundamental axioms for objective measurement. The excluded items have been analyzed and reasons for exclusion discussed. Using a subset of items excluded from the DEP scale, a new scale, called SUI, has been identified that may represent a construct like “Attitude to Suicide”, whose (disattenuated) correlation with DEP is 0.76. The sample used here however lacked of clinical cases that could help to validate this scale with more precision: the person reliability was indeed as low as 0.70 and its separation was only 1.52, but the item reliability reached 0.97 with good separation indices, so that we may hope that applications to a more complete sample could bring good results for this new scale. For what it concerns the Mt, DEP and ANX scales, we applied to each two different Rasch models: the dichotomous Rasch model (1) and a Grouped Rating Scale model which combine the version (1) for some items expressed in dichotomy fashion, and the version (2) for group of items expressed with a likert scale 0=FALSE, 1=more FALSE than TRUE, 2=more TRUE than FALSE, 3=, in the case the MMPI-2 consider as 1 the item if FALSE; the reversed in the case the item would be considered 1 if TRUE. This second kind of model allows reducing the measurement errors reaching a lower level of standard deviation error for person measures and increasing the reliability of the measure that however was limited by the absence of clinical cases in the sample. The reduction of person measure standard errors is about 27% to 37% depending on the scale. All person reliability indices in the three scales are above 0.85 and 0.95 for items. Therefore, we may suggest that future application of Rasch analysis to MMPI-2 would use a likert scale to collect the data, at least for the items, from this analysis, seem to support it well. The analysis showed the presence of Differential Item Functioning for some of the items included in the scales, some for SEX and other for AGE: we take account of that in the final models specifying item separated for Male and Female persons and for Sons and Parents and we discussed the results for the different level of difficulty. For what it concerns this sample we found that Male persons are less maladjusted, less depressed and less anxious than Female persons are. Sons are more maladjusted, more depressed and more anxious than Parents are. Sons Mt measures are positively correlated with that of their mothers and negatively correlated with their fathers, the same is observed for DEP measures, but with lower level of correlation; ANX measures of sons are positively correlated both with mothers and with fathers. Mt is highly correlated (above 0.9) both with ANX and with DEP, a lower correlation (above 0.8) is observed between DEP and ANX. In general, we found good psychometric properties for the reduced set of items identified from the Rasch analysis that we hope might be extended to other samples and countries in order to verify the invariance property of the item measures.

References

1. Allen M.J., Yen W. M. (2002). *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press.
2. Andersen E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69-81.
3. Andrich D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
4. Barthlow D.L., Graham J.R., Ben-Porath Y.S., McNulty J.L. (2004). Construct validity of the MMPI-2 College Maladjustment (Mt) Scale, *Assessment*, 11(3):251-62.
5. Barthlow DL1, Graham JR, Ben-Porath YS, McNulty JL. (2004). Construct validity of the MMPI-2 College Maladjustment (Mt) Scale. [Http://www.ncbi.nlm.nih.gov/pubmed/15358881](http://www.ncbi.nlm.nih.gov/pubmed/15358881).
6. Battauz, M., Bellio, R., Gori, E. (2011). Covariate measurement error adjustment for multilevel models with application to educational data, *J. Educ. Behav. Stat.* 36, 283–306.
7. Bethune M. M. (2011). Predictors of Performance in a Professional Counselor Masters Program, Doctoral Thesis <http://gradworks.umi.com/34/93/3493607.html>.
8. Bond T. G., Fox M. C. (2007). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences. 2nd Edition*. Routledge, New York.
9. Bond T.G. (2003). Validity and assessment: a Rasch measurement perspective, *Metodología de las Ciencias del Comportamiento* 5(2), 179–194.
10. Borghans L., Duckworth A. L., Heckman J., Weel B. (2008). The Economics and Psychology of Personality Traits, *National Bureau of Economic Research, Working Paper* 13810 <http://www.nber.org/papers/w13810>.
11. Brentari E., Golia S. (2007). Unidimensionality in the Rasch model: how to detect and interpret, *STATISTICA*, anno LXVII, n. 3.
12. Butcher J. N., Williams C. L. (2009). Personality assessment with the mmpi-2: historical roots, international adaptations, and current challenges, *Applied Psychology: Health and Well-Being* 1 (1): 105–135.
13. Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *The Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.
14. Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology, *Professional Psychology: Research and Practice* 31 (2): 141–154.
15. Campbell N.R. (1920). *Physics: The Elements*. London: Cambridge University Press.
16. Carneiro P., Heckman J. J. (2003). Human Capital Policy, *National Bureau of Economic Research, Working Paper* 9495 <http://www.nber.org/papers/w9495>.
17. Cronbach, L. J.; Meehl, P.E. (1955). Construct Validity in Psychological Tests, *Psychological Bulletin* 52 (4): 281–302.
18. Cunha F., Heckman J. J., Lochner L., Masterov D. V. (2005). Interpreting the Evidence on Life Cycle Skill Formation, *National Bureau of Economic Research, Working Paper* 11331 <http://www.nber.org/papers/w11331>.
19. Farkas G. (2003). Cognitive Skills and Noncognitive Traits and Behaviors in Stratification Processes, *Annual Review of Sociology*, Vol. 29: 541-562.
20. Gori E. (2004). L'investimento in Capitale Umano attraverso l'Istruzione, in G. Vittadini (a cura di) (2004) *Capitale Umano. La ricchezza dell'Europa*. Guerini ed.
21. Gregory R. (2007). *Psychological Testing: History, Principles, and Applications - 5th ed*. Boston: Pearson. p. 392.
22. Guilfor J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.
23. Guttman L. (1950). The basis for scalogram analysis. In Stouffer et al. (Eds.), *Measurement and prediction*. New York: Wiley.
24. Heckman J. J., Dimitriy V. M. (2004). *Invest in Kids Working Group Committee for Economic Development*, Working Paper 5 http://jenni.uchicago.edu/Invest/FILES/dugger_2004-12-02_dvm.pdf.
25. Heckman J., Rubinstein Y. (2001). The Importance of Noncognitive Skills: Lessons from the GED Testing Program, *The American Economic Review*, Vol. 91, No. 2, *Papers and Proceedings of the Hundred Thirteenth Annual Meeting of the American Economic Association* (May, 2001), pp. 145-149.

26. Karabatsos G. (1999a). Rasch vs. two- and three-parameter logistic models from the perspective of conjoint measurement theory. Paper presented at the *Annual Meeting of the American Education Research Association*. Montreal, Canada.
27. Karabatsos G. (1999b). Axiomatic measurement theory as a basis for model selection in item-response theory. Paper presented at the *32th Annual Conference for the Society for Mathematical Psychology*. Santa Cruz, CA.
28. Karabatsos G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, 1, 2, 152-176.
29. Kleinmuntz, B. (1960). An extension of the construct validity of the Ego Strength scale. *Journal of Consulting Psychology*, 24, 463–464.
30. Kleinmuntz, B. (1961). The College Maladjustment Scale (Mt): Norms and predictive validity. *Educational and Psychological Measurement*, 21, 1029–1033.
31. Kuczka, T., Handal, P. J. (1990). Validity of the College Maladjustment scale for identification of distressed students. *Psychological Reports*, 67, 730.
32. Lauterbach D., Garcia M, Gloster A. (2002). Psychometric properties and predictive validity of the Mt Scale of the MMPI-2. <http://www.ncbi.nlm.nih.gov/pubmed/12462759>.
33. Linacre J. M. (2004). Rasch Model Estimation: Further Topics, *Journal of Applied Measurement*, 5(1),95-1 10, 101-.
34. Linacre J. M. (2009). Local Independence and Residual Covariance: A Study of Olympic Figure Skating Ratings, *Journal of Applied Measurement*, 10(2).
35. Luce R.D., Tukey J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, J, 1-27.
36. Marais I., Andrich D. (2008). Formalizing Dimension and Response Violations of Local Independence in the Unidimensional Rasch Model, *Journal of Applied Measurement* 9(3):200-15.
37. Masters G. (1982). A Rasch model for partial credit scoring. *Psychometrika* 1982, 47:149-174.
38. McNulty J. L. (2006). *Construct validity of the MMPI-2 College Maladjustment (Mt) Scale*. University of Tulsa.
39. Merker, Bradley M.; Smith, Janet V. (2001). Validity of the MMPI-2 College Maladjustment Scale, *Journal of College Counseling* , Vol. 4, No. 1 , Spring 2001
40. Messick S. (1995). Standards of validity and the validity of standards in performance assessment, *Educational Measurement: Issues and Practice*, 14 (4.): 5–8.
41. Perline R., Wright B.D., Wainer H. (1979). The Rasch model as additive conjoint measurement, *Applied Psychological Measurement*, 3, 237-256.
42. Popper K. (1934). *The Logic of Scientific Discovery*, (as *Logik der Forschung*, English translation 1959), ISBN 0-415-27844-9.
43. Poropat A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance, *Psychol Bull.*,135(2):322-38.
44. Rasch G. (1960). *Probabilistic models for some intelligence and attainment tests*, Chicago: University of Chicago Press.
45. Smith E.V. (2002). Detecting and evaluation the impact of multidimensionality using item fit statistics and principal component analysis of residuals, *Journal of Applied Measurement*, 3:205-231.
46. Stewart, D. W. (1994). Using the MMPI–2 College Maladjustment scale in a Canadian university setting, *Canadian Journal of Counseling*, 28, 135–141.
47. Svanum S., Ehrmann, L. C. (1993). Screening for maladjustment in college students: An application of receiver operating characteristics curve to MMPI scales, *Journal of Personality Assessment*, 60, 397–410.
48. Thurstone L.L. (1927). A law of comparative judgement, *Psychological Review*, 34, 273-286.
49. Wright B.D., Masters G.N. (1982). *Rating scale analysis*, Chicago: MESA Press.