TPM®

# THE CONSTRUCTION OF A SCALE TO MEASURE MATHEMATICAL ABILITY IN PSYCHOLOGY STUDENTS: AN APPLICATION OF THE RASCH MODEL

SILVIA GALLI
FRANCESCA CHIESI
CATERINA PRIMI
UNIVERSITY OF FIRENZE

The purpose of the present study was to develop a scale to measure the math ability that psychology students need to enrol introductory statistics courses inside their degree program. The Rasch model was applied to construct the instrument. The principal component analysis (PCA) of the residual showed a one-dimensional construct; the fit statistics revealed a good fit of each item to the model. The item difficulty measures were examined and the area of ability accurately assessed by the items was identified. The validity of the scale was assessed: the measures obtained by the scale correlated with attitude toward statistics and statistics anxiety (concurrent validity), and a relationship with statistics achievement was found (predictive validity).

Key words: Math basics for introductory statistics course; Mathematical ability scale; Rasch model.

*Correspondence concerning this article should be addressed to Silvia Galli, Dipartimento di Psicologia, Università degli Studi di Firenze, Via Cittadella 7, 50144 FIRENZE (FI), Italy. E-mail: silvia_galli@hotmail.it*

## INTRODUCTION

Psychology students are required to enrol statistics and quantitative research methodology courses inside their degree programs. The rationale for teaching methodology and data analysis is to enable students to handle, use, and interpret research or statistical data in their field of study. Many students encounter difficulties in these courses, and they typically experience a lower level of performance on statistics examinations than they do on all other examinations taken in their degree programs (Onwegbuzie, 2003). They often fail such examinations, which sometimes are an obstacle standing in the way of attaining the degree.

Researches have consistently reported a positive relationship between statistics course performance and basic mathematics skills (Harlow, Burkholder, & Morrow, 2002; Lalonde & Gardner, 1993; Schutz, Drogosz, White, & Distefano, 1999). In particular, Lalonde and Gardner (1993) found that previously acquired mathematical skills influenced introductory statistics course performance. Schutz et al. (1999) reported that mathematical ability was correlated with performance in statistics; Harlow et al. (2002), in an attempt to improve the performance of students enrolled in quantitative methods courses, showed a positive relationship between performance and mathematical skills.

Mathematical ability was also found to be related to variables, such as attitudes and anxiety toward statistics, which were deemed important by researchers addressing the learning of statistics (Gal, Ginsburg, & Schau, 1997; Onwuegbuzie, 2003; Sorge & Schau, 2002). Students of statistics courses often show negative attitudes and a high level of anxiety toward statistics. Negative attitudes have been found to correlate with previous negative experiences in mathematics (Sorge & Schau, 2002; Tremblay, Gardner, & Heipel, 2000); whereas a high level of mathematical skills are associated with positive attitudes toward statistics (Gal et al., 1997). Concerning anxiety, Onwengbuzie (2003) found that students with low mathematical ability frequently experienced a high level of anxiety toward statistics. Similarly, Zeidner (1991), as well as Onwuegbuzie and Seaman (1995), reported that the amount of prior exposure to mathematics and poor prior achievement in mathematics influenced the level of statistics anxiety.

Given the relationship between mathematical ability and statistics achievement, the aim of the present study was to develop a scale to measure the mathematical ability deemed necessary for psychology students to successfully complete introductory statistics courses. In previous studies two kinds of measures were used in order to assess the mathematical ability of students enrolled in statistics courses: scores derived from traditional tests generally developed with classical item analysis (Harlow et al., 2002; Schutz et al., 1999), and grades obtained in mathematics over high school years (Lalonde & Gardner, 1993). In our study, a scale was developed applying the Rasch Simple Logistic (RSL) model (Andrich, 1988), and the Rasch analytical procedures implemented in Winsteps software 3.59 (Linacre, 2005).

According to several researchers (Kline, 1996; Webster & Fischer, 2003), the RSL model offers advantages in educational measurement, in particular in mathematical skills assessment. In this model, the probability of a dichotomous response was modeled as a function of person ability and item difficulty. Specifically, the probability of a correct response was modeled as a logistic function of the difference between the person parameters and item parameters. So, the model was formalized through the following formula:

$$P_i(\theta)=\frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}},$$

where $P_i(\theta)$ is the probability for a subject with ability $\theta$ to respond correctly to an item *i*, while $b_i$ is the difficulty of the item *i*. The application of the model permitted the estimation of item difficulties and person abilities through an iterative process. Since the model is a mathematical expression of the theoretical relations that would hold between items and persons, neither of the two will perfectly fit the model. So we can compare the observed response with the Rasch model expectations through the fit diagnosis.

The RSL model was used in order to overcome the limitations of the classical approach (Hambleton & Jones, 1993; Hambleton, Swaminathan, & Rogers, 1991). The first limitation of the classical test theory (CTT) can be summarized as a situation of circular dependency: the item statistics (item difficulty and item discrimination) and the person statistics (observed test scores) are dependent. Namely, estimates of item difficulty and discrimination are dependent on the particular group of examinees completing the test, and the estimates of the person ability are dependent on the particular test item administered. The second major limitation concerns reliability, which is dependent on the number of the items composing the scale and on the sample. Moreover, the CTT assumes that the precision of the test is uniform across all levels of the person ability, so the same pool of items must be administered to every person.

The RSL model produces item statistics independent of examinee samples and person statistics independent of the particular set of items administered; concerning the measurement precision, this model gives an estimate of the test reliability that is sample independent and it is asserted not to be uniform across all levels of ability. Moreover, the item difficulty parameters and the person ability parameters are measured on the same scale, so their distribution can be compared.

In this study the RSL model was used with multiple choice items in accordance with some researches reporting that this model was appropriate for use with this type of items (Henning, 1989; Rentz & Rentz, 1979; Tinsley & Dawis, 1975). While some researchers disagreed on this point (Divgi, 1986; Goldstein & Blinkhorn, 1977), it was argued that the satisfactory fit of the RSL model to data collected using multiple choice items supported the conclusion that this model was also suitable when items were not really dichotomous, but the correct/incorrect dichotomy was obtained collapsing the options representing the wrong alternatives.

The specific purpose of the present research was to develop a scale to accurately measure low levels of ability in order to identify students with relevant difficulties in this domain, as well as the areas that were harder for them. Given the influence that mathematical ability has on statistics achievement, the information that might be obtained from the scale, such as identification of students with low level of ability and of their specific difficulties, could be useful to promote achievement and prevent failure. For instance, such students could be supported from the first day of the course with specifically-designed mathematics training courses.

The first step in the development of the scale was to identfy different areas on the basis of these courses curriculum. Afterwards, a pool of item was developed and evaluated at the qualitative level. Considering the scale developed, the unidimensionality (which is a fundamental criterion underlying the Rasch model), and item fit statistics were assessed. Moreover, the validity of the test was investigated. In particular the concurrent validity was investigated using correlations with measures of attitudes and anxiety toward statistics, on the basis of the relationship between mathematical ability and these constructs reported in a previous body of research. Finally, the predictive validity was investigated in order to assess the relationship between mathematical ability and statistics achievement through / by means of regression analysis.

## METHOD

### Participants

The participants were 788 psychology students of introductory statistics courses over a two-year period.[1] Ages ranged from 19 to 62, with a mean age of 21.18 (*SD* = 3.82) years. Most of the participants were women (81%). They came from a broad spectrum of high schools (49 % scientific studies, 33% humanistic studies, 18 % technical studies).

### Measure

The scale developed to measure math basics for introductory statistics courses was named Mathematical Prerequisites for Psychometrics (PMP: Prerequisiti di Matematica per la

Psicometria). On the basis of the contents of the introductory statistics course curriculum, which includes fundamental concepts of psychological research, descriptive statistics, and inferential statistics, six different mathematics domains were identified in order to collect information about the students' ability in arithmetic: Operations, Fractions, Set theory (inclusion-exclusion, and intersection concepts), first order Equations, Relations (less than, greater than, equal to relations among numbers ranging from 0 to 1, and numbers expressed in absolute values), and Probability (base-rates, independence notion, disjunction and conjunction rules). Fractions and Operations are employed both in descriptive and inferential statistics (e.g., to compute the standard deviation, as well as the $t$ or $z$ values). Equations are used, for instance, in the standardization procedure and in regression analysis. In order to test hypotheses (i.e., to compare the computed and the critical values to decide if the null hypothesis has to be accepted or rejected) it is necessary to establishing Relations among elements. Set theory principles help to understand probability rules. Probability issues are the prerequisite for hypothesis testing.

An initial pool of items was developed in order to operationalize these areas. Each item presented a multiple choice question (one correct choice among four alternatives). The items were evaluated at the qualitative level. Two math teachers evaluated the items content, in particular checking the wording to find out if they were suitable, intelligible, and unambiguous. The pool of items was then administered to a sample of psychology students ($N = 35$) in order to check response alternatives and replace with new alternatives those with a response lower than 3% . As a result of this analysis, some items were removed, others adapted, and some new ones were constructed.

The PMP scale included 30 items (see Appendix) equally allocated to the six areas previously defined (five items per area): *Fractions* (e.g., item 2), Operations (e.g., item 5), Set theory (e.g., item 3), Equations (e.g., item 12), Relations (e.g., item 13), and Probability (e.g., item 11).

In order to investigate the concurrent validity of the PMP scale, attitude and anxiety toward statistics were assessed. The Survey of Attitudes Toward Statistics (SATS) (Schau, Stevens, Dauphinee, & Del Vecchio, 1995) was administered to measure attitudes. It contains 28 Likert-type items rated on a 7-point scale ranging from *Strongly disagree* to *Strongly agree*. It was developed for students enrolled in introductory statistics courses, in two forms to be administered at the beginning of the course (pre-SATS) and at the end of it (post-SATS). The Statistical Anxiety Rating Scale (STARS) (Cruise, Casch, & Bolton, 1985) was used to measure statistics anxiety. It is a 51-item (5-point Likert format) instrument organized in two parts. The first part includes 23 items related to different aspects of statistic anxiety rated from *No anxiety* to *Very much anxiety*, and the second part includes 28 items related to the respondent's feeling toward statistics ranging from *Strong disagreement* to *Strong Agreement*. The Italian version of both the SATS and the STARS was obtained using a back-translation method. A translation from English into Italian was first conducted and this version was then back-translated into English by an English mother tongue teacher. The two English forms (the original,and the back-translated one) were compared to verify the similarity of the Italian and original version.

In order to investigate the predictive validity of the PMP scale, two statistics achievement measures were used: Final Examination Grade and Failure. The examination included a written task (three problem-solving questions and six open-ended conceptual questions) and an oral interrogation. The grade deriving from both the written and verbal assessment (range 0-30) allowed scoring both Final Grade (scores equal or greater than 18) and Failure (scores less than 18). The Failure variable was recoded as dichotomous: No Failure and Failure (one or more).

## Procedure

The PMP and the pre-SATS were administered at the beginning of the course during the first day of class. The post-SATS and the STARS were administered at the end of the classes. Each survey was presented briefly to the students and instructions for completion were given. Answers were collected in paper-and-pencil form and the time needed to complete them ranged from 15 to 30 minutes.

## Data Analysis

The unidimensionality of the construct, a fundamental criterion underlying the Rasch models, was assessed through an analytical procedure implemented in Winsteps software 3.59 (Linacre, 2005): the principal component analysis (PCA) of the standardized residuals. This is an unrotated principal components analysis. The standardized residuals for all observations were computed; then correlations matrix of standardized residuals across items was computed. In order to identify other components that may possibly affect response patterns, the correlation matrix was decomposed. If items had commonalities beyond those predicted by the Rasch model, then these might appear as shared fluctuations in their residuals. Additional components with eigen-values greater than three were considered as violation of unidimensionality (Linacre, 2005; Raiche, 2005; Smith & Miao, 1994).

After checking the unidimensionality assumption, the items fit statistics were calculated through Winsteps (Linacre, 2005) to test the fit between the items and the Rasch model. The estimation of fit began with the calculation of a response residual ($Y_{ni}$) for each person ($n$) when each item ($i$) was encountered, which indicated how far the actual response ($X_{ni}$) deviated from Rasch model expectations ($E_{ni}$):

$$Y_{ni} = X_{ni} - E_{ni}$$

Because there were too many residuals to examine in one matrix, the fit diagnosis was summarized in a fit statistic. In this study the fit diagnosis was assessed using the mean square infit statistic. The item infit statistic was a weighted average of the square standardized residuals ($Z_{ni}$) across item, in which relatively more impact was given to unexpected responses close to an item measure (more weight was given to the performances of persons closer to the item values). Residuals were weighted by their individual variance ($W_{ni}$) to lessen the impact of unexpected responses far from the measure:

$$\text{infit} = \frac{\sum z_{ni}^2 w_{ni}}{\sum w_{ni}}.$$

When infit values were reported as mean squares in the form of chi-square statistics divided by their degrees of freedom, their expected values was 1 and the suggested limits were between .7 and 1.3 (Bond & Fox, 2007; Linacre 2005; Wright, Linacre, Gustafson, & Martin-Loff, 1994).

The fit analysis was followed by the estimation of the item parameters and of the person parameters. The item difficulty and person ability measures were obtained employing two estimation methods implemented in Winsteps: Normal Approximation Estimation Algorithm (PROX) (Cohen, 1979) and Joint Maximum Likelihood Estimation (JMLE) (Wright & Panchapakesan, 1969). Initially, all parameter estimates were set to zero. The first phase of the estimation used the PROX method, which was employed to obtain rough estimates; each iteration through the data improved the PROX estimates until the increase in the range of the item measures was smaller than .5 logits. Then the PROX estimates became the starting values for JMLE and, again by iterating through the data, the final estimates were obtained. When the convergence criteria were satisfied, the iterative process ceased and the final estimation was obtained.

The reliability of the item difficulty estimates and of the person ability estimates was also examined.

Moreover, item difficulty and person measures were plotted along the same latent trait in order to compare the distribution of item difficulty and person ability. This allows identification of areas of ability that are accurately assessed by the items and identification of more difficult items for students with lower mathematical ability.

As last step the validity of the test was assessed. In particular, the concurrent validity was investigated using correlations between PMP, SATS, and STARS. The predictive validity was examined performing a linear regression analysis to asses the relationship between final examination grades and scores on PMP and exploring the difference on the PMP means scores between who passed and who failed the exam.

## Results

Concerning the unidimendionality analysis, the principal component analysis (PCA) of the standardized residuals revealed an additional factor with eigenvalue 1.9 accounting for the 3.3% of the unexplained variance, but lower than the criterion of 3 (Linacre, 2005; Raiche, 2005; Smith & Miao, 1994). The low percentage of the unexplained variance accounted by an additional factor showed that the items measured one dimension; it was therefore concluded that the basic mathematical competence was an unidimensional construct.

Given that this result allowed application of the Rasch model, checking the fit between the items and the model was the next step. The mean square infit statistic was assessed for each item, which showed mean square infit statistics within a .7 to 1.3 range. Item 16 (Probability) had the largest mean square infit (1.26). All other items had better mean square infit (between .83 and 1.13). Table 1 shows a summary of item infit. In examining the fit statistics misfitting items were not identified, so these statistic indices revealed that the empirical data met the model requirements.

TABLE 1
Item infit statistics

| Item | Infit | Item | Infit | Item | Infit |
|---|---|---|---|---|---|
| 1 Operations | 1.09 | 11 Probability | 1.07 | 21 Operations | 1.04 |
| 2 Fractions | 1.00 | 12 Equations | .90 | 22 Probability | 1.10 |
| 3 Set theory | .96 | 13 Relations | .96 | 23 Set theory | 1.05 |
| 4 Fractions | .84 | 14 Fractions | .92 | 24 Operations | 1.13 |
| 5 Operations | 1.06 | 15 Equations | .91 | 25 Probability | 1.11 |
| 6 Equations | 1.01 | 16 Probability | 1.26 | 26 Equations | .90 |
| 7 Fractions | .95 | 17 Equations | .87 | 27 Relations | .99 |
| 8 Fractions | .94 | 18 Relations | .98 | 28 Relations | .95 |
| 9 Set theory | .99 | 19 Probability | 1.07 | 29 Operations | .97 |
| 10 Relations | .86 | 20 Set theory | 1.08 | 30 Set theory | 1.22 |

Since the data fit the model chosen, it was possible to estimate the item parameters and the person parameters. The item difficulty measures covered a range between –1.72 and +2.16 logits. Table 2 shows a summary of item difficulty measures.

TABLE 2
Item difficulty measures along with standard errors

| Item | Difficulty | Item | Difficulty | Item | Difficulty |
|---|---|---|---|---|---|
| 1 Operations | 2.16 ± .09 | 13 Relations | 1.31 ± .08 | 3 Set theory | –.05 ± .09 |
| 21 Operations | 2.05 ± .08 | 27 Relations | 1.27 ± .06 | 20 Set theory | –.17 ± .09 |
| 24 Operations | 1.29 ± .06 | 10 Relations | –.56 ± .10 | 30 Set theory | –.26 ± .10 |
| 29 Operations | –.74 ± .11 | 18 Relations | –.59 ± .10 | 23 Set theory | –.82 ± .11 |
| 5 Operations | –.92 ± .11 | 28 Relations | –.62 ± .10 | 9 Set theory | –1.13 ± .06 |
| 25 Probability | 1.13 ± .06 | 26 Equations | .88 ± .08 | 8 Fractions | 1.02 ± .08 |
| 19 Probability | .46 ± .09 | 6 Equations | .40 ± .09 | 14 Fractions | –.21 ± .09 |
| 16 Probability | .32 ± .09 | 17 Equations | –.23 ± .10 | 2 Fractions | –.99 ± .11 |
| 11 Probability | .21 ± .09 | 12 Equations | –.43 ± .10 | 4 Fractions | –1.66 ± .14 |
| 22 Probability | –.85 ± .11 | 15 Equations | –.58 ± .10 | 7 Fractions | –1.72 ± .14 |

Three items of the Operations area (items 1, 21, and 24), one of the Probability area (item 25), and two of the Relations area (items 13 and 27) with a difficulty measure higher than one standard deviation above the mean difficulty (at 0 by default) were the most difficult items. Two items of the Fractions area (items 7 and 4) with a difficulty measure lower than one standard deviation below the mean difficulty were the easiest items.

The reliability of the item difficulty estimates was very high: .99 on a 0 to 1 scale (Wright & Masters, 1982). Item reliability refers to the ability of the test to define a distinction hierarchy of items along the measured variable. The higher the number, the more confidence we can place in the replicability of the item placement across other samples.

Concerning the person parameters the results showed that estimates on the ability scale ranged from –2.19 to 3.85 logits, with a mean of 1.30 (*SD* = 1.19). The person ability estimates mean higher than the mean of item estimates indicated that this sample found the test comparatively easy. The mean person estimates would be closer to 0 for a well-matched test.

The reliability of the person ability estimates was acceptable: .78 on a 0 to 1 scale (Wright & Masters, 1982). It estimated the replicability of person placement across other items measuring the same construct.

The item difficulty and person measures were plotted along the same latent trait in order to compare the distribution of item difficulty (location) and person ability. This allowed identification of areas of ability that were accurately assessed by the items. Figure 1 reports the distribution of both item location and person. Results showed that area of ability accurately assessed by items ranged from 1.30 (person ability estimates mean) to –1.08 (two standard deviations below person ability estimates mean).

So the PMP scale measured low levels of ability accurately, whereas it did not have the same accuracy in assessing high levels of ability (higher than 1.30) and extremely low levels of ability (lower than –1.08).

TPM Vol. 15, No. 1, 3-18
Spring 2008
© 2008 Cises



```
more>|<rare>
    4            .##  +
              .#####  |
                      |
                     T|
                      |
                      |
                      |
            .#######  |
    3                 +
                      |
                      |
        .##########   |
                     S|
                      |
            .#######  |
                      |  operation_1
    2                 +T operation_21
        .##########   |
                      |
       ############   |
                      |
        ###########   |
        .########## M|  relation_13        operation_24         relation_27
                      |  probability_25
    1         ######  +S fraction_8
            .#######  |  equation_26
                      |
             #######  |
             .######  |  probability_19
                      |  probability_16     equation_6
                 ###  |  probability_11
              .#### S|
    0           .###  +M set-theory_3
                      |  set-theory_20
                .###  |  fraction_14        equation_17         set-theory_30
               .####  |  equation_12
                  .#  |  relation_10
                  .#  |  equation_15        relation_18         relation_28
                      |  operation_29
                  .#  |  probability_22    set-theory_23        operation_5
   -1             .#  +S fraction_2
                     T|  set-theory_9
                  .#  |
                   .  |
                      |
                   .  |  fraction_4
                      |  fraction_7
                   .  |
   -2                 +T
                      |
                   .  |
                      |
                      |
                      |
                      |
                      |
   -3                 +
               <less>|<frequ>
```

FIGURE 1
This plot shows the distribution of persons and items. Items are shown on the right-hand side of the figure and Person measures (represented by the # symbol, each "#" represents six persons) on the left-hand side. The most difficult items and the most able persons are placed at the top. "M" markers represent the location of the mean; "S" markers are placed one standard deviation away from the mean; "T" markers are placed two standard deviations away.

Moreover, the distribution of item difficulty and person ability were compared to identify the most difficult items for students with lower mathematical ability. Students with a measure of ability lower than one standard deviation below sample average (1.30 ± 1.19) could not perform operations correctly that required understanding absolute value (item 1) and those including percentages (items 21 and 24). They seemed not able to answer items correctly regarding the relations when these presented absolute values and decimals among the response alternatives (items 27 and 13). Concerning the Probability area, they did not succeed in answering four items correctly: item 25 requiring the assessment of the probability of drawing two kings drawing two cards from a pack of cards, item 19 testing the ability of identifying how many combinations are possible rolling a dice and tossing a coin at the same time, item 16 referring to the lotto game asking which of the two chosen numbers has a greater chance of being selected, and item 11 asking to assess the probability of drawing one specific marble from a bag containing 10 marbles. Although students seemed able to correctly compute basic fractions, they could not answer an item correctly checking the ability to raise one fraction to the second power (item 8). In addition, they could not solve equations correctly with fractions (item 26) and those containing square roots among the response alternatives (item 6).

The last part of the research aimed to assess the validity of the PMP. The concurrent validity was investigated examining the relationship with SATS and STARS, so composite scores for these measures were calculated; high scores in the SATS (pre and post) indicated a more positive attitude toward statistics and high scores in the STARS indicated high anxiety experienced with statistics. A positive correlation between mathematical ability and attitude toward statistics measured at the beginning of the course was found ($r$ = .31, $p$ < .01), so students with high mathematical ability seemed to have a more positive attitude toward statistics. A positive relationship between mathematical ability and attitude measured at the end of the course was also found ($r$ =.41, $p$ < .01), so high scores in the PMP were associated to more positive attitude at the end of the course as well. Moreover, a negative correlation between mathematical ability and anxiety ($r$ = –.34, $p$ < .01) was found, indicating that students with high mathematical ability experienced lower anxiety.

In order to assess the predictive validity, the PMP scores of students who never failed (No Failure, NF) and students who failed the final examination one or more times (Failure, F) were compared. The significant difference ($t_{(305)}$ = 7.45, $p$ < .001) between the mean scores obtained by the two groups indicated that NF students showed higher mathematical ability ($M$ = 23.85, $SD$ = 4.49) than F students ($M$ = 19.21, $SD$ = 6.34). Moreover, a linear regression analysis revealed that mathematical ability was a significant predictor of Final Examination Grade (β = .32, $p$ < .01), accounting for 10% of the variance in statistics achievement.

## DISCUSSION

A scale to measure the mathematical ability deemed necessary for psychology students to successfully complete introductory statistics courses (PMP) was developed. The Rasch model was applied to construct the instrument.

On the basis of the contents of the statistics courses curriculum, six different mathematical areas have been identified and a pool of 30 items has been developed. The dimensionality analysis indicates that the PMP refers to a unidimensional construct. In particular, the principal component analysis used in the Rasch model reveals an additional factor accounting for a low percentage of unexplained variance; this leads to the conclusion that items measure only one dimension. It therefore follows that the identification of different areas does not influence the unidimensionality of the PMP scale, in accordance with many studies presenting unidimensional math tests composed by several areas similar to the ones here identified (Al-Hasan & Jaberg, 2003; Gierl, Bisanz, Bisanz, & Boughton, 2002; Neidorf, Binkley, Gattis, & Nohara, 2006).

The second part of the research intended to establish whether the data fit the model chosen (RSL model). In accordance with many studies reporting the satisfactory fit of the RSL model to data collected using multiple choice items (Hennings, 1989; Rentz & Rentz, 1979; Tinsley & Dawis, 1975), the present results show that the RSL model accurately explains the pattern of responses obtained by the PMP scale.

On the basis of the item difficulty measures, the Operations area, including three items with a high difficulty level, is the most difficult one. Whereas Fractions, including two items with low difficulty level, is the easiest area.

Comparing the distribution of item difficulty and person measures, the PMP assesses accurately low levels of mathematical ability, in line with the main purpose of the research which was to identify students with greater difficulties in this domain. An accurate assessment of low levels of ability is consistent with the aim to measure math basics necessary to pass the introductory statistics final examination.

The lower ability students' pattern of response shows that they encountered difficulties in all six areas; moreover these difficulties seem due to specific mathematical elements contained in the items. In particular, results show that these students did not succeed in answering items correctly that included absolute value, percentages, decimals, and square roots. It is interesting to underline that the exercises taken by students during introductory statistics courses frequently include these mathematical elements. As for the Probability area, students were not able to answer most of the items correctly (no one containing specific mathematical elements).

The relationship between mathematical ability and two correlated constructs such as attitude and anxiety toward statistics, supports the concurrent validity of the PMP (Onwuegbuzie, 2003; Sorge & Schau, 2002). A positive relationship with attitudes is found, along with a negative one with anxiety (i.e., students with higher mathematical ability have a more positive attitude toward statistics and experience lower anxiety). The analysis of the relationship between mathematical ability and achievement reveals the predictive validity of PMP; in accordance with previous studies (Harlow et al., 2002; Lalonde & Gardner, 1993; Schutz et al. 1999), students with high mathematical ability pass the final examination at the first attempt and obtain higher grades.

Considering the purpose of the present research, the last result is extremely important because it shows the advantages offered by using the PMP scale in introductory statistics teaching. Precisely, by administering the PMP at the beginning of the course, students who are most likely to fail the examination could be identified, and an ad hoc training course could be developed focusing on the mathematical contents required by the task. This kind of intervention could improve students' achievement helping them to obtain higher grades and to reduce the time necessary to pass the examination.

## NOTE

1. Four hundred and seven students were tested in the academic year 2004-2005, and 381 students in the academic year 2005-2006.

## REFERENCES

Al-Hasan, A. N., & Jaberg, P. (2003). Assessing the general education mathematics courses at a liberal arts college for women. *Supporting Assessment in Undergraduate Mathematics,* 59-63. Retrived from the Web January, 20, 2007. http://www.maa.org/saum/cases/al-hasen-jaberg1105-saum.pdf

Andrich, D. (1988). Rasch model for measurement. London: Sage.

Bond , T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology, 32*(1), 113-120.

Cruise, J. R., Cash, R. W., & Bolton, L. D. (1985). Development and validation of an instrument to measure statistical anxiety. *ASA Proceedings of the Section Statistical Education* (pp. 92-97). Alexandria, VA, EE. UU.: American Statistical Association.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice item? Not if you look closely. *Journal of Educational Measurement 23*(4). 283-289.

Gal, I., Ginsburg, L., & Schau, C. (1997). Monitoring attitudes and beliefs in statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 37-51). Amsterdan: IOS Press.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Buoghton, K. A. (2002, April). *Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), New Orleans, LA.

Goldstein, H., & Blinkhorn, S. (1977). Monitoring educational standards: An inappropriate model. *Bulletin of British Psychological Society, 30*, 309-311

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newberry Park, CA: Sage

Harlow, L. L., Burkholder, G. J., & Morrow, J. A. (2002). Evaluating attitudes, skill and performance in a learning-enhanced quantitative methods course: A structural modelling approach. *Structural Equation Modeling*, *9*, 413-430.

Henning, D. (1989). Does the Rasch model really work for multiple choice item? *Journal of Educational Measurement, 26*(1), 91-97.

Kline, P. (1996). *Manuale di psicometria* [*Psychometrics manual*]. Roma: Casa Editrice Astrolabio.

Lalonde, R. N., & Gardner, R. C. (1993). Statistics as a second language? A model for predicting performance in psychology students. *Canadian Journal of Behavioural Science*, *25*, 108-125.

Linacre, J. M. (2005). Winsteps (Version 3.59) [Computer Software]. Chicago: Winsteps.com.

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing mathematics content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessment* (NCES 2006 – 029). Department of Education, National Center of Education Statistics. Washington, DC: Government Printing Office.

Onwuegbuzie, A. (2003). Modeling statistics achievement among graduate students. *Educational and Psychological Measurement, 63*(6), 1020-1038.

Onwuegbuzie, A., & Seaman M. A. (1995). The effect of time constraints and statistics test anxiety on test performance in a statistics course. *Journal of Experimental Education, 63*(2), 115-124.

Raiche, G. (2005). Critical eigenvalue size in standardized residual Principal Component Analysis. *Rasch Measurement Transactions*, *19*(1), 1012.

Rentz, R. R., & Rentz, C. C. (1979). Does the Rasch model really work? *Measurement in Education*, *10*(2), 1-8.

Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective Measurement: Theory into Practice* (Vol. 2). Greenwich: Ablex.

Schau, C., Stevens, J. J., Dauphinee, T. L., & Del Vecchio, A. (1995). The developmental validation of the Survey of Attitudes Toward Statistics. *Educational and Psychological Measurement*, *55*, 868-875.

Schutz, P. A., Drogosz, L. M., White, V. E., & Distefano, C. (1999). Prior knowledge, attitude and strategy use in an introduction to statistics course. *Learning and Individual Differences*, *10*, 291-308.

Sorge, C., & Schau, C. (2002, April). *Impact of engineering students' attitudes on achievement in statistics: A structural model.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Tinsley, H. E., & Dawis, R. V. (1975). An investigation of Rasch simple logistic model: Sample-free item and test calibration. *Educational and Psychological Measurement, 35,* 325-339.

Tremblay, P. F., Gardner, R. C., & Heipel, G. (2000). A model of the relationships among measures of affect, aptitude and performance in introductory statistics. *Canadian Journal of Behavioural Science, 32*, 40-48.

Webster, B. J., & Fisher, D. L. (2003). School-level environment and student outcomes in mathematics. *Learning Environments Research*, *6*, 309-326.

Wright, B. D., Linacre, J. M., Gustafsson, J-E., & Martin-Loff, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

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

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free items analysis. *Educational and Psychological Measurement, 29*, 23 - 48.

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology*, *61*, 319-328.

## APPENDIX

Mathematical Prerequisites for Psychometrics
(PMP: Prerequisiti di Matematica per la Psicometria)

Read the following problems. Each problem presents four response alternatives (only one is correct). Indicate the correct response by ticking (X) the appropriate box.

---

**1.** Which of the following is equal to $\sqrt{x^2}$ ?

□ $x$ □ $\frac{x}{2}$ □ $2x$ □ $\pm x$

---

**2.** Which is the result of $\frac{2}{5} \div \frac{3}{2}$?

□ $\frac{5}{10}$ □ $\frac{4}{15}$ □ $\frac{5}{7}$ □ $\frac{3}{5}$

---

**3.** If set A includes numbers from 0 to 10, whereas set B includes even numbers lower than 18, which of the following relations is true?

□ B is included in A □ A is included in B
□ A and B do not share elements □ None of the above

---

**4.** Which is the result of $\frac{2}{3} + \frac{3}{4}$?

□ $\frac{5}{7}$ □ $\frac{6}{12}$ □ $\frac{17}{12}$ □ $\frac{5}{12}$

---

**5.** Which is the result of (–8)÷(–2)?

□ 4 □ –4 □ 16 □ 0.25

---

**6.** Which of the following equations is not possible?

□ $-(\sqrt{x}) = -3$ □ $\sqrt{x} = 3$
□ $\sqrt{-|x|} = 3$ □ None of the above

---

**7.** Which is the result of $\frac{2}{5} \times \frac{3}{2}$?

□ $\frac{5}{7}$ □ $\frac{3}{5}$ □ $\frac{5}{10}$ □ $\frac{6}{7}$

---

(appendix continues)

Appendix (continued)

**8.** The double of $\frac{3}{4}$ is:

□ $\frac{6}{8}$ □ $\frac{3}{2}$ □ $\frac{9}{16}$ □ $\frac{3}{8}$

**9.** Set A consists of all odd numbers between 8 and 20, whereas set B consists of all numbers less than 10. How many elements do the two sets share?

□ None □ 1 □ 2 □ 3

**10.** Which of the following relations is true?

□ $\frac{1}{2}\times\frac{1}{3}>\frac{1}{2}$ □ $\frac{1}{2}\times\frac{1}{3}>1$
□ $\frac{1}{2}\times\frac{1}{3}<\frac{1}{2}$ □ $\frac{1}{2}\times\frac{1}{3}<0$

**11.** Place five red marbles, three green marbles and two yellow marbles in a bag. Draw one red marble out of the bag. Without replacing the marble back into the bag, what is the probability of drawing one green marble in a second drawing?

□ $\frac{3}{10}$ □ $\frac{3}{9}$ □ $\frac{4}{10}$ □ $\frac{4}{9}$

**12.** Which is the result of the following equation: $(5+3)x = 0$?

□ $x = 5 - 3$ □ $x = 0$ □ $x = \frac{5}{3}$ □ $x = 5 + 3$

**13.** Which of the following relations is true?

□ $0.01 \times 0.01 < 0.01$ □ $0.01 \times 0.01 > 0.01$
□ $0.01 \times 0.01 = 0.01$ □ $0.01 \times 0.01 = 0.1$

**14.** The fraction $\frac{3}{7}$ is within:

□ 0 and 1 □ 1 and 2 □ –1 and 0 □ 2 and 3

**15.** Considering the following equation: $3x + 27 = 18$, which is the value of $x$?

□ 3 □ 15 □ –9 □ –3

(appendix continues)

Appendix (continued)

**16.** If I choose numbers 13 and 17 in a lotto game, which of the following numbers has a higher probability of being drawn?

- ☐ 13 or 17
- ☐ 13 and 17
- ☐ 13
- ☐ 17

**17.** Knowing that $xy = 3$, which of the following is true?

- ☐ $y = \frac{3}{x}$
- ☐ $y = 3 - x$
- ☐ $y = 3x$
- ☐ $\frac{xy}{3} = 0$

**18.** The value 0.05 is:

- ☐ lower than 0
- ☐ higher than 0.1
- ☐ within –1 and 0
- ☐ within 0 and 1

**19.** Rolling a dice and tossing a coin at the same time, how many combinations are possible?

- ☐ $6 + 2$
- ☐ $6 \times 2$
- ☐ $6 + 6$
- ☐ $6 \times 6$

**20.** If set A is composed by the letters A M A and set B by the letters A M A R E, which of the following relations is true?

- ☐ A and B are coincident
- ☐ A is included in B
- ☐ B is included in A
- ☐ A and B share elements

**21.** In a school there are 125 students; the students who passed an examination are 116. The percentage of the students that failed is:

- ☐ 6.2%
- ☐ 9%
- ☐ 7.8%
- ☐ 7.2%

**22.** What is the probability of drawing one ace from a pack of 40 cards?

- ☐ $\frac{1}{40}$
- ☐ $\frac{1}{4}$
- ☐ $\frac{4}{10}$
- ☐ $\frac{4}{40}$

**23.** If set A includes the first 10 letters of the alphabet and set B includes the vowels, which of the following relations is true?

- ☐ A and B are coincident
- ☐ A is included in B
- ☐ B is included in A
- ☐ A and B share elements

**24.** One hundred and forty students took a test. Seventy per cent of them passed the test. Which of the following is true?

- ☐ 98 students did not pass the test
- ☐ Students who did not pass the test were half of those who passed it
- ☐ 42 students did not pass the test
- ☐ None of the above

(appendix continues)

Appendix (continued)

**25.** Drawing two cards from a pack of 40, the probability if drawing two kings is:

- ◻ Equal to the probability of drawing one king
- ◻ Lower than the probability of drawing one king
- ◻ Higher than the probability of drawing one king
- ◻ None of the above

**26.** If $y = \dfrac{x-10}{2}$, which of the following is true?

- ◻ $x = 5y$
- ◻ $x = 2y + 10$
- ◻ $x = 2y - 10$
- ◻ $x = y + 5$

**27.** Which of the following relations is true?

- ◻ $|-2.1| < |-8.3|$
- ◻ $|2.1| > |-8.3|$
- ◻ $|-2.1| > |-8.3|$
- ◻ $|2.1| > |8.3|$

**28.** The value –0.98 is:

- ◻ within –1 and 0
- ◻ higher than 0
- ◻ lower than –1
- ◻ within –2 and –1

**29.** Which is the result of the following operation $[(13 - 10)^2 + (17 - 20)^2 + (10 - 10)^2]$?

- ◻ 0
- ◻ 18
- ◻ 6
- ◻ 9

**30.** If A is composed by the letters P O R T A, whereas B by the letters P A R T O, which of the following relations is true?

- ◻ A and B are coincident
- ◻ A is included in B
- ◻ B is included in A
- ◻ A and B share elements