ASSESSING MATHEMATICS COMPETENCE IN INTRODUCTORY STATISTICS COURSES: AN APPLICATION OF THE ITEM RESPONSE THEORY

Silvia Galli, Francesca Chiesi and Caterina Primi
Department of Psychology, University of Florence, Italy
silvia_galli@hotmail.it

A positive relationship between statistics achievement and mathematics competence has been consistently reported in literature. Given the influence that mathematical ability has on statistics achievement, it turns out to be useful to assess the mathematical competence in order to promote statistics course performance. In this work, we developed a scale to measure the mathematical ability deemed necessary for psychology students to successfully complete an introductory statistics course. The Item Response Theory was applied to construct the instrument in order to overcome the limitations of the classical approach. The predictive validity of the instrument was assessed considering achievement as the criterion measure. Advantages offered by using the scale in introductory statistics teaching are discussed.

INTRODUCTION

Many psychology students encounter difficulties in statistics and quantitative research methodology courses inside their degree programs. Several factors have been investigated in order to better understand the underlying mechanism of statistics achievement. Among them, basic mathematics abilities (i.e. basic computational skills and algebra) appeared to be a fundamental requirement to succeed in statistics. Specifically, researchers have consistently reported a significant and positive relationship between statistics achievement and mathematical competence (Harlow, Burkholder, & Morrow, 2002; Nasser, 2004; Onwuegbuzie, 2003; Schutz, Drogosz, White & Distefano, 1999).

Given the relationship between statistics achievement and mathematical ability, it is generally agreed that assessing mathematics competence can be useful to promote statistics course performance. Specifically, the information that can be obtained, such as identification of students with low level of ability and of their specific difficulties, could be useful to prevent failures. For instance, such students could be supported from the first day of the course with specifically-designed mathematics training courses.

Two kinds of measures have been used in order to assess mathematical ability of students enrolled in statistics courses: self-report measures of mathematics background, i.e. grades in mathematics over high school years (Nasser, 2004; Onwuegbuzie, 2003), and scores derived from tests developed for the purpose of the research whose psychometric properties were not investigated (Harlow et al., 2002; Schutz et al., 1999).

The aim of the present study was to develop a scale to measure the mathematical ability deemed necessary for psychological students to successfully complete introductory statistics courses. In our study a scale was developed using item response theory (IRT) in order to overcome the limitations of the Classical Test Theory (CTT) (Hambleton & Jones, 1993). IRT differs from CTT mainly because IRT models produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered. In order to investigate the predictive validity of the developed scale achievement was used as criterion measure.

METHOD

The sample of this study consisted of 600 psychological students (84% females) of the University of Florence. Age ranged from 19 to 58, with a mean age of 21.19 (*SD* = 4.01) years. They enrolled in introductory statistics courses over a two-years period (2007 and 2008).

The scale developed to measure mathematics 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 introductory statistics course curriculum, which includes fundamental concepts of descriptive statistics and inferential statistics, two teachers of introductory statistics identified the mathematical prerequisites (e.g., in order to investigate the

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

relationship between variables it is necessary to solve first order equations; in order to calculate probability it is necessary to solve operations with fractions). An initial pool of 48 items was developed in order to operationalize these prerequisites. Each item presented a multiple choice question (one correct among four alternatives). Two teachers of mathematics analyzed the items contents; as a result of this analysis, some items were removed, others adapted, and some new ones were constructed. The final version of the scale included 30 items (e.g. "*Which is the result of the following equation: (5+3)x= 0?*"; "*Which is the result of 2/5 × 3/2?*").

Achievement measure was the Final Examination Grade. The examination included a written task (three problem solving questions and six open-ended conceptual questions) and an oral interrogation. The final grade derived both from the written and verbal assessment (range 18 - 30).

The PMP was administered at the beginning of the course during the second day of the class. The final grade was registered in the academic years 2007 and 2008.

RESULTS

The unidimensionality of the construct, a fundamental criterion underlying the IRT models, was assessed through a Confirmative Factor Analysis (CFA). Cause of variables analyzed were dichotomous (the correct and incorrect dichotomy was obtained collapsing the options representing the wrong alternatives) a non linear factor analysis on thetracoric correlation matrix was performed using the WLSMV (*Weighted Least Square with Mean and Variance Adjustement*) estimator. The factor analysis confirmed the hypothesis that the item measured one dimension. Specifically, the chi square/df ratio was 1.85; moreover the CFI and the TLI were respectively .92 and .94 and the RMSEA was .03.

Given that these results allowed application of the IRT models, choosing the model was the next step. For dichotomous items, the 1, 2, and 3 parameter logistic models are available (1PL, 2PL, 3PL). In the 1PL model the probability of a dichotomous response is modeled as a function of person ability and item difficulty, in the 2PL model as a function of person ability, item difficulty and item discrimination, and in the 3PL model as a function also of item guessing (respondents' probability of getting a question correct simply by chance). Since the utility of the IRT model is dependent upon the extent to which the model accurately reflects the data, the IRT model-data fit was investigated. Specifically, the differences of -2loglikelihood for nested models were assessed using Multilog software (Thissen, 1991). Cause of a significant result implies that the model with more parameters provides a superior fit to the data, results showed that the 2PL model is the most suitable model to analyze the PMP (Table 1) indicating that observed items responses are affected by item difficulty and discrimination, but not by guessing.

Table 1. Comparison of -2loglikelihood

Model	***Δ-2loglikelihood***	***df***	***p***
1PL - 2PL	196.2	29	.001
1PL - 3PL	234.3	59	.001
2PL - 3PL	38.7	30	n.s.

After the choice of the model, the item fit statistics (S-Q^2) were calculated through Goodfit software (Orlando, 1997) in order to test the fit between the items and 2PL model. Results showed that each item had not significant S-Q^2 values indicating that all items fit the 2PL model.

The next step was to apply the 2PL model to estimate items parameters. The item difficulty and discrimination were obtained by employing the *Marginal Maximum Likelihood* (MML) estimation with the EM algorithm (Bock & Aitkin, 1981) implemented in Multilog software (Thissen, 1991). The IRT models use the logits unit to express the parameters. The item difficulty measures cover a range between -3.90 ± 1.27 and +1.00 ± .28 logits (Table 2). The most difficult items (item 1 e 21) have a difficulty measure higher than mean ability (it was located at 0 by default). The other ones have a difficult measure lower than mean ability.

Table 2. Item difficulty (b) measures along with standard errors

Item	*b*	*Item*	*b*	*Item*	*b*
1	1.00 ± .28	19	-1.18 ± .20	4	-1.75 ± .14
21	.15 ± .12	17	1.28 ± .14	7	-1.84 ± .19
24	-.15 ± .15	12	-1.30 ± .13	29	-1.94 ± .33
27	-.17 ± .08	16	-1.46 ± .34	30	-2.09 ± .31
13	-.24 ± .08	3	-1.49 ± .25	5	-2.10 ± .31
25	-.47 ± .15	10	-1.54 ± .17	9	-2.53 ± .51
8	-.47 ± .09	11	-1.55 ± .31	22	-2.74 ± .61
26	-.53 ± .11	18	-1.57 ± .22	20	-3.13 ± .91
6	-.92 ± .14	15	-1.61 ± .20	2	-3.87 ± 1.13
14	-1.14 ± .14	28	-1.70 ± .25	23	-3.90 ± 1.27

The range of item discrimination is between .57 ± .13 and 3.41 ± .79 (Table 3). By and large items have a high discriminative power (measure higher than 1.35) or a moderate discriminative power (measure higher than .65) (Baker, 2001). Three items (item 20, 23 e 21) with a measure lower that .65 were items with a low discriminative power (Baker, 2001).

Table 3. Item discrimination (a) measures along with standard errors

Item	*a*	*Item*	*a*	*Item*	*a*
4	3.41 ± .79	26	1.30 ± .17	25	.91 ± .15
7	2.47 ± .51	28	1.30 ± .26	24	.83 ± .14
12	1.87 ± .27	18	1.25 ± .21	22	.80 ± .20
10	1.84 ± 3.1	6	1.21 ± .18	11	.78 ± .16
27	1.83 ± .21	29	1.11 ± .23	30	.73 ± .17
17	1.64 ± .25	5	1.09 ± .23	2	.67 ± .21
13	1.61 ± .19	21	1.07 ± .15	16	.65 ± .15
15	1.60 ± .20	19	1.00 ± .16	20	.60 ± .21
8	1.49 ± .19	3	.98 ± .18	23	.58 ± .17
14	1.47 ± .22	9	.94 ± .23	1	.57 ± .13

Moreover in order to identify the area of ability that is accurately assessed by the PMP the *Test Information Function* (TIF) was analyzed. The TIF is used in the IRT as the measure of accuracy in ability estimations; this is equal to the inverse of the standard error, higher value indicate accurate ability estimates. Results showed that PMP scale measures accurately a low level of ability; specifically the area of ability accurately assessed by items ranges from .00 logits (mean ability) to–2.60 logits (measure of ability lower than two standard deviation below average) (Figure 1).

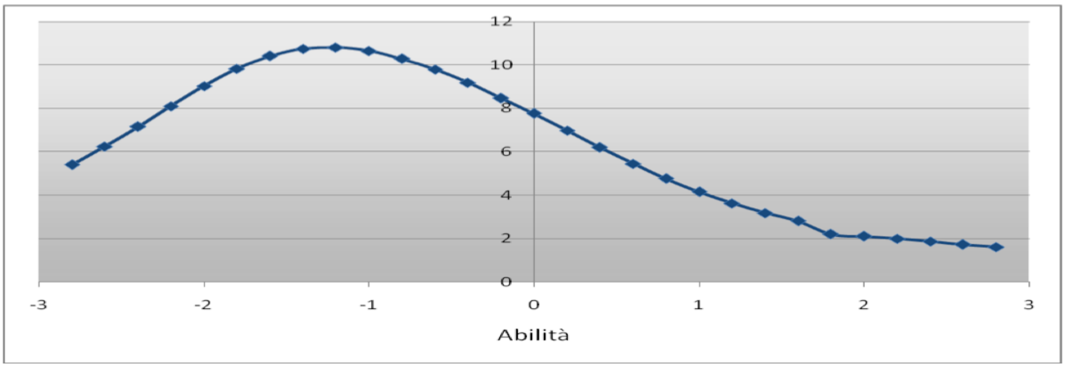

Figure 1. Test Information Function (TIF)

The last part of the research aimed to assess the predictive validity of the PMP. A linear regression analysis revealed that mathematical ability is a significant predictor of Final

Examination Grade (β = .32, *p* < .001), accounting for the 10% of the variance in statistics achievement.

DISCUSSION AND CONCLUSION

A scale to measure mathematical ability deemed necessary for psychology students to successfully complete introductory statistics courses (PMP) was developed using Item Response Theory (IRT). On the basis of the contents of statistics courses curriculum, mathematical prerequisites have been identified and a pool of 30 items has been developed. The dimensionality analysis indicates that the PMP refers a unidimensional construct. In accordance with many studies reporting the satisfactory fit to the 2PL model to data collected using multiple choice item (Hambleton, 1994), the present results showed that the 2PL model accurately explains the pattern of response obtained by PMP scale.

The item difficult measures showed that items contents are easy for the students. Nonetheless, the discriminative measures showed that items can discriminate students with different mathematical ability levels. In particular, the scale assesses accurately low levels of mathematical ability in accordance with the aim of the research which was to assess mathematical abilities necessary to pass introductory statistics exams, and to identify students with low level of ability with the aim to support them.

The analysis of relationship between mathematical ability and achievement attests the predictive validity of PMP. Considering the purpose of the present research, this result is extremely important because it shows the advantage offered by using the scale in introductory statistics teaching. Precisely, administering the scale at the beginning of the course it allows to obtain a measure of mathematical ability that can predict the performance in the final examination. So students who are most likely to fail the examination could be identified, and ad hoc training courses could be developed focusing on mathematical contents required by the task.

REFERENCES

Baker, F. B. (2001). *The Basic of Item Response Theory*. College Park: Erik.

Bock, R. D., & Aitkin M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters. Application of an EM Algorithm. *Psychometrika*, *46*, 443-459.

Hambleton, R. K. (1994). Item Response Theory: a broad psychometric framework for measurement advance. *Psicothema*, *6*(3), 535-556.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory*. Newberry Park: 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 modeling approach. *Structural Equation Modeling*, *9*, 413-430.

Nasser, F. (2004). Structural Model of the Effects of Cognitive and Affective Factors on the Achievement of Arabic-Speaking Pre-service Teachers in Introductory Statistics. *Journal of Statistics Education*, *12*(1). Online: www.amstat.org/publications/jse/v12n1/nasser.html]

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

Orlando, M. (1997). Item fit in the context of item response theory (Doctoral dissertation, University of North Carolina. *Dissertation Abstracts International*, *58/04-B*, 2175.

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

Thissen, D. (1991). *Multilog User's Guide*. Chicago: SSI.