MEASURING THE BASICS OF PROBABILISTIC REASONING: THE IRT-BASED CONSTRUCTION OF THE PROBABILISTIC REASONING QUESTIONNAIRE

Caterina Primi[1], Kinga Morsanyi[2], & Francesca Chiesi[1]
[1]NEUROFARBA – Section of Psychology, University of Florence, Italy
[2]School of Psychology Queen's University Belfast, UK
caterina.primi@unifi.it

Some of the difficulties students have in learning basic data analysis stem from a lack of basic understanding of probabilities. The aim of the present study was to develop a scale to measure basic probabilistic reasoning skills, which are deemed necessary to successfully complete introductory statistics courses. Specifically, our aim was to accurately measure low levels of ability in order to identify students with difficulties. Item Response Theory was applied to construct a scale. The validity of the scale was studied with reference to math skills and self-efficacy, attitude towards statistics, and numeracy. Finally, the equivalence of the scale across genders was assessed by analyzing Differential Item Functioning. The scale has advantages in introductory statistics teaching. Indeed, once students who are more likely to encounter difficulties are identified, ad-hoc training courses can be developed, focusing on probability knowledge that is required at the exam, in order to improve students' performance.

INTRODUCTION

The ability to think statistically about uncertain outcomes, and to make decisions on the basis of probabilistic information is important in many fields (e.g., for businesspeople, physicians, politicians, lawyers), and an inability to make optimal choices can be extremely costly, not only at the individual level, but also for society in general. Thus, statistics has been introduced as part of a wide range of curriculum programs in many countries. However, the discipline is viewed as a difficult and unpleasant topic. At the university level, students often encounter difficulties, experience stress and anxiety, and, eventually, many of them fail to pass the exams.

The main challenge in teaching and learning statistics is to ensure that students not only acquire the mechanics of statistical methods but also the concepts that underlie statistical reasoning (Garfield & Ben-Zvi, 2008). Indeed, even students who demonstrate their mastery of these techniques may have little understanding of statistical reasoning (Garfield, delMas, & Chance, 2007). To attain this goal it is essential to identify variables that attenuate or accentuate statistical reasoning and determine the nature of barriers faced by students. Konold and Kazak (2008) suggested that some of the difficulties students have in learning basic data analysis stem from a lack of basic understanding of probabilities.

In line with this claim, the aim of the present study was to develop a scale to measure the basics of probabilistic reasoning deemed necessary to successfully complete introductory statistics courses. In particular, the purpose of this research was to develop a scale to accurately measure low levels of probabilistic reasoning ability. The information obtained from the scale, that is the identification of students with difficulties in this domain, should be useful to improve achievement and prevent failure, as students with difficulties could be supported from the first day of the course with specific training activities.

The scale was developed using Item Response Theory (IRT), overcoming some of the limitations of Classical Test Theory (CTT). The IRT analytical methods and procedures offer a different value of test precision for every specific level of the underlying latent variable that is being measured, and it does not assume that a single estimate of reliability is sufficient to describe the precision of measurement over all levels of ability. Moreover, IRT parameters are invariant with respect to the sample characteristics from which they are generated. IRT methods can quantify the information value of both individual items and the overall test, and this information can be evaluated at any level of the latent trait. In other words, IRT is appropriate for developing instruments aimed at accurately measuring a specific level of the assessed ability (Embreston & Reise, 2000).

IRT measurement modeling can bear important implications for the assessment of probabilistic reasoning. Our aim was to develop a scale that makes it possible to measure low

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

levels of probabilistic reasoning skills (latent trait) with higher precision. It means that the scale has to be helpful in the accurate identification of respondents with weak probabilistic reasoning ability. In doing this, IRT is especially useful given that, assuming that measurement precision is not constant across the entire trait range, this method allows to choose items with good properties (discriminative and difficulty) for each level of the latent trait.

In order to develop the scale, as a first step, a definition of probabilistic reasoning ability was agreed on, and a pool of items was created. The items were tested in pilot studies at both qualitative and quantitative levels. As a result of these studies, we obtained a version composed of 16 items (*Probabilistic Reasoning Questionnaire*, PRQ; Primi et al, submitted). As a next step, the items were calibrated, and the test information function, (i.e., the reliability of the scale for different levels of ability) was investigated. Afterwards, the measurement equivalence of the scale across genders was tested. Indeed, conventional methods would also be problematic in addressing the issue of whether group (e.g., gender) differences in mean levels of probabilistic reasoning skills reflect true differences between groups or, rather, they arise because the items combined in aggregated scores have different measurement properties in different groups. Within the IRT framework, the measurement equivalence of the scale was assessed analyzing Differential Item Functioning (DIF). An item is considered to exhibit DIF, if respondents of two different groups who have equal levels of the ability that is being measured do not have the same probability of responding to the item correctly. Finally, the validity of the scale was studied. Relying on previous research on statistics achievement (for a review see Zieffler et al., 2008), we expected that probabilistic reasoning would be positively related to math skills, math self-efficacy, attitudes toward statistics, and numeracy.

METHODS

Participants

Participants were 1032 high school (42%) and university (58%) students (38% male, M = 19.31, SD = 4.08).

Measures

Probabilistic Reasoning Questionnaire (PRQ). The scale was composed of 16 multiple-choice probabilistic reasoning tasks (one correct out of three alternatives). Items included simple, conditional and conjunct probabilities, and data were presented both in frequencies and percentages (for examples: “A ball was drawn from a bag containing 10 red, 30 white, 20 blue, and 15 yellow balls. What is the probability that it is neither red nor blue? a. 30/75; b. 10/75; c. 45/75; and “60% of the population in a city are men and 40% are women. 50% of the men and 30% of the women smoke. We select a person from the city at random. What is the probability that this person is a smoker? a. 42%, b. 50%, c. 85%).

Mathematics Prerequisites for Psychometrics (PMP, Galli, Chiesi & Primi, 2011). This test was developed with the aim of measuring the mathematics skills needed by students enrolling in introductory statistics courses. The test consists of 30 problems, and it has a multiple-choice format (one correct out of four alternatives).

Solution of Math Problems of the *Mathematics Self-Efficacy Scale revised* (MSES-R; Kranzler & Pajares, 1997). The scale measures students’ self-reported level of confidence in successfully solving mathematics problems. The scale consists of specific mathematics problems for which students are asked to rate their confidence that they are able to solve them successfully, using a five-point scale ranging from 1 (no confidence) to 5 (complete confidence).

Numeracy Scale (Lipkus, Samsa & Rimer, 2001). The scale contains 11 items that assess basic probability and mathematical concepts, including simple mathematical operations on risk magnitudes, using percentages and proportions.

RESULTS

Calibration

The unidimensionality of the construct, a fundamental criterion underlying IRT models, was assessed through a Confirmatory Factor Analysis (CFA). Results confirmed that the items measured one dimension. Specifically, the chi-square/df ratio was 2.33, the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) were .94 and .95, respectively, and the Root Mean Square Error of Approximation (RMSEA) was .03. Factor loadings were all significant (p<.001) ranging from .40 to .80. Given that these results allowed for the application of IRT models, choosing the best model was the next step. Results showed that the 2PL model was the most suitable model to analyze the scale. This means that the observed item responses were affected by item difficulty and discrimination, but not by guessing. After selecting the model, the item fit statistics were calculated in order to test the fit between the items and the 2PL model. Results showed that each item fitted the 2PL model.

The next step was to apply the 2PL model to estimate item parameters. Item difficulty and discrimination were assessed by employing the Marginal Maximum Likelihood (MML) estimation with the EM algorithm implemented in the IRTPRO software (Cai, Thissen, & du Toit, 2011). The item difficulty measures covered the low range of the trait and, by and large, the items had a high discriminative power. Moreover, in order to identify the level of ability that is accurately assessed by the scale the Test Information Function (TIF) was analyzed. Results showed that the scale accurately measured low levels of ability (ability levels from three standard deviations below average to average; see Figure 1).

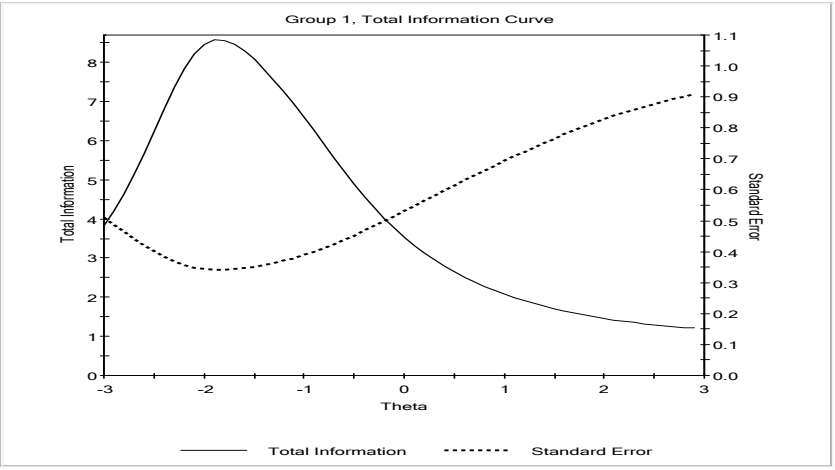

Figure 1. Test Information Function of the *Probabilistic Reasoning Questionnaire* (PRQ).

Differential Item Functioning (DIF) across gender

A preliminary CFA revealed that the items measured a single dimension, that the 2PL model provided a superior fit to the data, and all items fit the 2PL model in both groups. Analyses of DIF across genders was performed applying Item Response Theory Likelihood Ratio test approach implemented in the IRTPRO software (Cai, Thissen, & du Toit, 2011). The IRTLR is based on a nested model comparison approach. For each item, two models are compared, one in which all parameters (discrimination and difficulty) are constrained to be equal across groups, and one with separate estimation of all parameters. For each item, the fit of a model constraining the item parameters to be equal between the two groups was compared with a model allowing the parameters to be estimated freely in the two groups. No items exhibited cross-group DIF with respect to discrimination values. To test DIF in the difficulty parameters, two models were compared, one in which the parameters were constrained to be the same and one in which the parameters were freely estimated. Three items displayed a significant difference, but the effect size of DIF indicated a low magnitude. In sum, no items showed considerable cross-gender DIF.

Validity

As expected, there was a positive correlation between probabilistic reasoning score and mathematical ability, numeracy and math self-efficacy.

	PMP	***MSES-R***	***NUM***
PRQ	.59**	.42**	.52**
	(*n*=436)	(*n*=418)	(*n*=341)

** $p <. 01$

CONCLUSION

In this work we presented a new scale for evaluating probabilistic reasoning, and we used IRT analyses to evaluate the measurement properties of the scale. The PRQ accurately measures low levels of ability, and it is helpful in identifying students with difficulties in this domain. That is, the scale could be useful for identifying students who struggle with probabilistic reasoning at the beginning of an introductory statistics course. Once identified, these students could be aided in improving their probability skills through ad-hoc training courses focusing on the probability content that is required by the exam.

REFERENCES

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO 2.1 for Windows*. Chicago: Scientific Software International.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Galli, S., Chiesi, F., & Primi, C. (2011). Measuring mathematical ability needed for "nonmathematical"majors: The construction of a scale applying IRT and differential item functioning across educational contexts. *Learning and Individual Differences, 21*, 392–402.

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing students statistical reasoning: Connecting research and teaching practice*. Dordrecht, The Netherlands: Springer Publishing.

Garfield, J., delMas, R., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M. Lovett and P. Shah (Eds.), *Thinking with Data (Proceedings of the 33rd Carnegie Symposium on Cognition)* (pp. 117-147). New York: Erlbaum.

Konold, C., & Kazak, S. (2008). Reconnecting Data and Chance. *Technology Innovations in Statistics Education, 2*(1). http://repositories.cdlib.org/uclastat/cts/tise/vol2/iss1/art1

Kranzler, J., & Pajares, F. (1997). An exploratory factor analysis of the Mathematics Self-Efficacy Scale-Revised, MSES-R. *Measurement and Evaluation in Counseling and Development, 29,* 215-228.

Lipkus,I.M., Samsa, G., & Rimer, B.K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*, 37-44.

Primi, C., Morsanyi, K., Donati, M., Galli S., & Chiesi, F. (submitted). Measuring probabilistic reasoning: The construction of a new scale applying IRT. *Contemporary Education Psychology.*

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of literature. *Journal of Statistics Education, 16*(2) Retrieved from: http//www.amstat.org /publication/jse/v16n2/zieffler.html