Measuring probabilistic reasoning: the development of a brief version of the Probabilistic Reasoning Scale (PRS-B)

Caterina Primi, Maria Anna Donati, Sara Massino, Elisa Borace, Edoardo Franchi, Kinga Morsanyi

Measuring probabilistic reasoning: the development of a brief version of the *Probabilistic Reasoning Scale (PRS-B)*

Caterina Primi[1], Maria Anna Donati[1], Sara Massino[1],

Elisa Borace[1] Edoardo Franchi[1] and Kinga Morsanyi[2]

[1]NEUROFARBA – Section of Psychology, University of Florence, Florence, Italy; primi@unifi.it

[2]School of Psychology, Queen's University Belfast, Belfast, UK

The assessment of probabilistic reasoning skills is important in various context. The aim of this study was to develop an abbreviated, open-ended version of the Probabilistic Reasoning Scale (PRS-B), applying Item Response Theory (IRT). The analyses based on the open-ended version of the 16-item scale suggested the exclusion of seven items that did not perform well in measuring the latent trait. The resulting 9-item scale (PRS-B), which included highly discriminative items, covered a wider range of the measured trait than the original scale and it showed high measurement precision. Concerning validity, the results showed the expected correlations with numerical skills, math anxiety, and statistics achievement. In conclusion, the PRS-B can be is a shorter, open-ended version of the PRS that maintains excellent reliability and validity.

Keywords: Assessment, item response theory, short form, probabilistic reasoning, validity.

Introduction

The ability to think about uncertain outcomes and to make decisions on the basis of probabilistic information is relevant in many fields (e.g., business, medicine, politics, law, and psychology). However, people often struggle with interpreting probability information (see Chernoff & Sriraman, 2014). In particular, when they solve probability problems, people tend to make some typical mistakes instead of applying formal analytic procedures (see e.g., Gilovich, Griffin, & Kahneman, 2002). These mistakes, including heuristics, biases and misconceptions, make it difficult for people to interpret and critically evaluate probabilistic information, understand data-related arguments, and make reasoned judgments and decisions (e.g. Garfield & del Mas, 2010, Morsanyi et al., 2009; Pratt & Kazak, 2017).

Regarding education in probabilistic reasoning, previous studies demonstrated the difficulty of improving probabilistic reasoning ability or, rather, a resistance to eliminate probabilistic reasoning biases once these had consolidated (for a summary of the literature, see e.g., Gilovich et al., 2002; for adolescents, see Klaczynski, 2004). Nevertheless, training activities that target specific difficulties can reduce some misconceptions (e.g. Fong & Nisbett, 1991). Additionally, in statistic courses, which have been incorporated into a wide range of school and university programs in many countries, students often encounter difficulties and, eventually, many of them fail to pass the exams. Some authors (Konold & Kazak, 2008) pointed out that one of the reasons students have difficulties in learning basic data analysis stem from a lack of basic understanding of probability and in grasping the fundamental ideas of probability.

Given the important role of probabilistic reasoning skills in various contexts, it is important to accurately measure probabilistic reasoning skills, in order to improve these skills. This, in turn, can

lead to other beneficial outcomes, such as better academic achievement. For this reason, recently we developed a new scale, the *Probabilistic Reasoning Scale* – PRS (Primi, Morsanyi, Donati, Galli, & Chiesi, 2017) to measure probabilistic reasoning ability, which can be used to identify people with difficulties in this domain. The PRS was developed with a focus on some typical biases and fallacies that are known to lead to incorrect responses. Indeed, the PRS is a useful tool in educational contexts to identify individuals who could be targeted by specific interventions.

However, the scale has the limitation that the items measure probabilistic reasoning skills most precisely in the lower ability ranges and it is not ideal for the measurement of ability levels around the mean. Additionally, the scale is composed of 16 items and it does not seem appropriate for large, multivariate studies in which many tests and scales need to be administered at the same time. Indeed, with research questions becoming increasingly complex, and involving a growing number of constructs, such as when investigating the role of probabilistic reasoning in decision making (e.g., Schiebener & Brand, 2015) or in risk taking (Donati, Primi, & Chiesi, 2014), shorter scales potentially offer added value.

In sum, the aim of this study was to develop an abbreviated, open-ended version of the PRS. To achieve this goal, we used Item Response Theory (IRT) analyses, which make it possible to select items that offer the most information in measuring the targeted underlying trait, that is, probabilistic reasoning. Specifically, IRT has potential benefits in shortening a scale because it makes it possible to evaluate the amount of information provided by each item of the scale for each trait level on the trait dimension through the Item Information Function (IIF). In other words, if the amount of information is large, the trait level can be estimated with high precision, if the amount of information is small, the trait cannot be accurately estimated. Thus, on the basis of item information, it is possible to select items that convey the higher amount of information along the entire range of the measured trait. Through the selection of items that perform better and assure adequate information along the different levels of the trait, a well-performing shortened scale can be obtained.

Additionally, IRT provides the Test Information Function (TIF), which evaluates the precision of the test at different levels of the measured construct instead of providing a single value (e.g., Cronbach's α) for reliability (Embretson & Reise, 2000). More precisely, the TIF provides information on how accurate the test is at estimating a trait along the whole range of trait scores. The more information the test provides at a particular trait level, the smaller the error associated with ability estimation, and the higher the local reliability. Since the TIF is generated by aggregating the IIFs, in general, longer tests will measure an examinee's attribute with greater precision than shorter tests. Nonetheless, in the IRT framework, item selection can be done ensuring that the TIF of the shortened scale maintains an adequate amount of information along the trait continuum, which is similar to the original scale.

Finally, given that a short form of a test should meet the same standards of validity as the full form (Smith, McCarthy, & Anderson, 2000), validity measures were administered to provide evidence that the abbreviated scale still measures probabilistic reasoning adequately. Thus, we expected to replicate the pattern of relationships established for the construct as measured by the long form of

the test. In particular, we investigated the relationship between the PRS-B, numerical skills, anxiety and self-confidence related to numbers, and statistics achievement.

Methods

Participants

The participants in this study were 316 psychology students (mean age = 20.53; SD = 2.9; 76% female) enrolled in an undergraduate introductory statistics course at the University of Florence in Italy. All students participated on a voluntary basis.

Materials

The original Probabilistic Reasoning Scale (PRS, Primi et al., 2017) consists of 16 multiple-choice questions. In this study, the scale was administered in an open-ended format, which required some minor changes in the wording of the items. The items include questions about simple, conditional and conjunct probabilities, and the numerical data are presented in frequencies or percentages (e.g., "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?").

The Mathematics Prerequisites for Psychometrics (MPP; Galli, Chiesi, & Primi, 2011) 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). A single composite score, based on the sum of correct responses, was calculated. In the present sample, Cronbach's α was .74. We used this measure as an estimate of students' math knowledge.

The *Abbreviated Math Anxiety Scale* (AMAS; Hopko, Mahadevan, Bare, & Hunt, 2003; Italian version: Primi, Busdraghi, Tomasetto, Morsanyi, & Chiesi, 2014) measures math anxiety experienced by students in learning and test situations. Participants have to respond on the basis of how anxious they would feel during the events specified (for example, "Listening to another student explain a math formula"). High scores on the scale indicate high math anxiety. A single composite score was obtained, based on participants' ratings of each statement. In the present sample, Cronbach's α was .84.

The *Subjective Numeracy Scale* (Fagerlin et al., 2007) is a subjective measure (i.e., self-assessment) of quantitative ability. An example item is "How good are you at working with fractions?" The items have to be rated on a 6-point Likert scale. A single composite score was computed based on participants' ratings of each item. Coefficient α in the current sample was .78.

Measure of statistics achievement. As a measure of achievement, we used the final examination grade. The exam consisted of a written task that included three problems to be solved by a paper-and-pencil procedure without the support of a statistics computer package, 5 multiple-choice and 2 open-ended questions (e.g., describe the properties of a normal distribution) and 1 output of data analyses conducted with *R-Commander* to interpret. For the problems, students were given a data matrix (3-4 variables, 10-12 cases) and they had to compute descriptive indices, draw graphs, and choose and apply appropriate statistical tests (identifying the null and the alternative hypotheses, finding the critical value, calculating the value of the test, and making a decision regarding

statistical significance). Grades range from 0-30. From 0 to 17 the grade is considered insufficient based on the Italian University Grading System. Thus, only students who obtain 18 or higher grades pass the examination.

Procedure

Participants completed the measures individually in a self-administered format in the classroom. Each task was briefly introduced, and instructions for completion were given. The answers were collected in a paper-and-pencil format. All participants completed the scale during the first week of an introductory statistics course. *Achievement* in statistics was measured at the end of the course during the exam session.

Data Analysis

Preliminarily, we tested the unidimensionality of the PRS evaluating local dependence (LD). LD is an excess of covariation among item responses that is not accounted for by a unidimensional IRT model and it was assessed using the χ^2 LD statistic (Chen & Tiessen, 1997), computed by comparing the observed and the expected frequencies in each of the two-way cross tabulations between responses to each pair of items. This diagnostic statistic is approximately distributed as standardized χ^2. Given this approximation, as a rule of thumb, values of 10 or greater indicate the presence of LD.

After having verified this assumption, unidimensional IRT analyses were performed. IRT models use the original response data for estimating probabilities of responses as a function of the latent trait θ (i.e., in the current study, probabilistic reasoning), which is defined as a continuous variable that conventionally has a mean of zero and *SD* of 1.0. This function describes the relation between the probability of endorsing a response given not only the respondent's level of θ but also the item's characteristics. A model with two parameters (2PL) was tested in order to estimate the item difficulty and discrimination parameters. The parameters were estimated by employing the marginal maximum likelihood estimation method with the Expectation-Maximization (EM) algorithm implemented in the IRTPRO software (Cai, Thissen, & du Toit, 2011). In the 2PL model, the two item parameters are item difficulty and item discrimination. The item difficulty parameter (β) or "location" represents the latent trait level corresponding to a .50 probability of correctly endorsing the item. The item discrimination parameter (a) or "slope" represents the item's ability to differentiate between people at contiguous levels of the latent trait. This parameter describes how rapidly the probabilities change with trait levels. In order to test the adequacy of the model, the fit of each item under the 2PL model was tested computing the S χ^2 statistics. Additionally, for each item was calculated the IIF, graphically represented by the Item Information Curve (IIC), that describes the amount of information that a particular item provides across the entire continuum of the latent construct, and it depends on both the discrimination and location parameters. Thus, we used IIFs to select the items that conveyed the higher amount of information along the range of the trait measured by the PRS, looking at the area above the IICs, which equals both the size of the *a* parameters and the spread of the *b* parameters. Once the shortened scale was defined, all the above described analyses were repeated for the brief scale in order to confirm the item and test psychometric properties. In particular, we investigated the reliability of the shortened scale. IRT

makes it possible to assess the measurement precision of the test through the TIF.TIF is generated by aggregating the IIFs of items in a single measure and it allows to compute the information (*I*), that is, the expected value of the inverse of the standard error (*SE*), provided by the test at each level of the trait. Thus, the more information the test provides at a particular trait level, the smaller the error associated with trait estimation and the higher the test's reliability. Graphically, the TIF shows how well the construct is measured at different levels of the underlying construct continuum, and the peak of the TIF is where measurement precision is greatest. Regarding validity, Pearson product-moment correlations were computed.

Results

Preliminarily, we tested the unidimensionality of the probabilistic reasoning construct, a fundamental criterion underlying IRT models, evaluating the presence of LD. The results confirmed that a single factor model adequately represented the structure of the scale, as none of the LD statistics were greater than 10. Additionally, the factor loadings ranged from .54 to .80. Each item had a non-significant S-χ^2 value, indicating that all items fit under the 2PL model. Concerning the difficulty parameters (*b*), the results showed that the parameters ranged across the continuum of the latent trait from -2.68±.53 to .49±.12 logits (i.e., the logarithm of the odd, that is, the ratio of the probability of producing a correct response and the probability of responding incorrectly). Compared to the difficulty of the original scale (-2.97±.5 to -.07±.08 logits), results showed that the parameters spanned a wider range of the latent trait. With regard to the discrimination parameters (*a*), following Baker's (2001) criteria, all items showed adequate discrimination levels (*a* values over .60). Having verified the preliminary assumptions for IRT modelling, we looked at the IICs to select the items that conveyed the higher amount of information along the range of the trait measured by the open-ended PRS (see Figure 1). The figure clearly shows that items 1, 2, 3, 5, 13,15 and 16 provide lower amount of information.

Consequently, we retained nine items and we repeated the analyses for this shortened version of the PRS scale with an open-ended format. None of the LD statistics were greater than 10, indicating the absence of LD. All factor loadings were significant, ranging from .49 to .83. Each item had a non-significant S-χ^2 value, indicating that all items fit under the 2PL model. Concerning the difficulty parameters (*b*), the results showed that the parameters ranged from -1.57 (±.29) to .46 (±.11) logits across the continuum of the latent trait. With regard to the discrimination parameters (*a*), all items were above .60, indicating adequate discriminative power. Thus, still concerning item difficulty, the shortened version showed that the items had a low-medium level of difficulty. Nonetheless, the discriminative measures showed that the items could discriminate individuals with different trait levels.

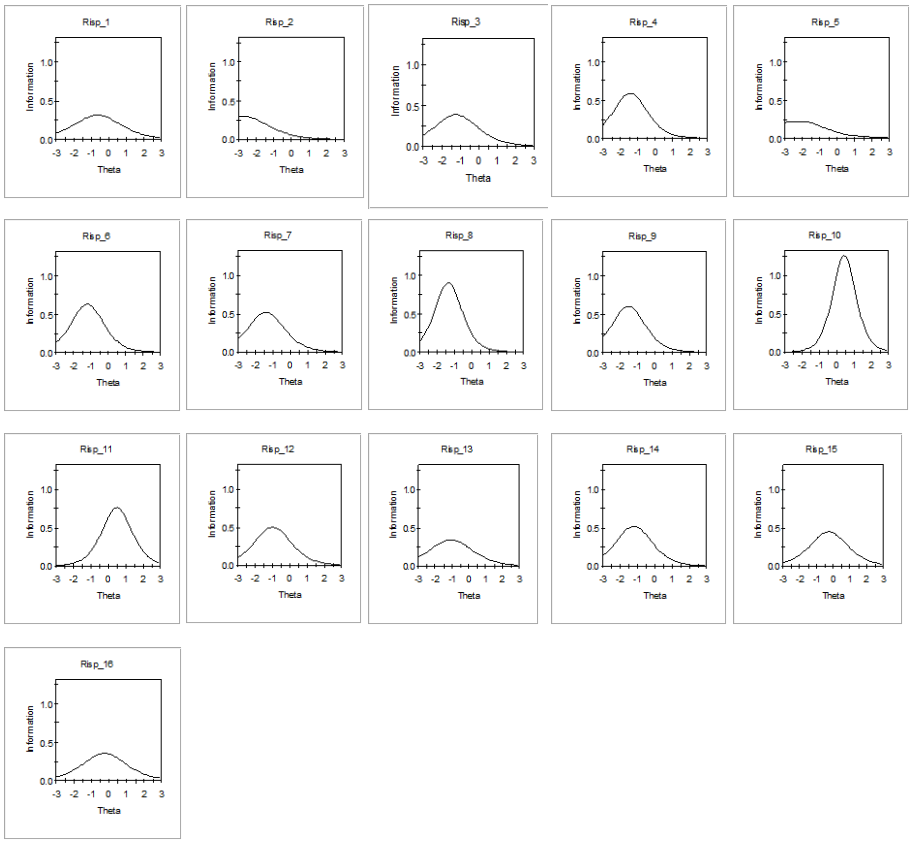

Figure 1: Item information curve (IIC) of the original 16-item PRS open format. Latent trait (theta) is shown on the horizontal axis and the amount of information and the SE yielded by the test at each trait level is shown on the vertical axis

Next, we investigated the TIF that provides test reliability estimations indicating the precision of the whole test for each level of the latent trait. As shown in Figure 2, from 2 standard deviations below the mean to 0.5 above the mean (fixed at 0), the amount of test information was equal to or greater than 4 indicating that the instrument was sufficiently informative for this range of the trait. Thus, compared to the original scale that adequately measured low levels of probabilistic reasoning ability, the PRS-B accurately measures a wider range of the underlying trait.

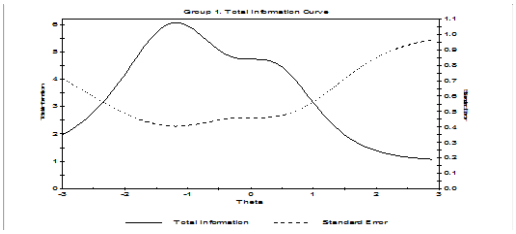

Figure 2: Test information function of the Probabilistic Reasoning Scale-Brief

Finally, we looked at the validity of the scale. Table 1 presents descriptive statistics for all measures in the study, and the relations between the PRS-B, and all other variables. As expected, in line with the original scale, the PRS-B was significantly and positively correlated with mathematics skills and subjective numeracy, and negatively with mathematics anxiety. Concerning the relationship between PRS-B and statistics achievement, we found a significant positive correlation.

	1	2	3	4	5
1. PRS-B	--				
2. PMP	.53***	--			
3. AMAS	-.19*	-.37***	--		
4. SN	.45***	.56***	.-.42***	--	
5. STATISTICS ACHIEVEMENT	.28**	.18*	-.15*	.09	--
M (*SD*)	5.99 *(2.25)*	22.38 *(3.69)*	23.39 *(5.76)*	32.14*(6.70)*	22.63 *(3.49)*

* $p<.05$; **$p<.01$;*** $p<.001$

Table 1: Descriptive statistics for the measures, and correlations between the PRQ-B and MPP, AMAS, SN, and statistics achievement

To ascertain the predictive value of probabilistic reasoning as measured by the PRS-B, we tested a regression model in which the PRS-B was entered as a predictor of statistics achievement along with math competence (MPP), math anxiety (AMAS), and subjective numeracy (SN). The results showed that when all predictors were entered in the regression analysis together ($F(4,133)=3.29$, $p=.013$; $R=.30$; $R^2=.09$), the PRS-B was a significant predictor of statistics exam performance ($\beta=.26$; $p=.009$), whereas the MPP, SN, and AMAS did not significantly predict statistics achievement.

Conclusion

In this study, we presented the PRS-B scale to assess probabilistic reasoning skills. Applying IRT, we obtained a shorter version of the original PRS, where all remaining items had good discriminative power and they measured a large spectrum of the trait which covered a wider range than the original scale. Additionally, the PRS-B accurately measures average levels of probabilistic reasoning ability, and it is helpful in identifying individuals who have difficulties in this domain. Finally, the validity results confirm that the shortened form replicates the pattern of relationships established for the construct as measured by the long form, with positive relations with mathematics and statistics skills and subjective confidence, and a negative relationship with mathematics anxiety. In conclusion, the PRS-B could be a useful tool in educational contexts, as well as in research.

References

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. *Chicago, IL: Scientific Software International*.

Chen, W. H., & Tiessen, D., (1997). Local dependence indices for items pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289. doi: 10.2307/1165285

Chernoff, E. J., & Sriraman, B. (2014). Commentary on Probabilistic Thinking: Presenting Plural Perspectives. In *Probabilistic Thinking* (pp. 721–727). Springer, Dordrecht.

Donati, M. A., Primi, C., & Chiesi, F. (2014). Prevention of problematic gambling behavior among adolescents: Testing the efficacy of an integrative intervention. *Journal of Gambling Studies*, *30*(4), 803-818.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672–680.

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General*, *120*(1), 34.

Galli, S., Chiesi, F., & Primi, C. (2011). Measuring mathematical ability needed for "non-mathematical" majors: the construction of a scale applying IRT and differential item functioning across educational contexts. *Learning and Individual Differences*, *21*(4), 392–402.

Garfield, J., & DelMas, R. (2010). A web site that provides resources for assessing students' statistical literacy, reasoning and thinking. *Teaching Statistics*, *32*(1), 2–7.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.

Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The abbreviated math anxiety scale (AMAS) construction, validity, and reliability. *Assessment*, *10*(2), 178–182.

Klaczynski, P. A. (2004). A dual-process model of adolescent development: Implications for decision making, reasoning, and identity. *Advances in child development and behaviour*, *32*, 73–125.

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education*, *2*(1).

Morsanyi, K., Primi, C., Chiesi, F., & Handley, S. (2009). The effects and side-effects of statistics education: Psychology students'(mis-) conceptions of probability. *Contemporary Educational Psychology*, *34*(3), 210–220.

Primi, C., Busdraghi, C., Tomasetto, C., Morsanyi, K., & Chiesi, F. (2014). Measuring math anxiety in Italian college and high school students: validity, reliability and gender invariance of the Abbreviated Math Anxiety Scale (AMAS). *Learning and Individual Differences*, *34*, 51–56.

Primi, C., Morsanyi, K., Donati, M. A., Galli, S., & Chiesi, F. (2017). Measuring probabilistic reasoning: The construction of a new scale applying item response theory. *Journal of Behavioral Decision Making*, *30*(4), 933–950.

Pratt, D., & Kazak, S. (2018). Research on Uncertainty. In *International Handbook of Research in Statistics Education* (pp. 193–227). Springer, Cham.

Schiebener, J., & Brand, M. (2015). Decision making under objective risk conditions–a review of cognitive and emotional correlates, strategies, feedback processing, and external influences. *Neuropsychology review*, *25*(2), 171–198.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological assessment*, *12*(1), 102.