Measuring probabilistic reasoning: The construction of a new scale applying Item Response Theory

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*. Advance online publication. https://doi.org/10.1002/bdm.2011

Published in:
Journal of Behavioral Decision Making

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Measuring probabilistic reasoning:

The construction of a new scale applying Item Response Theory

Caterina Primi*, Kinga Morsanyi°, Maria Anna Donati*, Silvia Galli & Francesca Chiesi*

*NEUROFARBA – Section of Psychology, University of Florence (Italy)

°School of Psychology Queen's University Belfast (UK)

Corresponding author:

Caterina Primi
Neurofarba – Section of Psychology
University of Florence (Italy)
Via S.Salvi 12 – Padiglione 26
50135 Florence - Italy
primi@unifi.it

Please cite as the following:
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 (in press)* DOI:10.1002/bdm.2011

Abstract

Probabilistic reasoning skills are important in various contexts. The aim of the present study was to develop a new instrument (the *Probabilistic Reasoning Scale* - PRS) to accurately measure low levels of probabilistic reasoning ability in order to identify people with difficulties in this domain. Item Response Theory was applied to construct the scale, and to investigate Differential Item Functioning (i.e., whether the items were invariant) across genders, educational levels and languages. Additionally, we tested the validity of the scale by investigating the relationships between the PRS and several other measures. The results revealed that the items had a low level of difficulty. Nonetheless, the discriminative measures showed that the items can discriminate between individuals with different trait levels, and the Test Information Function (TIF) showed that the scale accurately assesses low levels of probabilistic reasoning ability. Additionally, through investigating Differential Item Functioning, the measurement equivalence of the scale at the item level was confirmed for gender, educational status and language (i.e., Italian and English). Concerning validity, the results showed the expected correlations with numerical skills, math-related attitudes, statistics achievement, IQ, reasoning skills and risky choices both in the Italian and British samples. In conclusion, the PRS is an ideal instrument for identifying individuals who struggle with basic probabilistic reasoning, and who could be targeted by specific interventions.

Key words: Decision Making; Differential Item Functioning; Gender Differences; Item Response Theory; Probabilistic Reasoning; Scale; Validity

Introduction

In everyday life, an inability to make optimal choices on the basis of probabilistic information can be extremely costly, not only at the individual level, but also for society in general. Indeed, 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, psychology, etc.). Nevertheless, people often struggle with interpreting probability information (see Chernoff & Sriraman, 2014, for a recent review). For example, Bramwell, West and Salmon (2006) reported that only a small proportion of both healthcare professionals and patients were able to interpret the results of a genetic screening test correctly. An additional alarming finding of Bramwell et al. (2006) was that most people were highly confident about their incorrect judgments.

Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz and Woloshin (2008) highlighted the phenomenon of collective statistical illiteracy, and pointed out the associated societal costs. Statistical literacy is a key ability expected of citizens of information-laden societies, and it is generally considered a necessary component of numeracy (Gal, 2002; Rumsey, 2002; Utts, 2003). It enables people to interpret and critically evaluate probabilistic information, understand data-related arguments, and make reasoned judgments and decisions, i.e., to become reflective, engaged and constructive citizens. For this reason, education in statistics has been incorporated into a wide range of school and university programs in many countries. Probabilistic reasoning skills are considered to be necessary prerequisites of statistics learning (Konold & Kazak, 2008).

Given that people generally struggle with probabilistic reasoning, it is unsurprising that they often rely on simplifying heuristics (see e.g., Gilovich, Griffin, & Kahneman, 2002; Stanovich, 2004). The operation of some of these heuristics and biases can be demonstrated by asking people to make judgments about a sequence of random events (e.g., Ayton & Fischer, 2004; Bar-Hillel & Wagenaar, 1991; Kahneman & Tversky, 1972). Interestingly, although people have an intuitive grasp of some properties of random processes (e.g., Gauvrit & Morsanyi, 2014; Hahn & Warren, 2009), there are also some typical mistakes that they tend to make. For example, the gambler's

fallacy is the tendency for people to expect that, for random events, runs of a particular outcome will be balanced by a tendency for the opposite outcome (e.g., when tossing a coin, a series of heads should be followed by tails).

Probabilistic reasoning ability and its development has also been extensively studied in young people. One topic that has attracted particular interest is the role of probabilistic reasoning ability in tasks measuring decision making under risk in adolescence (e.g. Beyth-Marom, Austin, Fischhoff, Palmgren, & Jacobs-Quadrel, 1993; Halpern-Felsher & Cauffman, 2001). It has been shown that proportional reasoning ability (e.g., the ability to integrate the probability and amount of loss/reward; Huizenga, Crone, & Jansen, 2007; van Duijvenvoorde, Jansen, Visser, & Huizenga, 2010) and probability estimation (i.e., deciding which choice option has the greatest chance of resulting in a reward; van Leijenhorst, Westenberg & Crone, 2008) are important in making advantageous choices.

Probabilistic reasoning ability has also received increasing attention in research into the cognitive factors related to pathological gambling among adolescents (e.g., Derevensky, Gupta, & Baboushkin, 2007). In particular, adolescent problem gamblers' factual knowledge of probabilities of events in gambling, and their susceptibility to reasoning biases related to gambling outcomes, have been investigated. It has been found that adolescent problem gamblers are more prone to holding mistaken views about randomness when compared with non-problem gamblers, and they have erroneous beliefs about their chances of winning (Delfabbro, Lahn, & Grabosky, 2006; Delfabbro, Lambos, King, & Puglies, 2009; Turner, Zangeneh, & Littman-Sharp, 2006). Additionally, Donati, Chiesi, and Primi (2013) found that the ability to correctly reason about probabilities showed a significant negative relationship with problematic gambling behaviour.

Given the important role of probabilistic reasoning skills in various contexts, the aim of the present study was to develop a scale that accurately measures the basics of probabilistic reasoning. This would offer benefits for researchers and educators alike. For instance, in the domain of statistics education, identifying students with difficulties in probabilistic reasoning can help with

improving achievement and preventing failure, as students with difficulties can be supported with training activities from the start of their courses. In other settings, for example, in research into judgment and decision making, the scale could be used as a predictor of the quality of risky choices. The role of numeracy in decision-making abilities has been extensively studied and several numeracy scales exist (Hibbard, Peters, Slovic, Finucane, & Tusler, 2001; Thaler & Sunstein, 2003; Woloshin, Schwarz, & Welch, 2004). Nevertheless, these scales capture only certain aspects of probabilistic reasoning. For example, Lipkus, Samsa and Rimer's (2001) highly popular numeracy scale mostly concerns people's ability to transform information presented in percentages into a frequency format and vice versa.

The instrument that we present in the current paper – the *Probabilistic Reasoning Scale* (PRS) - was developed to measure various aspects of basic probabilistic reasoning. By basic skills, we mean skills that all individuals need to possess to be able to operate confidently and independently in everyday life, educational settings and work. To put it in another way, people who struggle with the problems presented in the test can be described as statistically illiterate (see Gigerenzer et al., 2008).

Regarding the content of the test, the basic function of probabilistic reasoning, and the standard interpretation of the concept of probability is that it answers the question of "how many out of how many[1]". Lipkus et al.'s (2001) numeracy test offers a good way of assessing people's ability to work with percentages and proportions. Nevertheless, this test does not measure some other important aspects of basic probabilistic reasoning, such as understanding conditional probabilities. The Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, Garcia-Retamero, 2012) assesses reasoning about more complex probability problems (including conditional probabilities), but does not consider more basic probabilistic reasoning skills. Thus, currently, there is no test available that would provide a general assessment of probabilistic reasoning ability. In order to

[1] In the studies that we present in this paper, we only investigated frequentist interpretations of probability, and did not include Bayesian problems.

address this issue, we developed a test that measures various aspects of probabilistic reasoning, including the understanding of basic and conditional probabilities presented in text and tables, reasoning about random sequences of events, and the ability to resist some typical fallacies and biases.

As we described above, people commonly hold both correct and incorrect intuitions and beliefs about randomness and probability. Indeed, there are strong traditions of research that demonstrate both the ways how intuitions about probability can be misleading (Baron, 2000; Gilovich et al., 2002; Stanovich, 2004), and how simple heuristics can lead to smart judgments and decisions (Gigerenzer, Todd, & the ABC Research Group, 1999; Hertwig, Herzog, Schooler & Reimer, 2008). When developing the current scale, we focussed exclusively on some typical biases and fallacies that are known to lead to incorrect responses (and did not consider the situations where heuristics might be useful). We did this, because our aim was to identify typical areas of weakness in probabilistic reasoning, rather than to explore all possible ways how intuitions and heuristics can influence probabilistic reasoning.

The scale also measures the ability to interpret probability information presented both in the form of natural frequencies and percentages. Whereas some authors (e.g., Cosmides & Tooby, 1996; Gigerenzer et al., 2008 Gigerenzer & Hoffrage, 1995) claim that using natural frequencies can help in interpreting probabilistic information, other researchers did not find that frequency formats improve probabilistic reasoning (e.g., Evans, Handley, Perham, Over & Thompson, 2000; Neace, Michaud, Bolling, Deer & Zecevic, 2008; Sloman, Over, Slovak, & Stibel, 2003). Specifically, when an advantage for frequency formats is found, this is not because of the presentation format *per se*, but because in the case of natural frequencies, the nested sets that are relevant to the computations are presented in a more transparent format.

In the PRS, we have included items using both frequency and probability formats. Specifically, we included three pairs of problems that were very similar in the information presented, with the exception that one item within each pair included frequency information and the

other item similar information in percentages. Although these pairs were matched in terms of the types of questions that they asked, we did not aim to eliminate some confounds that typically make it easier to process information that is presented in a frequency format. For this reason, we expected that the item within each pair that includes frequencies should be easier for participants. Overall, by including items using various formats for presenting probability information, our aim was to create a scale that helps in identifying people who struggle with probabilistic reasoning, and to also identify the specific skills and knowledge that they miss.

The scale was developed applying Item Response Theory (IRT) that overcomes some limitations of classical test theory (CTT) in test construction and scale evaluation (Embreston & Reise, 2000). IRT models assume that each examinee responding to a test item possesses some amount of the underlying ability and at each level of ability there will be a certain probability, denoted by $P(\theta)$, to give a correct answer to the item. Thus, this probability will be small for examinees of low ability and large for examinees of high ability. This approach derives the probability of each response as a function of the latent trait and some item parameters. In the 2PL model the two item parameters are, respectively, item difficulty and item discrimination. The item difficulty parameter (β) or " location" represents the latent trait level corresponding to a .50 probability of endorsing the item correctly. The item discrimination parameter (α) 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. It is important to note that an item's ability to discriminate between people changes across the trail-level region corresponding to item difficulty. That is, items are not equally informative across the entire trait range. In sum, IRT provides information that makes it possible to evaluate the performance of each item.

Another advantage of the IRT approach over CTT, is that IRT can provide unbiased estimates of item parameters even when a sample is unrepresentative (Embreston & Reise, 2000). This advantage is derived from the invariance property of IRT. Item parameters are independent of the

particular sample and person parameters are independent of the particular items. As a consequence, IRT parameters are invariant with respect to the sample characteristics from which they are generated.

Finally, IRT offers a better assessment of the measurement precision of scales than CTT. Instead of providing a single value (e.g., coefficient alpha) for reliability, through the *amount of information*, 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 **(**Embreston & Reise**,** 2000; Hambleton, Swaminathan, & Rogers, 1991). The more information the test provides at a particular ability level, the smaller the error associated with ability estimation is. Indeed, as IRT makes it possible to estimate the standard error of measurement for each level of the latent trait, it is possible to evaluate the quality of measurement at any level of the trait. This information can be used to develop or revise measures, so that they provide a high amount of information at particular levels of the latent trait.

Given that there are no existing instruments that were created with the aim to comprehensively measure various aspects of basic probabilistic reasoning, the purpose of the present work, organized in two studies, was to develop a unidimensional instrument to measure with high precision low levels of the latent trait of probabilistic reasoning skill. That is, the scale has to be appropriate for the accurate identification of respondents with weak probabilistic reasoning ability. In doing this, IRT is particularly useful, given that this psychometric 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, we constructed the items operationalized to measure basic probabilistic reasoning skills. 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. Subsequently, the measurement equivalence of the scale across genders, educational levels and languages was tested. 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 trait that is being measured do not have the same probability of responding to the item correctly. DIF makes it possible to test if the items measure the same trait dimension when administered to two or more distinct groups controlling for true group mean differences. Indeed, to compare groups of individuals with regard to their level of a trait, one must be sure that the numerical values that quantify that trait are on the same measurement scale.

Although IRT analyses can be used to confirm the accuracy of the scale in measuring the underlying construct, validity measures are needed to provide evidence of the appropriateness of the scale for measuring probabilistic reasoning by examining the statistical relationships with other constructs that are related to, but distinct from, probabilistic reasoning.

Study 1

The aim of this study was to construct and calibrate the scale, i.e., to estimate the item parameters, and to examine the reliability of the scale at different levels of ability. Once we established the properties of the scale applying IRT, the equivalence of item parameter estimates across genders and educational status was investigated in order to provide further evidence for the soundness of the scale. Finally, we explored the validity of the PRS as a measure of probabilistic reasoning as detailed below.

In line with 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 and statistics achievement. Indeed, Garfield and Ahlgren (1988) pointed out that one of the reasons why students might struggle with grasping the fundamentals of probability is that they have underlying difficulties with rational number concepts and basic concepts involving fractions, decimals and percentages.

We also investigated the relationship with attitudes (i.e., one's feelings and emotions) toward mathematics. Students' attitudes toward mathematics play an important role in learning mathematics, and positive attitudes correlate with higher achievement (Hemmings, Grootenboer, &

Kay, 2011). Moreover, we measured the relationship with math anxiety (a feeling of tension, dread or fear that arises when a person is required to undertake tasks involving math; Ascraft, 2002). We expected to find a negative correlation between probabilistic reasoning and math anxiety.

Regarding decision making, we measured the relationship between the PRS and advantageous decisions making under explicit risk (e.g., Brand, Labudda, & Markowitsch, 2006; Schiebener & Brand, 2015) among adolescents. We used the *Game of Dice Task* (GDT; Brand, Fujiwara, Borsutzky, Kalbe, Kessler, & Markowitsch, 2005), a task used previously in neuropsychology research, in which individuals have to decide between different alternatives that are linked to a specific amount of gain or loss. As the winning probabilities are explicit and stable over time, individuals have the chance to calculate the risk associated with each alternative from the very beginning. Given that probabilistic reasoning should help with understanding the likelihood of the different options, we expected a positive correlation with advantageous choices. Additionally, we measured the relationship between probabilistic reasoning and severity of problem gambling. We expected a negative correlation, with lower probabilistic reasoning skills in adolescents relating to higher levels of problem gambling habits.

Finally, individual differences in probabilistic reasoning are related to cognitive abilities, which play an important role in recognizing relevant information to avoid erroneous intuitions and heuristic responses (e.g., De Neys, Schaeken, & d'Ydewalle, 2005; Stanovich & West, 2008). Kahneman and Frederick (2002) pointed out that people with high cognitive abilities are more likely to possess relevant logical rules and recognize the need to apply these rules under various circumstances. Prior studies substantiated this relationship in adolescents and showed that those with high cognitive abilities were more likely to provide normatively correct responses to probabilistic reasoning tasks (Chiesi, Primi, & Morsanyi, 2011; Klaczynski, 2001; Kokis, MacPherson, Toplak, West, & Stanovich, 2002). Thus, we explored the relationship between the PRS and fluid intelligence, that is, the general factor of intelligence reflecting reasoning abilities

that operate across a variety of domains (Hicks, Harrison, & Engle, 2015) and we expected to find a positive correlation.

Method

Participants

The participants were 822 (52% male, *Mean age* = 18.43, *SD* = 3.3) students (61% from high school and 39% attending university). The adolescents were recruited from high schools in a suburban area in Italy (Tuscany). A detailed study protocol that explained the goal and methodology of the study was approved by the institutional review boards of each school. Students received an information sheet, which assured them that the data obtained would be handled confidentially and anonymously. All university students were enrolled in the first year of a psychology course at the University of Florence, and were recruited using opportunity sampling from various lectures and seminars. All students participated on a voluntary basis.

In the gender DIF analyses, the male group included 427 students (*Mean age* = 17.8, *SD*= 2.6), and the female group 395 students (*Mean age* = 19.1, *SD*= 3.8). In the educational status DIF analyses, the high school group included 505 students (70% males; *Mean age* = 16.9, *SD*= 1.6; range from 14.08 to 21.83) and the university group 317 students (23% males; *Mean age* = 20.7, *SD*= 3.9; range from 18.50 to 52.75).

Materials

The Probabilistic Reasoning Scale (PRS) *and its development*

The PRS was developed to measure basic probabilistic reasoning skills that are necessary to successfully interpret probability information in everyday settings, as well as to complete introductory statistics courses. Thus, the items were developed in order to operationalize low levels of the probabilistic reasoning latent trait. Items included simple, conditional and conjunct probabilities, and data were presented both in frequencies and percentages. Some of the items were based on existing materials (Chiesi & Primi, 2014; Morsanyi, Primi, Chiesi & Handley, 2009), others were constructed following the literature on this topic. Each item presented a multiple choice

question (one correct among three alternatives). In the case of some items, among the incorrect options, one corresponded to a specific bias (e.g. the gambler's fallacy or ratio bias). Thus, some items of the scale measure not only the ability to normatively reason in probabilistic terms but they also indicate susceptibility to specific heuristics and biases. The development of the scale was conducted in two steps. In the first step, a preliminary 12-item version of the scale was prepared (see Appendix A). This version was presented to several teachers of introductory statistics courses. In particular, they were asked to judge if the wording of the items was simple and understandable and if the items were suitable for measuring basic probabilistic reasoning skills. Moreover, we administered this version of the scale to a sample of high school students (n=95). The Equiprobability Bias (Compound events) item and the Law of Small Numbers item did not correlate with any other items of the scale. Additionally, the Law of Small Numbers item was considered by experts as unsuitable to measure basic probabilistic reasoning. As a consequence, these two items were omitted. In the next phase, six new items including conditional and conjunct probabilities were added, with data presented as frequencies (either in text or a table) or percentages. The new items were again evaluated by several teachers of introductory statistics courses. They considered the items appropriate to measure probabilistic reasoning using base-rates, percentages, and tree diagrams. The wording of the items was also considered appropriate. We administered the new set of items together with the previously selected ones to a new sample of high school and university students (n=87). As the two sets of items correlated strongly (r=.56, p<.01), we combined them and obtained the final, 16-item version of the *Probabilistic Reasoning Scale* (PRS). All items had a multiple choice format (one correct out of three response options). The items included simple, conditional and conjunct probabilities, and data were presented as frequencies or percentages (for example: "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%) (see full scale in Appendix B).

Measure of mathematical skills

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). For example: "The value 0.05 is" a) lower than 0; b) between -1 and 0; c) higher than 0.1; d) between 0 and 1; and "Knowing that xy = 3 which of the following is true? a) y=3/x; b) y=3-x; c = 3x; d) xy/3. A single composite score, based on the sum of correct responses, was calculated. In the present sample Cronbach's alpha was .87. We used this measure as an estimate of students' math knowledge.

Measure of attitudes toward math

The *Attitudes Toward Mathematics Inventory* (ATMI; Tapia & Marsh, 2004) is a 40-item scale to measure students' attitudes toward mathematics. The Italian version of the ATMI was obtained from the English version using the forward-translation method. Two non-professional translators worked independently, and then they compared their translations with the purpose of assessing equivalence. Subsequently, a group of five people read this first version, revised it, and eventually obtained a final form. Students were asked to respond to a series of statements (for example: "*Mathematics is important in everyday life*" or "*I feel a sense of insecurity when attempting mathematics*") using a scale from 1 (Strongly Disagree) to 5 (Strongly Agree). Negatively worded items were reverse scored and items were summed to create a single composite score, with higher scores indicating higher levels of attitudes toward math. In our sample, Cronbach's alpha was .96.

Measure of math anxiety

The *Abbreviated Math Anxiety Scale* (AMAS; Hopko et al., 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*" or "*Starting a new chapter in a math book*") using a 5-point response scale (ranging from "strongly agree" to "strongly disagree"). 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 alpha was .86.

Measure of statistics achievement

As a measure of achievement, we considered the final examination grade. The exam consisted of a written task and an oral exam. The written task included six problems – to be solved by a paper-and-pencil procedure without the support of a statistics computer package –, and four conceptual, open-ended questions (e.g., defining the null hypothesis in hypothesis testing). For the problems, students were given a data matrix (3-4 variables, 10-12 cases) and they had to compute descriptive indices, report data in a two-way table or 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). The marks for the written exam ranged from 0 to 30. Students obtaining a mark of 18 or higher were admitted to the oral exam. The oral exam provided the students with an opportunity to improve their performance (from 0 to 3 points). They were asked to comment on the mistakes made in the written task and/or to complete the tasks with missing answers. The final grade – derived both from the written and verbal parts – ranged from 18 to 30 in accordance with the Italian University Grading System.

Measure of decision making under risk

The *Game of Dice Task* (GDT; Brand et al., 2005) was used to measure risky decision making. In this task, participants were instructed to maximize their fictitious starting capital of €1,000 within 18 throws of a single virtual die. In each trial, before the die is thrown, participants must bet on the outcome of the die throw choosing one option: a single number, or combinations of two, three, or four numbers. If they choose one of the six possible single number options (from "1" to "6,"

winning probability = .17), they receive a fictitious gain of €1,000 when the chosen number is thrown but a fictitious loss of €1000 when one of the five unchosen numbers is thrown. Choosing one of the three possible combinations of two numbers ("1,2"–"3,4"–"5,6"; winning probability = .33) is linked to a gain of €500 when one of the numbers included in the chosen combination is thrown, but a loss of €500 when one of the numbers not included in those combinations is thrown. A further alternative is to choose one of the two possible combinations of three numbers ("1, 2, 3"–"4, 5, 6"; winning probability = .50) linked to a potential gain/loss of €200. Finally, participants may choose one of the three possible combinations of four numbers ("1, 2, 3, 4"–"2, 3, 4, 5"–"3, 4, 5, 6"; winning probability= .67) that will lead to a gain of €100 in the event that one of the four numbers chosen is thrown, but a loss of €100 when one of the numbers included in the other two unchosen combinations is thrown. The participants can choose each of the 14 different alternatives (clustered in four groups) in each trial. After each throw, the gain or loss in money is indicated on the screen accompanied by a distinct sound (the jingle of a cash machine for gains; a dull tone for losses). The current total capital and the number of remaining rounds are also displayed on the computer screen. In line with other studies (e.g., Brand & Schiebener, 2013; Brand, Laier, Pawlikowski, & Markowistch, 2009; Schiebener, Zamarian, Delazer, & Brand, 2011), a net score was calculated by subtracting the number of disadvantageous or high risk choices from the number of advantageous or low risk choices. Thus, the higher the net score, the more advantageous choices were made by the participant.

Measure of problem gambling severity

Problem gambling severity was measured by the *South Oaks Gambling Screen-Revised for Adolescents* (SOGS-RA; Winters, Stinchfield, & Fulkerson, 1993; Italian version: Colasante et al., 2013), one of the most widely used measures of adolescent gambling behavior (Edgren et al., 2016). Item Response Theory analyses have shown that the SOGS-RA is an effective screening tool (Chiesi, Donati, Galli, & Primi, 2013). Participants were initially asked to indicate their frequency of gambling based on a list of gambling activities. Then they were presented with 12 items related

to pathological gambling from which the total score was derived. In the present study, internal consistency of the SOGS-RA was satisfactory with a Cronbach's alpha of .71.

Measure of fluid intelligence

Set I of the *Advanced Progressive Matrices* (APM-Set I; Raven, 1962) is a measure of fluid intelligence, and it was used as a short form of *Raven's Standard Progressive Matrices* (SPM, Raven, 1941). Set I of the APM includes 12 items that increase in their level of difficulty, and cover the full range of difficulty of the SPM (Raven, 1962). The items consist of a series of perceptual analytic reasoning problems, each presented in the form of a matrix. In each case, the lower right corner of the matrix is missing and the participant's task is to determine which piece out of eight possible options fits into the missing space, so that the row and column transformations are satisfied. Using IRT analysis procedures, the short form of the SPM was found to have high reliability and validity (Chiesi, Ciancaleoni, Galli, & Primi, 2012).

Procedure

University students (N=317, 23% male, *Mean age* = 20.74, *SD* = 3.9) 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 PRS, MPP, ATMI and AMAS during a statistics class at the start of the term. The order of test administration was fully randomized (we obtained 4! = 24 versions of the questionnaire) and there was about an hour and a half to complete the scales. For a subsample of these students (*n* = 129, 20% male, *Mean age* = 20.42, *SD* = 2.8) the final statistics exam grade was collected at the end of the term. The written exam was timed (2 hours) and was followed by an oral examination. Concerning high school students, as in other studies (e.g., Panno, Lauriola, & Figner, 2013; Panno, Pierro, & Lauriola, 2013), participants (*n*=505; 70% male, *Mean age* = 16.97, *SD* = 1.6) were tested in two separate sessions, which were presented as two unrelated studies. In the first session, participants completed the PRS, the SOGS-RA, and the APM–Set I. The measures were administered in the classroom by trained experimenters. In the second session, a subsample of the

participants (*n*=352, 70% male, *Mean age* = 17.12, *SD* = 1.6) completed the GDT in an individual setting on a desktop computer.

Results

Calibration and Test Information Function

We tested the unidimensionality of the probabilistic reasoning skill construct, a fundamental criterion underlying IRT models, evaluating the presence of local dependence (LD). LD is a term used to describe excess covaration among item responses that is not accounted for by a unidimensional IRT model. To investigate the LD, the χ^2 LD statistic (Chen & Thissen, 1997) was used. 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. The factor loadings were all significant (p<.001) ranging from .34 to .81 (Table 1).

Having verified the assumption that a single continuous construct accounted for the covariation between item responses, unidimensional IRT analyses were performed. The two-parameter logistic model (2PL) was tested in order to estimate the item difficulty and discrimination parameters. 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 (α) 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.

The parameters were estimated by employing the marginal maximum likelihood (MML) estimation method with the EM algorithm (Bock & Aitkin, 1981) implemented in IRTPRO software (Cai, Thissen, & du Toit, 2011). 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. As large samples lead to a greater likelihood of significant chi-square differences, the critical value of p=.01, rather than p=.05, was used (Stone & Zhang, 2003). 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 -2.97 ± .5 to -0.07 ± .08 logit[2] across the continuum 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; Table 1).

Insert TABLE 1 here

We investigated the *Test Information Function* (TIF) which provides test reliability estimations indicating the precision of the whole test for each level of the latent trait (Embretson & Reise, 2000). The information (I) is the expected value of the inverse of the error variances for each estimated value of the underlying construct [$I(\theta) \approx 1/SE^2(\theta)$]. This means that the more information the test provides at a particular ability level, the smaller the error associated with ability estimation is and the higher the test's reliability. In terms of graphical presentation, the test information curve 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.

As shown in Figure 1, from 3 to -.05 standard deviations below the mean to the mean (fixed at 0), the amount of test information was equal to or greater than 4 (which yields a standard error of estimate equal to or less than .50) indicating that the instrument was sufficiently informative for this range of the trait. We can interpret the amount of information by computing the associated reliability (r = 1-1/Information). Reliability was equal to or greater than .80 (alfa coefficient) within the described range[3]. Thus, the PRS adequately measured low levels of probabilistic reasoning ability.

Insert FIGURE 1 here

[2] The logit is the logarithm of the *odd*, that is, the ratio of the probability of producing a correct response and the probability of responding incorrectly.

[3] The reliability of the total scale calculated with the traditional Cronbach's alpha was .70.

Differential Item Functioning (DIF) across gender and educational status

In order to investigate the invariance property of the items of the scale, analyses of differential item functioning (DIF) across genders were performed, applying an IRT likelihood ratio test approach implemented in IRTPRO software (Cai, Thissen, & du Toit, 2011). The DIF detection procedure 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. This procedure involves comparing differences in log-likelihoods (distributed as chi-square) associated with nested models. Since multiple tests were performed, the level of significance of .05 was adjusted by Bonferroni correction to .003 (0.05/16).

In the first step of the gender DIF analyses, we used the male group as the reference group. We found gender equivalence for the discrimination parameter (*a*) and, concerning difficulty (*b*), we found nonequivalence only for Items 14 and 16 ($p<.003$). Then, using all the other items as "anchor" items, the analysis was repeated. Anchor items are assumed without DIF, and are used to estimate the trait, and to link the two groups being compared in terms of trait levels. Anchor items are selected through a process of log-likelihood comparisons performed iteratively, and called a "purification" procedure. During this iterative process, the DIF status of items may change as a result of using a less than optimal conditional variable at various steps in the analyses. During this iterative process, the DIF status of items 14 and 16 did not change. Thus, we can conclude that both items showed uniform DIF that indicates that the difference is in the same direction across the entire spectrum of the construct to be measured (i.e., at all levels of the trait one group is consistently more likely than another to endorse an item). Nonetheless, given that the PRS only exhibits minor

non-invariance[4] (only 12% of the total number of items that compose the scale are non-invariant) it would be highly useful in research into gender differences.

Concerning educational status DIF, comparing high school students (reference group) and university students, we found educational status equivalence for the discrimination parameter (*a*) and, concerning difficulty (*b*), we found nonequivalence only for Item 4 and item 7 ($p<.003$). Then, using all the other items as "anchors", the analysis was repeated. During this iterative process, the DIF status of item 4 and item 7 did not change. Nonetheless, as the PRS exhibits only minor non-invariance (only 12% of the total number of items that compose the scale are non-invariant) it would be highly useful in research comparing groups with different levels of education.

Having preliminarily attested the measurement equivalence of the scale, a 2 (gender) X 2 (educational level) analysis of variance (ANOVA) was conducted with the PRS score as the dependent variable. When comparing female students ($M= 12.11$; $SD = 2.5$) and male students (M= 12.45; $SD = .2.6$), the main effect of gender was significant ($F(1,818) = 20.46$; $p<.001$ $\eta^2 = .02$). Similarly, when comparing high school students ($M= 12.01$; $SD = 2.6$) and university students ($M= 12.73$; $SD = 2.59$), the main effect of educational level was significant ($F(1,818) = 32.37$; $p<.001$ $\eta^2 = .04$). Finally, no significant interaction effect was found ($p=.13$).

Validity

First we present the relationships between the measures that were administered to the university students. As expected, there was a significant positive correlation between probabilistic reasoning skills, mathematical ability and attitudes toward math, and a significant negative correlation with math anxiety. Concerning the relationship with statistics achievement, we found a significant positive correlation (Table 2).

Insert TABLE 2 here

[4] A high level of Differential Test Functioning (DTF) exists if 25% or more of the items are categorized as having a significant DIF (Penfield & Algina, 2006).

To ascertain the predictive role of probabilistic reasoning as measured by the PRS, we tested a regression model in which the PRS was entered as a predictor of statistics achievement along with math competence (MPP), attitudes toward math (ATMI), and math anxiety (AMAS).

As some of the performance measures were highly correlated, we also conducted a multi-collinearity analysis for the regression analysis. According to the criteria proposed by Myers (1990), which specifies that a variance inflation factor (VIF) of ten or greater is cause for concern, the VIFs obtained for each predictor were at acceptable levels (the VIFs ranged from 1.49 to 1.64). Results showed that, when entered in the regression analysis together with the PRS, the MPP, ATMI and AMAS did not significantly predict achievement, whereas the PRS was a significant predictor (Table 3).

Insert TABLE 3 here

Next we analyzed the results regarding the measures administered to high school students. As expected, there was a significant positive correlation between probabilistic reasoning score and advantageous choices on the game of dice task. PRS scores were also significantly and positively correlated with fluid intelligence. Concerning the relationship with problem gambling behavior, we found a significant negative correlation (Table 4).

Insert TABLE 4 here

To ascertain the predictive role of probabilistic reasoning, we tested a regression model in which the PRS was entered as a predictor of advantageous choices (GDT- net score) together with fluid intelligence (APM) and problem gambling behavior (SOGS). The results showed that the PRS

was a significant predictor whereas the APM and SOGS did not significantly predict advantageous choices (Table 5).

Insert TABLE 5 here

Discussion

In this study, we presented a new scale to assess probabilistic reasoning skills, and we used IRT analyses to evaluate the measurement properties of the scale. The choice of using IRT analyses to develop the scale is consistent with the theoretical standpoint which identifies IRT as an appropriate approach to construct scales aimed to accurately measure specific levels of the assessed ability (Embreston & Reise, 2000). Concerning item difficulty, the results revealed that the items had a low level of difficulty, indicating that their contents are easy for respondents. Nonetheless, the discriminative measures showed that the items can discriminate individuals with different trait levels, and the TIF showed that the scale accurately assesses low levels of probabilistic reasoning in accordance with the purpose of constructing a scale to identify individuals with low levels of ability.

This approach can be used not only to identify people who struggle with probabilistic reasoning, but also to establish the specific concepts that they have difficulties with, and the particular heuristics and biases that they are susceptible to. Additionally, through investigating Differential Item Functioning inside the IRT framework, the measurement equivalence of the scale at the item level was confirmed for gender and educational status. These findings revealed that the scale can be considered equally suitable and fair for males and females, and high school and university students. Thus, group differences or similarities in mean levels of the construct reflect true differences or similarities among groups, and they are not the artifacts of the measurement process. In the current sample, males showed a small advantage in probabilistic reasoning over females, and probabilistic reasoning ability increased with age.

Regarding item difficulty, items 2 and 3, 12 and 10, and 13 and 11 were designed in such a way that they were very similar in their structure, but the first item within each pair was presented in a frequency format, and the second item included probability information in percentages. Comparing the difficulty parameters within each pair, it is apparent that the difficulty of items with frequency information is consistently lower. This supports the notion that presenting items in a frequency format might boost performance (e.g., Gigerenzer et al., 2008), although a careful comparison of items within each pair also highlights that items including frequencies are somewhat easier to process, and they require simpler computations, which could explain the difference in item difficulties. In particular, as suggested by Evans et al. (2000) frequency versions of the problems present set inclusion relationships in a more transparent manner than percentages. For example, when people compute the 30% of 90% (see item 11), they have to understand that the parent population has changed (i.e., 90% of the original population represents the full relevant population for the purposes of the computation of 30%). They also need to be able to retrieve the relevant formula to compute the result (for example, 0.3x0.9 or 0.3/100x90). By contrast, when they work on item 13 (i.e., "out of 1000 inhabitants of a village 600 people own a pet, and amongst pet owners 1 in 3 own more than 1 pet"), it is clearly stated that the 1 in 3 proportion should be applied to the 600 pet owners, rather than the full sample. Although there are still additional computations needed (for example, 600/3), this computation will result in a whole number, rather than a fraction, which makes the interpretation easier.

Items 4-9 measured some typical heuristics related to randomness. Based on the item difficulty parameters, these items were amongst the easiest. Finally, participants found it relatively difficult to process probability information presented in a table (items 14-16). A recent paper by Clinton et al. (2016) used eye-tracking data to investigate the processing of probability information presented in tables. These researchers concluded that, although information laid out in tables might look particularly clear and well-organized, integrating information from text and tables might pose a challenge, as it requires switching attention back and forth between the two presentation formats.

Clinton et al. (2016) also found that using an interactive interface that attaches meaningful labels to each cell of the table can help with processing probability information, possibly because it makes it easier to integrate the information presented in the table with the corresponding information in the accompanying text.

The validity measures of the scale provided evidence for the appropriateness of the scale for measuring probabilistic reasoning skills. In line with previous studies on statistics achievement, we found that probabilistic reasoning shows a positive correlation with mathematical ability and attitudes toward math, and a negative correlation with math anxiety. Moreover, it is a significant predictor of statistics achievement, whereas mathematical skills, attitude towards math and math anxiety were not significant when they were included in a regression model together with the PRS. Additionally, the PRS longitudinally predicted statistics exam results. These findings are extremely important because they confirm the uniqueness of the construct measured by the PRS and they show the usefulness of the scale in introductory statistics teaching. Indeed, once students who are more likely to encounter difficulties in their statistics course are identified, training exercises could be offered, focusing on the probabilistic reasoning skills required by the exam in order to improve students' understanding and performance (e.g., Zieffler et al., 2008).

Additionally, regarding the role of probabilistic reasoning in decision making under explicit risk conditions (e.g., Brand et al., 2006; Schiebener & Brand, 2015), we measured the relation between the PRS and advantageous decision making. The results were in line with previous studies confirming that probabilistic reasoning abilities are related to reasoning about the likelihood of the different options (e.g., Huizenga et al., 2007; van Duijvenvoorde et al., 2010; van Leijenhorst et al., 2008). Probabilistic reasoning also acts as a protective factor against developing problem gambling habits among adolescents (e.g., Delfabbro et al., 2006; Delfabbro et al., 2009; Donati et al., 2013; Turner et al., 2006). Moreover, we confirmed the relationship between probabilistic reasoning ability and cognitive abilities (Chiesi et al., 2011; Klaczynski, 2001; Kokis et al., 2002).

In conclusion, the PRS could be useful in real-life settings (including educational contexts), as well as in research. Whereas the role of probabilistic reasoning in statistical thinking is well-established, probabilistic reasoning is also increasingly studied as an explanatory factor in research into judgment and decision making. Based on its utility in these areas, the scale could be used to develop common international intervention strategies aiming to promote basic probabilistic reasoning. In line with this suggestion, in Study 2 we tested the invariance of the PRS scale across languages. Indeed, it is important to develop scales that are not specific for a single educational system (Hambleton, 2004; Hambleton & de Jong, 2003) or cultural context.

Study 2

One aim of Study 2 was to investigate the measurement equivalence of the PRS across languages (Italian and English) applying DIF analysis. We also measured the validity of the English version of the scale by administering it together with various different measures, including self-report questionnaires, measures of reasoning skills, and statistics achievement. We used scales and tasks that partially overlapped with the ones used in Study 1, in order to demonstrate the robustness of our findings across educational/cultural contexts.

Specifically, Study 1 demonstrated that performance on the PRS was related to mathematical ability, math-related attitudes and statistics exam performance. It was also related to fluid intelligence, risky decision making, and self-reported gambling problems. Additionally, the PRS was negatively related to math anxiety.

In the current study we again measured the relationship between performance on the PRS and math anxiety. We also administered the subjective numeracy scale (Fagerlin, Zikmund-Fisher, Ubel, Jankovic, Derry & Smith, 2007), a self-reported measure of numeracy skills that does not require actual computations. Another self-report questionnaire that we administered was the need for cognition scale (Cacioppo, Petty, Feinstein, & Jarvis, 1996), a measure of individual differences in the tendency to engage in and enjoy effortful cognitive activity, even without external motivation (Heijltjes, van Gog, Leppink, & Paas, 2014). Need for cognition has been found to be positively

related to academic performance and course grades (Leone & Dalton, 1988; Sadowski & Gulgoz, 1996). Students who score high on the scale are able to comprehend material requiring cognitive effort better (Leone & Dalton, 1988), and they are also more effective information processors (Sadowski & Gulgoz, 1996). Need for cognition has also been found to be related to probabilistic reasoning performance (Morsanyi et al., 2009; Clinton et al., 2016).

In addition to the self-report measures, we also administered two reasoning tests: the conditional probability task (based on Gauffroy & Barrouillet, 2009) and the CRT-Long (Primi, Morsanyi, Chiesi, Donati & Hamilton, 2016). The conditional probability task measures hypothetical thinking about the probability of events, but it requires almost no computations. Instead it measures participants' ability to correctly judge the truth or falsity of conditional statements about the probability of events that are displayed in a pictorial format.

The CRT-Long is an extended version of the cognitive reflection test (CRT; Frederick, 2005). It is related to cognitive ability, and to the tendency to resist making tempting but incorrect inferences. In other words, the CRT is a measure of rational, unbiased thinking (Toplak, West & Stanovich, 2011). Similar to the CRT, all tasks in the CRT-Long include numerical information, and performance on the task has been found to be moderately related to numeracy (Primi et al., 2016). We expected that both the CRT-Long and the conditional probability task should be positively related to probabilistic reasoning performance.

Finally, we also investigated the relationship between performance on the PRS and achievement on a statistics exam in a subsample of our participants. As in Study 1, the PRS was administered to all students at the start of the academic year. The statistics exam took place at the end of the academic year. Study 1 showed that in the case of Italian students, the PRS was a better predictor of statistics exam marks than another scale that was developed to measure basic mathematical skills that are necessary for learning statistics. In Study 2, we compared the predictive value of the PRS with the CRT-Long, as well as statistics lecture attendance.

In summary, in Study 2 we further examined the validity of the PRS as a test of numerical and reasoning skills. Furthermore, we tested the usefulness of the scale as a predictor of academic achievement in statistics among UK students. In Study 1 performance on the PRS at the start of the academic year explained a significant proportion of variance in exam marks at the end of the academic year in the case of Italian students. If this result replicates in a different educational context, it would be an indication that the scale could be used to identify students who are likely to struggle with their statistics courses early on. These students could then be offered additional support with their studies.

Method

Participants

The participants were 1060 undergraduate university students (Mean age = 19.8 years; *SD* = 3.7; 66% female) at the School of Psychology and the School of Medicine in Florence (Italy) (47%; Mean age = 18.8 years; *SD* = 3.5; 64% female) and in Belfast (United Kingdom) (53%; Mean age = 20.8 years; *SD* = 3.9; 68% female). All students participated on a voluntary basis. The UK participants obtained ungraded course credit for their participation.

Materials

The *Probabilistic Reasoning Scale* (PRS) was administered to a new Italian sample. An English version was prepared for the British sample (Appendix B)

Measure of math anxiety: The AMAS (Hopko et al., 2003) was also administered to the UK sample (Cronbach's alpha was .86).

The *Subjective Numeracy Scale* (SNS; Fagerlin et al., 2007) is a subjective measure (i.e. self-assessment) of quantitative ability that was developed with the aim of distinguishing between low-numerate and high-numerate individuals but in a less aversive and quicker way than it is possible with objective tests of numeracy. An example item is: 'How good are you at working with fractions?' The items have to be rated on a 6-point Likert scale, which are labelled differently for different questions (e.g. ranging from 1=not good at all to 6=extremely good; or 1=never to 6=very

often). A single composite score was computed based on participants' ratings of each item. Coefficient alpha in the current sample was .62.

The *Need for Cognition scale* (NFC; Cacioppo et al., 1996) measures a tendency to engage in and enjoy effortful cognitive activities. The scale consists of 18 items, and it uses a 5-point Likert scale that ranges from 1=not at all like me to 5=very much like me. A single composite score was computed based on participants' ratings of each item. Cronbach's alpha for the current sample was .75.

Measures of reasoning skills

In the *conditional probability task* (based on Gauffroy & Barrouillet, 2009) participants were presented with a visual display of eight cards, and they were given the following information. 'There are 8 cards in a pack. All the questions concern these 8 cards.' Some of the cards presented to participants were black and some were white, and some of them had a circle printed on them others had a square printed on them. The participants were then asked four questions about the cards, which had the following format: 'How likely are the following statements to be TRUE/FALSE of a card drawn randomly from the pack? If the card is black then there's a square printed on it.' Participants had to provide a response in the format of: '__ out of __' (i.e. participants had to fill the gaps with the relevant numbers). Two questions were asked about the likelihood of the statement being true, and two questions were asked about the likelihood of the statements being false. In the present sample, a single composite score was computed, based on the overall number of correct responses (out of 4). Cronbach's alpha was .82.

The *CRT-Long* (Primi et al., 2016) is an extended version of the cognitive reflection test (CRT; Frederick, 2015) that consists of 6 questions. Although the questions are open-ended, almost all participants produce either the correct response or a typical incorrect (i.e., heuristic) response. An example item is the following: 'If three elves can wrap three toys in one hour, how many elves are needed to wrap six toys in two hours? [correct answer = 3 elves; heuristic answer = 6 elves] Cronbach's alpha in the current study was .76. When we investigate the validity of the PRS, we

report the results both for the CRT-Long and for the 3 items included in the original CRT (Frederick, 2005).

Measures of statistics achievement and lecture attendance

Lecture attendance: The UK psychology students attended a 6-week statistics course that consisted of a 2-hour lecture each week. Attendance was recorded during weeks 2, 4 and 6 (thus, the recorded number of classes attended ranged from 0-3).

Statistics achievement: The students participated in a written exam session which required them to read descriptions of psychology experiments and interpret the statistics output related to the experiments that was produced by a software package. The students had to recognize the statistical test used, and they also had to find relevant information in the output tables (including descriptive and test statistics). Additionally, they had to state the null and the alternative hypotheses for the tests, make decisions regarding statistical significance, and summarize and report the results following the APA style. Performance was measured as a percentage of the questions answered correctly (i.e., it ranged from 0-100).

Procedure

The participants individually completed the measures 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. The Italian participants (n =500) only completed the PRS. The scale was administered at the start of the academic year during a statistics class. The UK participants (n=560) also completed all tasks and questionnaires during lecture time at the start of the academic year (at the end of September). The order of administration was the following: PRS, AMAS, CRT-Long, conditional probability task, SNS and NFC. In the case of a subgroup of the UK students (n=160) statistics lecture attendance was monitored during a 6-week period in February and March[5]. The exam performance of these students was also recorded at the

[5] The rest of the UK students were not enrolled on this statistics course.

end of May. Students in this subgroup were asked for permission that the researchers can access their academic records for the purpose of this study. After excluding students who did not give consent for their records to be used or who did not participate in the statistics exam, the final sample size was *n*=124 for lecture attendance and statistics achievement.

Results

Differential Item Functioning (DIF) across languages

Preliminarily, the unidimensionality of the scale and item fit under the 2PL model in each group (Italian and UK students) were tested in order to verify the possibility of using unidimensional IRT. The results showed that the unidimensionality assumption was met. Specifically, in each group none of the LD statistics were greater than 10 and factor loadings were all significant for each group (Italian from .37 to .87 and English from .38 to .79). Finally, all items had an acceptable fit under the 2PL model for each group. The DIF analysis across languages revealed equivalence for the discrimination parameter (a) and for difficulty (b). The results confirmed the equivalence of the scale across languages.

Validity

Table 6 presents the descriptive statistics and the correlations between the PRS and all other measures. As expected, the PRS was positively related to subjective numeracy, need for cognition, conditional probability reasoning, the CRT and CRT-Long and the result of the statistics exam.

Insert TABLE 6 here

We run a regression analysis to investigate the strength of each of these measures in predicting statistics exam results (Table 7). When the predictors were entered together into the regression equation, both the PRS and lecture attendance explained a significant proportion of variance, whereas the effect of all other measures was non-significant. As some of the performance measures were highly correlated, we also conducted a multicollinearity analysis for the regression

analysis. The VIFs obtained for each predictor were at acceptable levels (the VIFs ranged from 1.07 to 1.88).

Insert TABLE 7 here

Discussion

Study 2 confirmed the equivalence of the Italian and English versions of the PRS, attesting the appropriateness of the scale to be used in both educational contexts. This study also examined further the validity of the PRS. As expected, the test correlated with measures of both numerical and reasoning skills. Specifically, we found a positive relationship between the PRS and subjective numeracy, and a negative relationship with math anxiety. Although these relationships were not very strong, it should be noted that these scales are based on self-report, whereas the PRS measures actual performance. An additional self-report measure that was found to be related to performance on the PRS was need for cognition.

We also studied the relationship between the PRS and two measures of reasoning skills: the conditional probability task and the CRT-Long. As expected, performance on the PRS showed moderate positive relationships with these ability measures. Additionally, we investigated the suitability of the PRS to be used as a predictor of performance in introductory statistics courses. In line with the results of Study 1, performance on the PRS at the start of the academic year was related to statistics exam results at the end of the year. Importantly, the PRS was a better predictor of statistics performance than the CRT-Long (a test that also requires sound reasoning about numerical information). The PRS also explained additional variance in statistics performance when lecture attendance was taken into account. Together with the results of Study 1, and the finding of equivalence of the Italian and English versions of the scale, these results provide evidence for the appropriateness of the PRS to be used as a screening tool to identify students who are likely to struggle with their statistics courses in various countries and educational contexts.

Conclusion

In the current studies we developed a new scale for measuring probabilistic reasoning skills, and we used IRT analyses to evaluate the measurement properties of the scale. The PRS accurately measures basic probabilistic reasoning ability, and it is helpful in identifying individuals who have difficulties in this domain. Performance on the PRS was related to a broad range of measures, including measures of numeracy (the MPP), fluid intelligence (the APM), reasoning and decision making skills (the CRT, conditional probability reasoning, the game of dice task), and various self-report questionnaires (e.g., measures of mathematical anxiety and gambling related cognitions). These correlations were mostly of medium strength, which indicated that the PRS only partially overlapped with existing related measures. The usefulness of the PRS was also demonstrated by the fact that it predicted variance in some important real-life outcomes beyond the variance explained by other relevant tests and scales. In terms of practical implications, given the relationship between probabilistic reasoning and decision making under risk when explicit information about gains and losses is provided, the PRS may be used in interventions designed to promote probabilistic reasoning ability with the final goal of improving decision making abilities. Specifically, the psychometric properties of the scale allow practitioners for using it to identify adolescents with poor probabilistic reasoning abilities and in planning educational interventions.

The PRS also longitudinally predicted statistics achievement over the course of an academic year in both Italian and UK psychology students. This highlights the usefulness of the scale in another context: introductory statistics education. Given its links with statistical literacy (e.g., Gigerenzer et al., 2008) another field where measuring probabilistic reasoning skills could be useful is in understanding health-related risks, and making treatment choices on the basis of probabilistic information by both patients and medical professionals (e.g., Politi, Han & Col, 2007; Reyna & Brainerd, 2007). As the scale measures basic skills, it is appropriate for people with lower levels of education, or older adults, who finished their relevant education a long time ago.

Going back to the educational implications of this work, in addition to identifying adolescents who might be at risk of developing problem gambling habits or university students who might struggle with statistics courses, using the PRS can also help in identifying specific gaps and reasoning biases that could be targeted by *ad-hoc* training activities. Regarding education in probabilistic reasoning, previous studies demonstrated the difficulty of improving probabilistic reasoning ability, or rather, a resistance to change once probabilistic reasoning biases consolidated (for a summary of the literature, see e.g., Gilovich et al., 2002; for adolescents, see Klaczynski, 2004). For instance, Donati, Primi, and Chiesi (2014) tested the effectiveness of an intervention method to reduce problem gambling among adolescents. Donati et al. (2014) found that their intervention produced the hypothesized effects for knowledge and misconceptions on gambling, superstitious thinking, and economic perception of gambling, with the exception of the susceptibility to the gambler's fallacy. Similarly, although a training developed by Morsanyi, Handley and Serpell (2012) successfully decreased the equiprobability bias, the susceptibility to the representativeness heuristic increased as a side effect. Nevertheless, training in certain concepts, such as the law of large numbers, appears to be relatively easy (Fong, Krantz & Nisbett; 1986; Fong & Nisbett, 1991). A potential future direction of this work could be to test some training activities that could target the specific difficulties revealed by the scale. It would be particularly interesting to see whether improvements in these skills would lead to other beneficial outcomes (e.g., better academic achievement or a reduction in gambling-related misconceptions).

Although it could be considered a limitation of the scale that the items measure probabilistic reasoning skills most precisely in the lower ability ranges, it should be noted that many people struggle with basic probabilistic reasoning. It has been reported that up to 25% of economically active individuals in more economically developed countries lack basic numerical knowledge, skills and understanding that would be essential for them to operate confidently and independently in everyday life, educational settings and work (Snyder & Dillow, 2012).

In summary, the PRS is a valid and reliable instrument that provides a comprehensive measure of basic probabilistic reasoning skills. Focusing on low levels of probabilistic reasoning ability makes it an ideal instrument for identifying individuals who could be targeted by specific interventions (e.g., to improve statistics learning or to prevent the development of problem gambling habits). This instrument could also be useful in research into other types of risky decisions, such as choosing between medical treatments or financial investment options.

References

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness?. *Memory & cognition*, *32*(8), 1369-1378.

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

Chernoff, E.J. & Sriraman, B. (Eds). *Probabilistic Thinking: presenting Plural Perspective - Advances in Mathematics Education*. Springer Science + Business Media, Dordrecht, Netherlands

Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae. net/irt/baker..

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness.*Advances in applied mathematics*, *12*(4), 428-454.

Baron, J. (2000). Thinking and deciding (3rd ed.). Cambridge, UK: Cambridge University Press.

Beyth-Marom, R., Austin, L., Fischhoff, B., Palmgren, C., & Jacobs-Quadrel, M. (1993). Perceived consequences of risky behaviors: Adults and adolescents. *Developmental Psychology*, *29*, 549-563.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bramwell, R., West, H., & Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: experimental study. *BMJ*, *333*(7562), 284.

Brand, M., Fujiwara, E., Borsutzky, S., Kalbe, E., Kessler, J., & Markowitsch, H. J. (2005). Decision-making deficits of Korsakoff patients in a new gambling task with explicit rules: associations with executive functions. *Neuropsychology*, *19*, 267-277. doi:10.1037/0894-4105.19.3.267

Brand, M., Labudda, K., & Markowitsch, H. J. (2006). Neuropsychological correlates of decision-making in ambiguous and risky situations. *Neural Networks*, *19*, 1266-1276. doi:10.1016/j.neunet.2006.03.001

Brand, M., Laier, C., Pawlikowski, M., & Markowitsch, H. J. (2009). Decision making with and without feedback: The role of intelligence, strategies, executive functions, and cognitive styles. *Journal of Clinical and Experimental Neuropsychology*, *31*, 984–998. doi: 10.1080/13803390902776860

Brand, M., & Schiebener, J. (2013). Interactions of age and cognitive functions in predicting decision making under risky conditions over the life span. *Journal of Clinical and Experimental Neuropsychology*, *35*(1), 9–23. doi:10.1080/13803395.2012.740000

Cacioppo, J.T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*, 197–253.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*,*22*(3), 265-289.

Chiesi, F., Donati, M. A., Galli, S., & Primi, C. (2013). The suitability of the SOGS-RA as screening tool: Item response theory-based evidence. *Psychology of Addictive Behaviors*, *27*(1), 287–293. http://dx.doi.org/10.1037/a0029987.

Chiesi, F. & Primi, C. (2014). The Interplay Among Knowledge, Cognitive Abilities and Thinking Styles in Probabilistic Reasoning: A Test of a Model. In E. J. Chernoff & B. Sriraman (Eds). *Probabilistic Thinking: presenting Plural Perspective - Advances in Mathematics Education*, pp. 195-214 Springer Science + Business Media, Dordrecht, Netherlands

Chiesi, F., Primi, C., & Morsanyi, K. (2011). Developmental changes in probabilistic reasoning: The role of cognitive capacity, instructions, thinking styles and relevant knowledge. *Thinking and Reasoning*, *17*, 315–350. doi:10.1080/13546783.2011.59840

Chiesi, F., Ciancaleoni, M., Galli, S., & Primi, C. (2012). Using the Advanced Progressive Matrices (Set I) to assess fluid ability in a short time frame: An item response theory–based analysis. *Psychological assessment*, *24*(4), 892.

Clinton, V., Morsanyi, K., Alibali, M.W., & Nathan, M.J. (2016). Learning about probability from text and tables: Do color coding and labeling through an interactive-user interface help? *Applied Cognitive Psychology*, *30*, 440-453.

Colasante, E., Gori, M., Bastiani, L., Scalese, M., Siciliano, V., & Molinaro, S. (2013). Italian adolescent gambling behaviour: Psychometric evaluation of the South Oaks gambling screen-revised for adolescents (SOGS-RA) among a sample of Italian students. *Journal of Gambling Studies*. http://dx.doi.org/10.1007/s10899-013-9385-6.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, *7*(1), 25.

Cosmides, L. J., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1–73.

Delfabbro, P., Lahn, J., & Grabosky, P. (2006). It's not what you know, but how you use it: statistical knowledge and adolescent problem gambling. *Journal of Gambling Studies*, *22*, 179–193. http://dx.doi.org/10.1007/s10899-006-9009-5.

Delfabbro, P., Lambos, C., King, D., & Puglies, S. (2009). Knowledge and beliefs about gambling in Australian secondary school students and their implications for education strategies. *Journal of Gambling Studies, 25*, 523–539. http://dx.doi.org/10.1007/s10899-009-9141-0.

Derevensky, J. L., Gupta, R., & Baboushkin, H. R. (2007). Underlying cognitions in children's behavior: can they modified? *International Gambling Studies, 7*(3), 281–298. http://dx.doi.org/10.1080/14459790701601448.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking & Reasoning*, *11*(4), 349–381. doi:10.1080/13546780442000222

Donati, M. A., Chiesi, F., & Primi, C. (2013). A model to explain at risk/problem gambling among male and female adolescents: Gender similarities and differences. *Journal of Adolescence, 36*, 129–137. doi:10.1016/j.adolescence.2012.10.001

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*, 803-818. doi: 10.1007/s10899-013-9398-1

Edgren, R., Castrén, S., Mäkelä, M., Pörtfors, P., Alho, H., & Salonen, A. H. (2016). Reliability of Instruments Measuring At-Risk and Problem Gambling Among Young Individuals: A Systematic Review Covering Years 2009–2015.*Journal of Adolescent Health*, *58*(6), 600-615.

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

Evans, J. S. B., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, *77*(3), 197-213.

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*, 672–680.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*, 253–292. doi:10.1016/0010-0285(86)90001-0

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120*, 34–45. doi:10.1037/0096-3445.120.1.34

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 25–42.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1-25.

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.

Gauffroy, C., & Barrouillet, P. (2009). Heuristic and analytic processes in mental models for conditionals: An integrative developmental theory. *Developmental Review, 29*, 249–282.

Gauvrit, N., & Morsanyi, K. (2014). The Equiprobability Bias from a Mathematical and Psychological Perspective. *Advances in Cognitive Psychology*, *10*(4), 119.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704.

Gigerenzer, G., Hoffrage, U., & Goldstein, D. G. (2008). Fast and frugal heuristics are plausible models of cognition: Reply to Dougherty, Franco-Watkins, and Thomas (2008).

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, *8*(2), 53-96.

Gigerenzer, G., Todd, P.M., & the ABC Research Group. (1999). Simple heuristics that make us smart. New York: Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, UK: Cambridge University Press.

Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: why three heads are better than four. *Psychological review*, *116*(2), 454.

Halpern-Felsher, B. L., & Cauffman, E. (2001). Costs and benefits of a decision: Decision-making competence in adolescents and adults. *Journal of Applied Developmental Psychology*, *22*(3), 257-273.

Hambleton, R. K., & de Jong, J. H. (2003). Advances in translating and adapting educational and psychological tests. *Language testing*, *20*(2), 127-134.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Measurement methods for the social sciences series, Vol. 2).

Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2014). Improving critical thinking: Effects of dispositions and instructions on economics students' reasoning skills. *Learning and Instruction, 29*, 31–42.

Hemmings, B., Grootenboer, P., & Kay, R. (2011). Predicting mathematics achievement: The influence of prior achievement and attitudes. *International Journal of Science and Mathematics Education*, *9*(3), 691-705.

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: a model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1191.

Hibbard, J. H., Slovic, P., Peters, E., Finucane, M. L., & Tusler, M. (2001). Is the informed-choice policy approach appropriate for Medicare beneficiaries?.*Health Affairs*, *20*(3), 199-203.

Hicks, K. L., Harrison, T. L., & Engle, R. W. (2015). Wonderlic, working memory capacity, and fluid intelligence. *Intelligence,* 186-195. http://dx.doi.org/10.1016/j.intell.2015.03.005

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.

Huizenga, H. M., Crone, E. A., & Jansen, B. J. (2007). Decision-making in healthy children, adolescents and adults explained by the use of increasingly complex proportional reasoning rules. *Developmental Science*, *10*(6), 814–825. doi:10.1111/j.1467-7687.2007.00621.x

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), Heuristics and biases (pp. 49–81). New York, NY: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*(3), 430-454.

Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decisionmaking. *Child Development*, *72*(3), 844–861. doi:10.1111/1467-8624.00319

Klaczynski, P. A. (2004). A dual-process model of adolescent development: Implications for decision making, reasoning, and identity. In R. V. Kail (Ed.), *Advances in child development and behavior* (pp. 73–123). San Diego, CA: Academic Press.

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

Kokis, J. V., MacPherson, R., Toplak, M. E., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, *83*, 26–52.

Leone, C., & Dalton, C. (1988). Some effects of the need for cognition on course grades. *Perceptual and Motor Skills, 67*, 175–178.

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.

Myers, R. (1990). *Classical and modern regression with applications.* (2nd ed.). Boston, MA: Duxbury.

Morsanyi, K., Handley, S. J., & Serpell, S. (2012). Making heads or tails of probability: An experiment with random generators. *British Journal of Educational Psychology, 83*(3), 379–395. doi:10.1111/j.2044-8279.2012.02067.x

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. doi: 10.1016/j.cedpsych.2009.05.001

Neace, W. P., Michaud, S., Bolling, L., Deer, K., & Zecevic, L. (2008). Frequency formats, probability formats, or problem structure? A test of the nested-sets hypothesis in an extensional reasoning task. *Judgment and Decision Making*, *3*(2), 140.

Panno, A., Lauriola, M., & Figner, B. (2013). Emotion regulation and risk taking: Predicting risky choice in deliberative decision making. *Cognition & Emotion*, *27*, 326–334. doi:10.1080/02699931.2012.707642

Panno, A., Pierro, & Lauriola, M. (2013). Self-regulation predicts risk taking through people's time horizon. *International Journal of Psychology*, 49(3), 211-215. doi:10.1002/ijop.12026

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring global differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*, 295-312.Politi, M. C., Han, P. K., & Col, N. F. (2007). Communicating the uncertainty of harms and benefits of medical interventions. *Medical Decision Making*, *27*, 681-695.

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., Chiesi, F., Donati, M.A. & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). Journal of Behavioral Decision Making. (in press) DOI: 10.1002/bdm.1883

Raven, J. C. (1941). Standardization of progressive matrices. *British Journal of Medical Psychology, 19*, 137–150. doi:10.1111/j.2044-8341.1941.tb00316.x

Raven, J. C. (1962). Advanced progressive matrices. London: Lewis & Co. Ltd..

Reyna, V.F. & Brainerd, C.J. (2007). The importance of mathematics in health and human judgment: numeracy, risk communication, and medical decision making. *Learning and Individual Differences. 17*, 147–59.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education, 10*(3), 6-13.

Sadowski, C. J., & Gulgoz, S. (1996). Elaborative processing mediates the relationship between the need for cognition and academic performance. *Journal of Psychology, 130*, 303–308.

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. doi:10.1007/s11065-015-9285-x

Schiebener, J., Zamarian, L., Delazer, M., & Brand, M. (2011). Executive functions, categorization of probabilities, and learning from feedback: What does really matter for decision making under explicit risk conditions? *Journal of Clinical and Experimental Neuropsychology*, *33*(9), 1025-1039. doi:10.1080/ 13803395.2011.595702

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes, 91,* 296–309.

Snyder, T. D., & Dillow, S. A. (2012). *Digest of education statistics 2011*. National Center for Education Statistics.

Stanovich, K. E. (2004). Metarepresentation and the great cognitive divide: A commentary on Henriques'"Psychology Defined". *Journal of clinical psychology*, *60*(12), 1263-1266.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Personality Processes and Individual Differences, 94*(4), 672–695, doi:10.1037/0022-3514.94.4.672

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*(4), 331-352.

Tapia, M., & Marsh, G. E. (2004). An instrument to measure mathematics attitudes. *Academic Exchange Quarterly*, *8*(2), 16-22.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *The American Economic Review*, *93*(2), 175-179.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics and biases tasks. *Memory & Cognition, 39*, 1275–1289.

Turner, N. E., Zangeneh, M., & Littman-Sharp, N. (2006). The experience of gambling and its role in problem gambling. *International Gambling Studies, 6*, 237–266.

Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician, 57*(2), 74-79.

van Duijvenvoorde, A. C. K., Jansen, B. R. J., Visser, I., & Huizenga, H. (2010). Affective and cognitive decision-making in adolescents. *Developmental Neuropsychology*, *35*, 539–554. doi:10.1080/87565641.2010.494749

van Leijenhorst, L., Westenberg, P. M., & Crone, E. A. (2008). A developmental study of risky decisions on the Cake Gambling Task: Age and gender analyses of probability estimation and reward evaluation. *Developmental Neuropsychology*, *33*(2), 179–196. doi:10.1080/87565640701884287

Winters, K. C., Stinchfield, R. D., & Fulkerson, J. (1993). Toward the development of an adolescent gambling problem severity scale. *Journal of Gambling Studies*, *9*, 63–84.

Woloshin, S., Schwartz, L. M., & Welch, H. G. (2004). The value of benefit data in direct-to-consumer drug ads. *Health Affairs*, W4.

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

Appendix A

	Format	Bias	*I° step*	*II° step*	PRS
Proportion	Frequency		✓		Item 2
	Percentage		✓		Item 3
Simple probability	Frequency		✓		Item 1
	Frequency in Table			✓	Item 14
	Frequency	Equiprobability Bias	✓		Item 7
	Frequency	Equiprobability Bias (Compound event)	✓		
Probability of a sequence of independent events	Frequency	Random Similarity Bias	✓		Item 4
Comparison of simple probabilities	Frequency	Ratio bias with unequal proportions	✓		Item 5
	Frequency	Ratio bias with equal proportions	✓		Item 8
Probability of independent events	Frequency	Gambler's fallacy with equiprobable events	✓		Item 6
	Frequency	Gambler's fallacy with not equiprobable events)	✓		Item 9
Disjunctive probability	Percentage		✓		Item 10
	Frequency			✓	Item 12
Conjunct probability	Percentage			✓	Item 11
	Frequency			✓	Item 13
	Frequency in Table			✓	Item 15
Conditional probability	Frequency in Table			✓	Item 16
Law of small numbers	Frequency		✓		

Appendix B

1. 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

2. The proportion of left-handed people is 4 out of 100. In a school there are 300 students. How many left-handed students are there?

a. 20
b. 12
c. 25

3. Smokers are about 35% of the population. At the airport in the waiting room there are 200 passengers. How many smokers are there?

a. 70
b. 35
c. 20

4. A fair coin is tossed nine times. Which of the following sequence of outcomes is a more likely result of nine flips of the fair coin? (H: Head, T: Tails)

a. THHTHTTHH
b. HTHTHTHTH
c. Both sequences are equally likely

5. Two containers, labeled A and B, are filled with red and yellow tokens in the following quantities. Container A contains 100 tokens, 65 red and 35 yellow. Container B contains 10 tokens, 6 red and 4 yellow. Each container is shaken vigorously. After choosing one of the containers, you must draw a token (without peeking, of course). Which container gives you a better chance of drawing a yellow token?

a. Container A (with 65 red and 35 yellow)
b. Container B (with 6 red and 4 yellow)
c. Equal chances from each container

6. A marble bag contains 15 blue and 15 green marbles. After you drew 5 marbles (the marble drawn was always put back into the bag), a sequence of 5 green marbles was obtained. What is the most likely outcome if a marble is drawn a sixth time?

a. a green marble
b. a blue marble
c. blue and green are equally likely

7. A bingo game is played with 25 numbers (from 1 to 25). At the first draw, which of the following results is the most likely?

a. It is more likely to be an even number
b. It is more likely to be an odd number
c. It is just as likely to be an even or an odd number

8. Two decks, labeled A and B, are composed of cards with a star (star cards) and cards without any figure (white cards) on the reverse side. Deck A contains 100 cards, 80 white cards and 20 with a star. Deck B contains 10 cards, 8 white cards and 2 with a star. After choosing one of the decks, you must draw a card (without peeking, of course). Which deck gives you a better chance of drawing a star card?

a. Deck A (with 80 white and 20 star cards)
b. Deck B (with 8 white and 2 star cards)
c. Equal chances from each deck

9. A marble bag contains 10 blue and 20 green marbles. After you drew 5 marbles (the marble drawn was always put back into the bag), a sequence of 5 green marbles was obtained. What is the most likely outcome if a marble is drawn a sixth time?

a. a green marble
b. a blue marble
c. blue and green are equally likely

10. 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%

11. According to a recent survey, 90% of the population in a city usually lie and 30% of those usually lie about important matters. If we pick a person at random from this city, what is the probability that the person usually lies about important matters?

a) 60%
b) 30%
c) 27%

12. In a choir there are 100 children: 30 boys and 70 girls. Half of the boys and 1 in 10 girls learn to play the piano. We select a child from the choir at random. What is the probability that he/she plays the piano?

a) 22 out of 100
b) 30 out of 100
c) 50 out of 100

13. A village has 1000 inhabitants. 600 people own a pet, and amongst pet owners 1 in 3 own more than 1 pet. If we select one person from this village at random, what is the probability that they own more than one pet?

a) 333 out of 1000
b) 500 out of 1000
c) 200 out of 1000

In a medical center a group of people were interviewed with the following results:

	55 years-old or younger	Over 55	Total
Previous heart attack	29	75	104
No previous heart attack	401	275	676
Total	430	350	780

Suppose we select a person from this group at random. Based on the table:

14. What is the probability that the person had a heart attack?

a) 104 out of 780
b) 104 out of 676
c) 390 out of 780

15. What is the probability that the person had a heart attack and, at the same time is older than 55?

a) 104 out of 350
b) 75 out of 350
c) 75 out of 780

16. When the person had a heart attack, what is the probability that they are over 55?

a) 75 out of 780
b) 75 out of 104
c) 104 out of 350

Scoring (correct answers):

1. c
2. b
3. a
4. c
5. b
6. c
7. b
8. c
9. a
10. a
11. c
12. a
13. c
14. a
15. c
16. b

Table 1. S*tandardized factor loadings, fit statistics, and parameters for each item of the Probabilistic Reasoning Scale (PRS) based on Study 1.*

Item	λ	*S-*χ^2	*b (SE)*	*a (SE)*	*Mean*
1	.52	.09	-1.58 (.18)	1.03 (.15)	.80
2	.81	.71	-2.39 (.23)	2.37 (.48)	.97
3	.68	.31	-1.84 (.17)	1.56 (.23)	.89
4	.35	.82	- 2.97 (.57)	0.64 (.14)	.85
5	.42	.17	-1.49 (.21)	0.79 (.12)	.74
6	.49	.47	-2.45 (.34)	0.95 (.17)	.88
7	.57	.44	-1.27 (.13)	1.19 (.16)	.77
8	.67	.74	-1.56 (.14)	1.52 (.21)	.85
9	.42	.32	-2.04 (.29)	.80 (.13)	.81
10	.50	.02	- 0.07 (.08)	.99 (.12)	.51
11	.50	.02	-0.34 (.09)	.98 (.12)	.57
12	.52	.30	-2.23 (.28)	1.03(.17)	.87
13	.55	.71	-1.66 (.18)	1.13 (.16)	.82
14	.59	.19	-1.79 (.19)	1.25 (.18)	.86
15	.39	.06	-0.04 (.11)	.72 (.11)	.51
16	.34	.51	- 0.69 (.16)	.61(.10)	.60

Note. Standardized factor loadings λ are all significant at $p = .001$. The parameters were computed under the 2PL model (*a* = discrimination, *b* = difficulty).

Table 2. *Descriptive statistics for the measures, and correlations between Probabilistic Reasoning and Mathematics Skills, Math Anxiety, Attitudes Toward Mathematics, and statistics achievement.*

	1	2	3	4	5
1. *Probabilistic Reasoning*	--				
2. *Mathematics Skills*	.59*** (N=317)	--			
3. *Math Anxiety*	-.15* (N=317)	-.18** (N=317)	--		
4. *Attitudes Toward Mathematics*	.22*** (N=317)	.27*** (N=317)	.-.53*** (N=317)	--	
5. *Statistics Achievement*	.32*** (N= 129)	.31*** (N= 129)	.01 (N= 129)	.13 (N= 129)	--
M (*SD*)	12.73 *(2.59)*	22.80 *(4.56)*	24.39 *(6.55)*	124.10 *(25.42)*	24.30 *(3.88)*

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

Table 3. *Multiple regression with statistics achievement as a dependent variable and Probabilistic Reasoning and Mathematics Skills, Math Anxiety, and Attitudes Toward Mathematics, as independent variables.*

Predictors	*B*	*β*	*t*	*p*
Probabilistic reasoning	..41	.26	2.40	.018
Mathematics skills	.18	.17	1.57	.120
Math anxiety	.03	.04	.39	.693
Attitudes toward math	.01	.06	.51	.614

$F(4,113)=5.35$, $p=.001$; $R=.40$; $R^2=.16$

Table 4. *Descriptive statistics for the measures, and correlations between Probabilistic Reasoning and Problem Gambling Severity, Fluid Intelligence and Risky decision making (advantageous choices)*

	1	2	3	4
1. *Probabilistic Reasoning*	--			
2. *Problem Gambling Severity*	-.16* (N=505)	--		
3. *Fluid Intelligence*	.38*** (N=505)	.21*** (N=352)	--	
4. *Risky decision making (advantageous choices)*	30*** (N=352)	.01 (N=352)	. 21*** (N=352)	--
M	11.35	.949	10.07	7.92
(SD)	*(2.91)*	*(1.54)*	*(1.84)*	*(9.34)*

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

Table 5. *Multiple regression with advantageous decisions as a dependent variable and Probabilistic Reasoning, Problem Gambling Severity and Fluid Intelligence as independent variables.*

Predictors	*B*	*β*	*t*	*p*
Probabilistic Reasoning	.99	.28	3.69	.001
Problem Gambling Severity	.33	.05	.77	.450
Fluid Intelligence	.12	.03	.34	.732

$F(3,194)=5.96$, $p=.001$; $R=.29$; $R^2=.08$

Table 6. Descriptive statistics, and correlations between Probabilistic Reasoning Skills and the other measures

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. PRS	--								
2. Math Anx.	-.13**	--							
3. Subj. Num.	.13**	-.33**	--						
4. NFC	.12**	-.10*	.31**	--					
5. Cond. Prob.	.24**	-.21**	.11**	.12**	--				
6. CRT	.27**	-.41**	.39**	.14**	.34**	--			
7. CRT-Long	.30**	-.43**	.42**	.14**	.35**	.93**	--		
8. Attendance	-.04	.05	.07	.05	.05	.06	.09	--	
9. Exam result	.26**	-.01	-.06	.05	.16	.17	.22*	.29**	--
M	13.23	13.96	34.74	61.47	2.94	1.59	3.79	2.45	76.58
(SD)	*(4.01)*	*(5.98)*	*(8.27)*	*(13.58)*	*(1.40)*	*(1.16)*	*(1.74)*	*(.81)*	*(16.48)*

AMAS= *Abbreviated Math Anxiety Scale*, Subj. Num.= *Subjective Numeracy Scale*, NFC=*Need for Cognition*, Cond. Prob. =*Conditional Probability Reasoning;* CRT=*Cognitive Reflection Test*

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

Table 7. *Multiple regression analysis with statistics exam performance as a dependent variable and the PRS, the CRT-Long and attendance at statistics lectures as independent variables.*

Predictors	*B*	*β*	*t*	*p*
PRS	.57	.26	2.54	.012
CRT-Long	.15	.06	.53	.600
Lecture attendance	2.29	.28	3.10	.003
Subjective Numeracy	-.055	-.12	-1.08	.282
Need for Cognition	.04	.10	.89	.377
AMAS	.04	.05	.48	.634
Conditional Prob.	.04	.01	.12	.902

AMAS=Abbreviated Math Anxiety Scale; Conditional Prob.= Conditional Probability Reasoning

$F(7,104107)=2.76$, $p=.012$; $R=.41$; $R^2=.17$

Figure 1. Test Information Function of the *Probabilistic Reasoning Scale* (PRS)