Eurasia Journal of Mathematics, Science & Technology Education, 2016, **12**(8), 2013-2038
doi: 10.12973/eurasia.2016.1265a

Assessing Probabilistic Reasoning in Verbal-Numerical and Graphical-Pictorial Formats: An Evaluation of the Psychometric Properties of an Instrument

Mirian Agus & Maria Pietronilla Penna
University of Cagliari, ITALY
Maribel Peró-Cebollero & Joan Guàrdia-Olmos
University of Barcelona, SPAIN

•*Received 16 July 2015*•*Revised 5 September 2015*•*Accepted 18 October 2015*

Research on the graphical facilitation of probabilistic reasoning has been characterised by the effort expended to identify valid assessment tools. The authors developed an assessment instrument to compare reasoning performances when problems were presented in verbal-numerical and graphical-pictorial formats. A sample of undergraduate psychology students (n=676) who had not developed statistical skills, solved problems requiring probabilistic reasoning. They attended universities in Spain (n=127; f=71.7%) and Italy (n=549; f=72.9%). In Italy 173 undergraduates solved these problems under time pressure. The remaining students solved the problems without time limits. Classical Test Theory (CTT) and Item Response Theory (IRT) were applied to assess the effect of two formats and to evaluate criterion and discriminant validity. The instrument showed acceptable psychometric properties, providing preliminary evidence of validity.

Keywords: probabilistic reasoning; format of problem presentation; verbal-numerical format; graphical-pictorial format; educational assessment; validity

INTRODUCTION

Probabilistic reasoning is a crucial aspect when using mathematics and statistics

Correspondence: Mirian Agus,
Department of Pedagogy, Psychology, Philosophy, Faculty of Humanistic Studies, University of Cagliari, Italy
E-mail: mirian.agus@unica.it

(e.g., Brase, 2009; Brase & Hill, 2015; Franklin et al., 2007; Mandel, 2014; Tubau, 2008). This reasoning is omnipresent in our society, supporting judgements about uncertainty in daily and scholastic contexts (Gigerenzer, 2008; Jones, 2006; Sharps,
Hess, Price-Sharps, & Teh, 2008). As Batanero and Sánchez (2006, p.241) stated, "(…) in high school students are expected to determine the likelihood of an event by constructing probability distributions for simple sample spaces, compute and interpret the expected value of random variables in simple cases (…), to identify mutually exclusive and joint events, understand conditional probability and independence, and draw on their knowledge of combinations, permutations, and counting principles to compute these different probabilities". In the "Guidelines for Assessment and Instruction in Statistics Education" (GAISE), among other indications, there are "understand and apply basic concepts of probability" (Franklin et al., 2007, p.5). It is a very specific type of thinking, which does not improve at a specific age; indeed, it is reliant on education, and individuals are expected to reach a satisfactory performance after an appropriate training (Agnoli & Krantz, 1989; Mandel, 2015).

The study of probabilistic reasoning applied to mathematical and statistical problems has interested many researchers, specifically in order to overcome the great troubles encountered by the students when trying to solve probabilistic problems. Some items about the probabilistic and Bayesian reasoning have been considered in literature as a useful type of tasks in order to evaluate the features of some kind of statistical reasoning (Brase, 2009; Brase & Hill, 2015; 2013; Mandel, 2014).

DelMas (2004) analysed in a deep way the features of statistical and mathematical reasoning, in order to shed light on the main characteristics of human reasoning. He stated that probability theory is an essential part of mathematics and statistics; moreover, he highlighted the differences between statistical and mathematical reasoning, mostly as concerns the key and strong role of the context of the problem in the application of statistical reasoning (DelMas, 2004; Franklin et al., 2007). Franklin et al. (2007) examined in detail the different approach to the solution of probabilistic problems in statistics and in mathematics. They highlighted that the probability problem solving in statistics (for example in the classical coin problem) generally starts from an unfamiliar situation. Often the solution is reached in an experimental way, deriving from specific practices (it is context related). Instead, in mathematics the solution of a probabilistic problem derives from the application of a known rule (that is, the solution is model determined) (Franklin et al., 2007). In this regard, it is useful to refer to the well-known Garfield and Ben-Zvi classification (Ben-Zvi & Garfield, 2004; Garfield & Ben-Zvi, 2008), which discriminated between *statistical literacy,*

State of the literature

- Many authors highlighted the difficulties encountered by undergraduates in probabilistic reasoning. The study of the features of this reasoning has been considered very useful in order to find methods to support the performances in these problems.
- The administration of the problems in specific formats (i.e., frequency formats, in a natural sampling framework, using graphical-pictorial representations) has been appraised as a useful way to improve this performance.
- Nevertheless, there are few suitable measures valuable to assess the performance in these problems, comparing different formats of presentation, specifically in the same inexperienced student.

Contribution of this paper to the literature

- This paper appraises the psychometric characteristics of an instrument settled to evaluate probabilistic reasoning in both verbal-numerical and graphical-pictorial formats.
- We focused on simple and conditional probabilities expressed as frequencies, consistent with classic studies in the literature.
- By applying the Classical Test Theory and the Item Response Theory, we evaluated the features of our items, the validity of criterion and discriminant. We assessed the two formats in the three samples of inexperienced psychology undergraduates in Italy and Spain, in presence versus absence of time pressure. We provided preliminary evidence of validity.

statistical reasoning and statistical thinking. Statistical literacy (Ben-Zvi & Garfield, 2004; Gal, 2002) derives from individual's education and consists of the understanding of the basic use of statistical language and simple statistical tools (such as the mean), discriminating among dissimilar data representations (Chance, 2002; Rumsey, 2002). *Statistical reasoning* is the process applied in the reflection on statistical data, in order to give meaning to aspects related to them. It may implicate the association of one concept with another, or the combination among dissimilar concepts inferred from the data and probability considerations (Garfield & Ben-Zvi, 2008; Garfield, 2003). The statistical reasoning presents specified relationships with probabilistic reasoning (Ben-Zvi & Garfield, 2004; Garfield & Ben-Zvi, 2008). Indeed, six main classes of statistical reasoning were identified (i.e., on data, on representations of data, on associations of variables, on statistical measures, on samples, and on uncertainty) (Garfield, 2003). Within these classes, the reasoning on uncertainty is extensively used in statistical and mathematical education. This kind of reasoning denotes the ability to infer and use notions of casualness, chance and probability. *Statistical thinking* denotes specifically the statistician's type of thought (Wild & Pfannkuch, 1999). It connotes detailed knowledge about how and why to apply a method, a measure, or a statistical model. Thus, it refers to awareness of the theories implied in the statistical process and methods (Chance, 2002). Garfield and Ben-Zvi (2008) clarify that statistical literacy, reasoning and thinking are facets of the same dimension, in which there are multiple intersections. All previous considerations allow understanding that the probabilistic reasoning, which is the object of our investigation, is the most important crossover point between the different forms and abilities of thinking so far mentioned.

These aspects have been often studied in the recent years, in order to define and describe the undergraduates' difficulties when dealing with these topics. Many papers have highlighted the troubles implied in this form of reasoning, especially in the academic context (Díaz & De La Fuente, 2006; Gal, 2005). These difficulties are relevant to all students, especially undergraduates in humanistic faculties, who must cope with courses that contain a strong mathematical component (e.g., statistics) (Chiesi & Primi, 2009; Guàrdia-Olmos et al., 2006). It has been observed that these students often do not possess foundations adequate to take advantage from these courses (Galli, Chiesi & Primi, 2011). Moreover, the students often do not appreciate the utility of these mathematical aspects for their future careers (Guàrdia-Olmos et al., 2006).

Then the teachers and the scholars attempted to devise specific ways or formats of problem presentation in order to overcome these difficulties on the study of probabilistic topics. This topic deserves further attention, in order to understand why and how the subjects are often affected by biases preventing the reach of a correct solution (Pessa & Penna, 2000). For these reasons, a large number of investigations have been performed to assess the effects produced on reasoning by the different formats of the problem presentation (modifying wording, type of numerical data presented, and illustrations) (e.g., Moro, Bodanza & Freidin, 2011; Sirota, Kostovičová, & Vallée-Tourangeau, 2015). In this context, many authors have attempted to develop approaches to overcome these well-known individuals' difficulties in probabilistic reasoning in different situations (García-Retamero, Galesic, & Gigerenzer, 2011; Gigerenzer & Hoffrage, 1995; Johnson, Pierce, Baldwin, Harris, & Brondmo, 1996). We could focus on the widespread use of graphics to support this type of reasoning (Bishop, 2008; Clements, 2014; De Hevia, Vallar, & Girelli, 2008; González, Campos, & Perez, 1997; Johnson & Tubau, 2013; Konold, Higgins, Russell, & Khalil, 2014; Lean & Clements, 1981; Tubau, 2008; Tufte, 2001; Zahner & Corter, 2010).

The use of graphical representation is related to a widely invoked effect in this body of literature: graphical facilitation (e.g., Brase 2009; Brase & Hill, 2015; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002; Moro & Bodanza, 2010). However, some authors do not agree that graphics pay a facilitating role and speak of graphical impediment (Knauff & Johnson-Laird, 2002; Castañeda & Knauff, 2013). The debate about this theme is strong in literature. Indeed, also when it seems recognised the effect of graphical facilitation, it is not clear what are the factors that could affect the application of probabilistic reasoning, enhancing the performance (Brase & Hill, 2015; Mandel, 2014; Moro & Bodanza, 2010).

The different opinions could find a theoretical basis in the old theory of dual coding (Paivio, 1971), which postulates that data can be represented in both verbal and visuo-spatial modalities. Consequently, the specific format of problem presentation could support the selection of an adequate solution strategy based on a suitable high-level problem representation.

Also the more recent Dual-Processes Theory (Evans & Stanovich, 2013) suggests important considerations on facilitation effect, supporting the idea that different problem presentation formats could enhance diverse levels of data processing (Type I – fast and automatic - and Type II –controlled and rule-based). These cognitive aspects could be evaluated in relation to the nested-sets approach (Barbey & Sloman, 2007; Girotto & Gonzalez, 2001; Sloman, Over, Slovak, & Stibel, 2003) that states that the facilitation could derive from the enlightening of relationships among sets. Coherently with this approach, some authors have highlighted the null iconicity effect (Sirota, Kostovičová, & Juanchich, 2014), affirming that the level of iconicity did not affect the facilitation effect, because only the structure of problem could have an impact on the probabilistic reasoning.

On the other hand, the solution processes are also undoubtedly related to many aspects, such as the features of the task (e.g., the presence or absence of time pressure) (e.g., Rieskamp & Hoffrage, 2008) and the individual characteristics (e.g., cognitive and non-cognitive dimensions, such as ability, anxiety and attitudes) (e.g., Galli et al., 2011; Onwuegbuzie & Seaman, 1995). These aspects might strongly affect the student performance, interacting with reasoning and problem solving (e.g., Chiesi, Primi, & Morsanyi, 2011).

Recently, we have observed an increasing attention to the study of probabilistic reasoning and problem presentation format (e.g., Kellen, Chan, & Fang, 2006; 2007; 2013; Sirota et al., 2014). Specifically it is cogent and demanding the requirement to study in more detail the probabilistic reasoning, in order to better understand its features and opportunities for improvement, as a function of different formats of problem presentation.

However, in the existing literature we can find only few specific assessment instruments that are useful for comparing reasoning performance as a function of problem format, e.g., verbal-numerical compared to graphical-pictorial formats (e.g., Chan & Ismail, 2014; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). Some previous studies on this topic often evaluated the effect of graphical facilitation by comparing single pairs of problems, which are frequently presented to different individuals (e.g., Moro & Bodanza, 2010).

Coherently with these considerations, to overcome these limitations, we conducted two pilot studies (Agus, Peró-Cebollero, Penna, & Guàrdia-Olmos, 2014; Penna, Agus, Peró-Cebollero, Guàrdia-Olmos, & Pessa, 2014), based on an ad hoc assessment instrument, in which the same student must solve problems presented in both verbal-numerical and graphical-pictorial formats. Indeed our work aimed at devising an agile and short assessment instrument of probabilistic reasoning, assessed by items presented in the two cited formats (verbal-numerical and graphical-pictorial). In order to identify our items in both formats, we referred to the classic studies reported in this kind of literature, selecting specific problems

related to simple and conditional probability (e.g., Brase, 2009; Gigerenzer & Hoffrage, 1995; Yamagishi, 2003).

In this paper, we report an assessment of the psychometrics features of this instrument used in a sample of psychology undergraduate students without any statistical expertise.

The data were collected in Italy and Spain, which are both members of the European Higher Education Area (EHEA), to evaluate the reliability of these measures in both countries.

Moreover, we administered the instrument preliminarily to Italian undergraduates to compare the effect of presence and absence of time pressure, reserving the extension of this evaluation among Spanish student if interesting results were identified. In this way, we assessed the validity of the measures under different timing conditions (Maule, Hockey, & Bdzola, 2000).

METHOD

Participants

Our research was conducted on 676 undergraduates in Psychology during the first year of the degree course in Spain (n= 127) and Italy (n= 549) (Table 1). The participants consisted of 91 females in Spain (71.7%) and 400 females in Italy (72.9%). The average age of the Spanish sample was 20.32 years (*SD* = 6.098) and 20.00 years for the Italian sample (*SD*=3.676).

Of the Italian participants, 173 solved problems under a time pressure condition (the remaining students worked without time limits). Our samples included students who voluntarily participated in the study (non-probability sampling) who were selected based on their accessibility (convenience sampling). The data were collected during regular classes by the same trained researcher during the first semester of academic activity (from September 2013 to January 2014). All universities involved in this study were located in metropolitan areas.

We evaluated the undergraduates' previous curricula to identify and exclude students who had already learned statistics from this study.

Table 1. Summary of demographic characteristics

	Spanish sample		**Italian sample**			
	without time pressure		**without time pressure**		**with time pressure**	
Sample size	127		376		173	
Percentage of women	71.70		78.70		60.10	
Mean age	20.32		20.14		19.69	
Standard deviation age	6.09		4.297		1.63	
Age range	17-52		18-62		18-36	
University affiliation by percentage	Barcelona 100%		Cagliari 7.40% Chieti 38.30% Genoa 15.40% Milan 13.30% Naples 9.30% Pavia 5.30% Trieste 10.90%		Rome 100%	
Mean university marks [1]	8.82		78.96		78.35	
Standard deviation of university marks	1.14		10.55		10.98	
Range of university marks	6.33-13		60-100		60-100	
SCALE	***M***	***SD***	***M***	***SD***	***M***	***SD***
PMA VISUO-SPATIAL	23.93	11.81	19.72	10.46	23.59	11.69
PMA NUMERICAL	17.11	6.31	17.20	6.69	18.50	6.17

Note: Access to university is regulated differently in Spain and Italy; moreover, the marks required for admission are different.

Instruments

The demographic characteristics were requested on a specific form. Then, participants completed subsequent sections of the protocol. The first and second sections, which were intended to assess the subjects' abilities, consisted of the Intermediate Form of the Primary Mental Abilities (PMA) test (L.L. Thurstone & T.G. Thurstone, 1981; 1987) appraising the visuo-spatial and numerical dimensions. These scales were administered according to the time limits and indications of the PMA manual.

The following protocol sections assessed probabilistic reasoning in verbal-numerical (N) and graphical-pictorial (G) formats. Through pilot studies (Agus, Peró-Cebollero, Guàrdia-Olmos, & Penna, 2013; Agus et al., 2014; Penna et al., 2014), we created a short questionnaire, including five problems for each format (see Appendix).

To create the problems, we first identified basic aspects of probabilistic reasoning, referring to simple and conditional probability, as stated by many authors in the typical works on this topic (Brase, 2009; Ben-Zvi & Garfield, 2004; Garfield & Ben-Zvi, 2008; Mandel, 2014; Moro & Bodanza, 2010). Specifically our items referred to the core domain of probability introduced by Moore (1990), which identified the basic probabilistic concepts developed, from the early grades of school, for the evaluation of the chance. This author specifically referred to the concepts related to simple and conditional probability, the independence of the data and the concept of random sampling. These aspects constituted in his theorisation the intermediate level of probabilistic reasoning (Langrall & Mooney, 2006; Jones, 2006; Moore, 1990).

The selection of appropriate problems occurred through collaboration with experts on this topic. Subsequent presentation of the problems to samples of inexperienced students allowed us to clarify and select the most appropriate problems (Agus et al., 2014). The development of the instrument was supported also by the application of a qualitative data analysis on the open responses given to the probabilistic problems presented. The solution to the problems required basic mathematical skills learned in high school. We used a numerical format of probabilistic problems, based on representations of frequencies, referring to classic works on this topic. We included the problems on medical diagnoses (e.g., Brase, 2008; 2009; Brase, Cosmides, & Tooby, 1998; Cosmides & Tooby, 1996; Evans, Handley, Perham, Over & Thompson, 2000; Gigerenzer & Hoffrage, 1995; Sloman, Over, Slovak & Stivel, 2003), on decks of cards (e.g., Kahneman & Tversky, 1974), on university examinations (e.g., Girotto & González, 2001), on dices (e.g., Watson & Moritz, 2003), on production defects (e.g., Yamagishi, 2003) (see Appendix). The G format of probabilistic problems included tree diagrams and simple drawings, as indicated and suggested by classical research on probabilistic reasoning (e.g., Brase, 2009; Corter & Zahner, 2007; Hoffrage et al., 2002; Moore, 1990; Zhu & Gigerenzer, 2006).

The performances in these kind of problems have been studied by many authors, which attempted to identify the features of probabilistic reasoning evaluating the subjects' performances in relation to different formats of presentation (e.g., Corter & Zahner, 2007; Evans, Handley, Neilens, & Over, 2010; Evans, Handley, Perham, Over, & Thompson, 2000; Hoffrage et al., 2002; Gigerenzer & Hoffrage, 1995; Gilovich, Griffin, & Kahneman, 2002; Girotto & Gonzalez, 2001; 2008; Moro & Bodanza, 2010; Moro, Bodanza, & Freidin, 2011; Sloman et al., 2003; Tversky & Kahneman, 1983; Yamagishi, 2003).

For each problem, we provided a short explanation and four response options (of which only one was correct). Subsequently, the students were asked to describe

their reasoning in an open-ended question. We analysed the results of problem solving by summing the number of correct responses for both N and G scales.

This test was part of a larger research project, and during the same work session, the students filled out other questionnaires assessing statistical anxiety and attitudes towards statistics (not used in this investigation).

Each student completed the protocol in his/her native language. We administered all paper-and-pencil questionnaires to large groups in a quiet lecture room.

Procedure

The performance of each participant was assessed for both problem formats to control for variability due to individual characteristics. The problems were presented in different format orders (e.g., NG, first N then G format, versus GN, first G then N format) and problem sequences (sequences 1 and 2, in which the latter was the reverse of the former sequence), ensuring that each problem was not always presented in the same position.

To investigate the potential effect of timing administration, we first selected a sample of Italian undergraduates (n=173) to who these problems were presented under time pressure. The time limit was 30' to solve 10 problems in the N and G formats and to complete the other two sections of the protocol (all outstanding undergraduates solved all problems before the time limit).

We obtained different samples, which were studied separately (the Spanish sample without time pressure and Italian students both without and with time pressure).

Analysis approach

We evaluated descriptive statistics to determine the number of correct responses. From the open-ended responses provided for each problem, we identified the students who obtained the correct answer randomly, which we coded as incorrect. We observed that many missing values were obtained for problems that were perceived as very difficult (as clarified by the open-ended responses). Consequently, we consider these missing values as the inability to provide the correct response.

Then, our analysis proceeded in several steps.

Initially, we evaluated the potential effects of format order and sequence of problem presentation through a one-way ANOVA of the scores by format. Variables identifying the four presentations (NG1, NG2, GN1, GN2) were used as factors. Because some variables did not meet the assumptions of the ANOVA, we applied the Kruskal-Wallis Test for some comparisons. We did not find significant effects of

Table 2. One-way Analysis Of Variance and Kruskal-Wallis Test for order and sequence of problem presentation in verbal-numerical and graphical-pictorial formats

		Levene			*F*			*Kruskal-Wallis*			NG1		NG2		GN1		GN2		Total	
sample	**format**	***Levene***	***df1; df2***	***p***	***F***	$df_{b;w}$	***p***	***KW***	***df***	***p***	***m***	***sd***	***m***	***sd***	***m***	***sd***	***m***	***sd***	***m***	***sd***
Spanish without time pressure	**N**	.60	3;123	.61	.72	3;123	.53	/			2.17	1.50	1.66	1.32	2.10	1.59	2.00	1.47	2.00	1.48
	G	4.01	3;123	.01	/			5.51	3	.13	3.00	1.15	2.73	1.61	2.36	1.30	2.36	1.22	2.62	1.34
Italian without time pressure	**N**	.29	3;372	.82	.01	3;372	.99	/			1.72	1.33	1.76	1.47	1.73	1.44	1.75	1.53	1.74	1.44
	G	1.13	3;372	.33	.95	3;372	.41	/			2.59	1.45	2.68	1.41	2.45	1.33	2.37	1.28	2.53	1.37
Italian with time pressure	**N**	1.50	3;169	.21	1.32	3;169	.26	/			2.58	1.27	2.22	1.34	1.95	1.58	2.33	1.63	2.27	2.57
	G	3.28	3;169	.02	/			6.74	3	.08	2.76	1.64	2.09	1.50	2.58	1.30	2.84	1.16	1.47	1.43

order or sequence (p>.05) (Table 2). Then, we conducted analyses for all problems jointly.

At this point, our work followed two main approaches, Classical Test Theory (CTT) and Item Response Theory (IRT). The use of both methods is helpful and has been identified in previous research as a respected method of developing assessment instruments (e.g., Frey & Seitz, 2009; Hays, Brown, Brown, Spritzer, & Crall, 2006; Kunina-Habenicht, Rupp, & Wilhelm, 2009; Pollard, Dixon, Dieppe, & Johnston, 2009). Specifically, the application of IRT overcomes some limits of CTT methods, which are essentially based on correlational data and processes that capitalise on Cronbach's alpha (Reeve & Fayers, 2005). Moreover, CTT is highly dependent on subject abilities, accessing only a small part of the underlying dimensions (Singh, 2004). The application of IRT offers supplementary evidence and information with respect to CTT, being the parameters that are independent from the sample of items to which the subjects respond and furnishing additional suggestions to disregard non-discriminating or redundant items (Embretson & Reise, 2000; Hartig & Höhler, 2009).

In relation to CTT, we applied Classical Item Analysis and evaluated reliability using Cronbach's Alpha. Furthermore, we explored construct validity using Confirmatory Factor Analysis (CFA). We also evaluated discriminant and concurrent validity. Specifically, to appraise discriminant validity, we considered the type of high school attended (in the Italian sample, we distinguish between students with "strong" and "weak" mathematical educations). To assess concurrent validity, we correlated our measures of probabilistic reasoning in both formats with the visuo-spatial and numerical PMA values and admission marks.

In the succeeding phase of the analysis, we applied Item Response Theory (IRT) to determine the difficulty and discrimination of problems in both formats. This method overcomes some limits of CTT related to sample dependency in the evaluation of item features (Fan, 1998).

The analyses were conducted using R 3.1.3, EQS 6.1 (Bentler, 1995) and IRTPRO 2.1 software (Cai, Thissen, & du Toit, 2011).

RESULTS

The first phase - CTT

We evaluated the scale's alpha reliability, the difficulty and discrimination indices, and the correlation item-total (Table 3).

The difficulty index indicates that most problems had some difficulty in both formats, and only one problem in the G format (B4) was considered easy to solve (Table 3). It is clear that this index "is not determined solely by the content of the items" (Ebel & Frisbie; 1991, p. 228). Indeed, the difficulty also reflects subject abilities to solve specific problems, unlike when IRT models are used (Fan, 1998).

The discrimination index was computed in two ways: first, based on the difference between the proportion of correct answers in the top 50% group and the bottom 50% group, and second, by the correlation between the item-total score excluding that problem (Table 3). The discriminating capacity of the problems was reasonably good (higher than 0.3) in both the N and G formats. The correlations between item-totals are considered acceptable if they are higher than 0.3 (e.g., Schinka & Velicer, 2003). However, we observed some problems with lower values. In particular, we detected a poor index value among all samples for problem A3. Problem A5 produced a poor index value only among the Italian students (with and without time pressure). In the G format, the values are low across samples for B5; moreover, poor values are identified for both Spanish and Italian students without time pressure for B3.

Table 3. Classical Item Analysis for problems in verbal-numerical and graphical-pictorial formats

Spanish sample without time pressure				**Italian sample without time pressure**				**Italian sample with time pressure**			
Verbal-numerical format											
Alpha reliability = .584; St. Alpha =.580				Alpha reliability = .576; St. Alpha =.574				Alpha reliability = .567; St. Alpha =.560			
Item label	Difficulty index	Discrimination index	r(item, total)	Item label	Difficulty index	Discrimination index	r(item, total)	Item label	Difficulty index	Discrimination index	R (item, total)
A1	.465	.738	.325	A1	.359	.648	.347	A1	.538	.842	.445
A2	.472	.833	.461	A2	.402	.728	.374	A2	.584	.719	.389
A3	.220	.380	.232	A3	.242	.504	.294	A3	.266	.385	.141
A4	.409	.690	.342	A4	.394	.728	.401	A4	.491	.719	.387
A5	.433	.666	.347	A5	.346	.520	.254	A5	.399	.649	.276
Graphical-pictorial format											
Alpha reliability = .522; St. Alpha = .509				Alpha reliability = .545; St. Alpha = .546				Alpha reliability = .602; St. Alpha = .602			
Item label	Difficulty index	Discrimination index	r(item, total)	Item label	Difficulty index	Discrimination index	r(item, total)	Item label	Difficulty index	Discrimination index	r(item, total)
B4	.874	.333	.151	B4	.816	.472	.324	B4	.838	.385	.356
B1	.386	.738	.418	B1	.359	.624	.327	B1	.422	.719	.335
B3	.575	.666	.232	B3	.582	.656	.288	B3	.572	.736	.369
B2	.504	.761	.392	B2	.513	.776	.385	B2	.491	.894	.499
B5	.283	.500	.255	B5	.263	.440	.226	B5	.249	.456	.244

Internal consistency (Cronbach's Alpha coefficient) was weak for both formats, with values ranging to .52 to .60. These values refer to dichotomy scores, which not does suggest that these index values are too poor (Ebel & Frisbie, 1991). Moreover, these low values might be related to the small number of problems involved in questionnaire construction (Kline, 2000) (Table 3). Our considerations also relate to the lively debate in the literature on the limitations of the application of Cronbach's Alpha (e.g., Laverdière, Morin, & St-Hilaire, 2013; Pastore, & Lombardi, 2014; Sijtsma, 2009a, 2009b).

Confirmatory factor analysis

We performed a CFA to assess the structure of our problems in each format. All analyses were performed using the EQS 6.1 software (Bentler, 1995). We used the covariance matrices to explore the data in the unifactorial solution (Table 4). We assessed univariate and multivariate normality. In this regard, the variable distribution showed a non-symmetrical curve and a non-normal multivariate trend; therefore, the solutions were estimated using Elliptical Least Squares (ELS). We fixed the factor variance at 1.0 and specified all factor loadings as free to be estimated (Kline, 2000).

On the base of our theoretical framework, we chose to analyse the items on probabilistic reasoning in two formats (N and G) separately, as different arrangements of the items inquiring the same construct; then we applied two CFAs in unifactorial solution, separately for each sample of undergraduates.

Many authors (e.g., Bentler, 1995; Hu & Bentler, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003) recommend using multiple fit indices to evaluate

Table 4. Confirmatory Factor Analyses, Goodness of Fit indices, and factor loadings for items in verbal-numerical and graphical-pictorial formats

sample	indices	Items N format						Items G format			
		A1	A2	A3	A4	A5	B4	B1	B3	B2	B5
Spanish sample without time pressure	Factor loadings	.45	.70	.27	.47	.44	.14	.72	.28	.64	.30
	χ^2			6.95					2.82		
	χ^2 (df)			5					5		
	χ^2 p			.22					.72		
	INDEPENDENCE AIC			63.32					35.15		
	MODEL AIC			-3.04					-7.18		
	RMSEA			.05					.00		
	RMSEA [90% CI]			[0.00, 0.14]					[0.00, 0.09]		
	SRMR			.04					.03		
	CFI			.97					1.00		
Italian sample without time pressure	Factor loadings	.49	.54	.38	.56	.34	.43	.49	.39	.59	.31
	χ^2			13.13					21.91		
	χ^2 (df)			5					5		
	p			.02					.01		
	INDEPENDENCE AIC			194.44					171.61		
	MODEL AIC			3.13					11.91		
	RMSEA			.03					.09		
	RMSEA [90% CI]			[0.02, 0.11]					[0.05, 0.13]		
	SRMR			.03					.04		
	CFI			.96					.90		
Italian sample with time pressure	Factor loadings	.67	.59	.18	.48	.36	.40	.49	.49	.72	.32
	χ^2			6.07					13.83		
	χ^2 (df)			5					5		
	p			.29					.01		
	INDEPENDENCE AIC			92.47					108.04		
	MODEL AIC			3.92					3.83		
	RMSEA			.03					.10		
	RMSEA [90% CI]			[0.00, 0.11]					[0.03, 0.16]		
	SRMR			.03					.05		
	CFI			.99					.92		

Note: AIC=Akaike Information Criterion; RMSEA=Root mean square error of approximation; SRMR=Standardised Root Mean Square Residual; CFI= Comparative Fit index; χ2=Chi-squared test

model fit. For these reasons, we compared the root mean squared error approximation (RMSEA), standardised square root mean residual (SRMR) and comparative fit index (CFI). To compare the models, we also followed Wang and Russell (2005), who suggest using the RMSEA and its confidence interval because this index is not sensitive to the number of indicators and factors, sample size or model complexity (Cheung & Rensvold, 2002). Table 4 includes the fits and statistics of the models. In relation to the RMSEA, SRMR and CFI, Schermelleh-Engel et al. (2003) indicate that values of at least .08 and .10 for RMSEA and SRMR, respectively, and between .95 and .97 for CFI are indicative of acceptable model fit.

Generally, all problems in both formats appeared useful to assess the dimensions of interest. The factor concerning probabilistic reasoning in the N format exhibited good or acceptable fit indices in each sample. For example, the RMSEA was not higher than .056. Additionally, the SRMR index was lower than .046 (Table 4). The factor loadings indicated a unique problematic value for A3 in the sample of Italians under time pressure, unlike the other samples.

The problems in the G format highlighted some problems; indeed, the RMSEA produced values higher than .08 in both Italian samples compared to the Spanish sample. The CFI appeared also problematic for Italians without time pressure. In relation to the factor loading, we observed a low value for the Spanish sample for problem B4 (which produced a higher value for Italians).

Table 5. Pearson's *r* correlations by format, university admission mark, PMA visuo-spatial and numeric scales

Sample		*r*		
	Format	**Admission mark**	**Visuo-Spatial scale**	**Numeric scale**
Spanish without time pressure	**N**	.21*	.19*	.10
	G	.18*	.20*	.03
Italian without time pressure	**N**	.17**	.19**	.14**
	G	.07	.25**	.11*
Italian with time pressure	**N**	.02	.26**	.27**
	G	-.01	.22**	.13

*Note: *p < .05. **p < .01*

Concurrent and discriminant validity

We assessed the criterion-related validity for the evaluation of concurrent validity. We computed the linear relationships (the Pearson's *r*) for performance in the N and G formats, numerical and visuo-spatial scales, and university admission marks. The choice of these dimensions was related to previous research indicating the existence of relationships between performance in probabilistic reasoning and these dimensions (e.g., Chiesi et al., 2011; Furlan & Agnoli, 2010; Guàrdia-Olmos et al., 2006; Tubau, 2008).

In both formats, probabilistic reasoning exhibited weak significant direct correlations with the PMA scales, and university admission marks were not correlated with probabilistic reasoning (Table 5). Moreover, the interpretation of these correlations required specific attention due to the content of evaluated reasoning. Low values of *r* could be ascribed to a mismatch among our dimensions and other instruments (Kubiszyn & Borich, 1990). Furthermore, the weak linear bivariate associations, found in our data among these external constructs and the probabilistic reasoning, may perhaps be related also to the effect of interacting dimensions (attitudes and anxiety, for example). The complex effects of other variables on reasoning have been identified, according to the literature (e.g., Chiesi & Primi, 2009; Lalonde & Gardner, 1993), in other studies conducted by the authors (Agus, Peró-Cebollero, Penna, & Guàrdia-Olmos, 2015). As result, the relationships among abilities and performances in probabilistic reasoning only partially may be described by the bivariate linear correlation. Consequently, we are confident that the validity of our items relies on the reference to the broad literature on the topic, and furthermore on qualitative analysis of open responses applied in our previous pilot studies (Agus et al., 2014). Therefore, as indicated by Kubiszyn and Borich (1990, p. 355), we considered the evidence of content validity to be "*most important*".

To assess the discriminant validity, we distinguished between students who attended high schools with a mathematical orientations and humanistic orientations. Then, we randomly extracted two groups of similar size from the Italian sample working without time pressure. We observed a significant effect. Namely, students whose math curricula were stronger obtained higher scores, compared to their colleagues who attended high schools that emphasised the humanities in both the N (t= 2.228, df=294; p=.027; *Partial* η^2= .017; m= 1.618, sd=1.413 versus m= 1.993, sd=1.479) and G (t=3.165; df=294; p=.002; *Partial* η^2=.033; m= 2.269, sd=1.317 versus m= 2.763, sd=1.368) formats.

The second phase - IRT

We used Item Response Theory (IRT), applying Marginal Maximum Likelihood Estimation (MMLE), using the IRTPRO 2.1 software (Cai et al., 2011). We estimated two-parameter logistic models (2PLM) (parameter of discrimination – a – and

difficulty – b) without considering a parameter for guessing (c) (1). The following equation provides the applied computation rule (Cai et al., 2011):

$$p_i(x_i \mid \theta, b, a) = \frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \quad (1)$$

where p_i is the probability of each answer for item i defined as a function of the latent dimension and item parameters, θ is the level of ability, b_i is the coefficient of difficulty for item i, a_i is the coefficient of discrimination for item i, and D is a constant. The 2PLM model was estimated because we previously individuated the students who had provided correct responses randomly, using the responses to the open-ended question associated with each problem. We evaluated the assumptions of unidimensionality and local independence (Baker, 2001; Paek & Han, 2013). Moreover, we estimated the model parameters, item fit statistics, population parameters as well as the standard error, item information function, marginal reliability estimate, -2 log likelihood, Akaike Information Criterion and Bayesian Information Criterion (Paek & Han, 2013) (Table 6 and Table 7).

The parameter of discrimination (a) highlighted indices with a wide range for both problem formats. De Ayala (2009) recommended that discrimination indices range from 0.8 to 2.5. Referring to these criteria, we observed that some problems had low discrimination (even if these values changed in our samples). In the N format, problem A3 exhibited low discrimination (a), especially among the Italians under time pressure; the same problem provided medium discrimination for the Spanish sample and Italian sample without time pressure. In the G format, we

Table 6 . Fit measures from IRT (2PLM) in N and G formats

Sample	Statistics	Format N	Format G
Spanish without time pressure	**-2 log likelihood**	780.75	723.98
	Akaike Information Criterion (AIC)	800.75	743.89
	Bayesian Information Criterion (BIC)	829.19	772.42
	G^2	22.22	38.03
	G^2 (df)	21	21
	G^2 p	.38	.01
	RMSEA	0.02	0.08
	X^2	16.83	59.47
	X^2 (df)	21	21
	X^2 p	.72	.01
Italian without time pressure	**-2 log likelihood**	2271.13	2196.48
	Akaike Information Criterion (AIC)	2291.13	2216.48
	Bayesian Information Criterion (BIC)	2330.43	2255.77
	G^2	28.05	29.45
	G^2 (df)	21	21
	G^2 p	.13	.10
	RMSEA	0.03	0.03
	X^2 (df)	27.75	30.51
	X^2 (df)	21	21
	X^2 p	.14	.08
Italian with time pressure	**-2 log likelihood**	1076.66	979.87
	Akaike Information Criterion (AIC)	1096.66	999.87
	Bayesian Information Criterion (BIC)	1128.19	1031.40
	G^2	18.24	36.46
	G^2 (df)	21	21
	G^2 p	.63	.01
	RMSEA	0.00	0.07
	X^2 (df)	16.55	35.73
	X^2 (df)	21	21
	X^2 p	.73	.02

Note: $\chi 2$=Chi-squared test; G^2= Test for maximum likelihood statistical significance; RMSEA=Root mean square error of approximation

observed a low index of discrimination only in the Spanish sample for problem B4. The same problem exhibited a medium index of discrimination in the Italian samples. Moreover, problems B3 and B5 had low discrimination values (below 0.8), particularly for the Spanish sample. High discrimination indices are observed for B1 and B2 in the G format and problem A2 in the N format for all samples.

In relation to the difficulty, in N format, problem A3 appeared difficult, particularly for the Italian sample under time pressure. None of the remaining problems registered high values of *b*. In the G format, we observe a low index of

Table 7. Parameters *a* and *b* resulting from IRT in verbal-numerical and graphical-pictorial formats

		Spanish without time pressure				**Italian without time pressure**				**Italian with time pressure**			
		Parameter a		***Parameter b***		***Parameter a***		***Parameter b***		***Parameter a***		***Parameter b***	
Format	**Item**	***a***	***se***	***b***	***se***	***a***	***se***	***b***	***se***	***a***	***se***	***b***	***se***
N	**A1**	1.13	.37	.16	.20	1.33	.28	.58	.13	2.36	.84	-.11	.12
	A2	2.81	1.44	.08	.13	1.54	.33	.37	.11	1.74	.47	-.29	.14
	A3	.71	.33	1.95	.81	1.02	.22	1.34	.25	.44	.23	2.39	1.22
	A4	1.20	.39	.39	.21	1.60	.38	.39	.11	1.28	.36	.04	.16
	A5	1.11	.37	.30	.21	.78	.18	.92	.23	.89	.27	.54	.24
G	**B4**	.41	.39	-4.88	4.35	1.60	.41	-1.32	.21	1.44	.47	-1.54	.35
	B1	2.52	1.64	.34	.15	1.33	.28	.58	.13	1.32	.36	.32	.16
	B3	.61	.29	-.54	.38	.98	.22	-.41	.14	1.28	.35	-.29	.17
	B2	2.29	1.61	-.03	.14	1.72	.41	-.05	.09	3.20	1.54	.03	.11
	B5	.74	.35	1.40	.58	.79	.20	1.47	.33	.86	.28	1.48	.44

Note: a: Parameter a; b: Parameter b; se: standard error

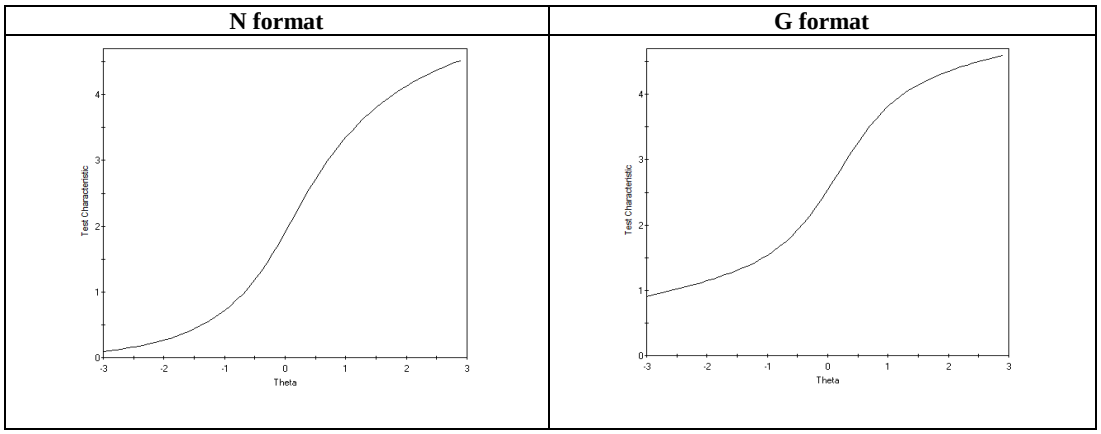


Figure 1. Test Characteristic Curve of the Spanish sample without time pressure by format

Note: abscissa axis: Theta parameter; Ordinate axis: Test Characteristic value

difficulty (*b*) (Baker, 2001) only for problem B4 (representing the classic "ballot box" problem) in all samples. On the other hand, B5 had high and similar index values for all samples. In the Spanish sample, problems B4, B3, and B2 (G format) had a low level of difficulty; instead, these values are higher for B1 and B5. These facts constitute a common trend for all samples. Then, we might assume that this trend might be related to the construction of specific problems.

Comparing the effects pressure compared to no time limits highlighted interesting differences. We observed a greater level of difficulty in the G format than in the N format for problems B1 and B5. We also observed a lower level of difficulty in G for the remaining problems (B2, B3, and B4). Comparing the difficulty among Italian students with and without time pressure, we observe that some problems in the N format exhibit lower indices of difficulty in the sample with time limits (A1, A2, A4 and A5). We observed a higher difficulty index only for A3 under time pressure. In the G format, the difficulty values are lower under time pressure for problems B4 and B1; however, the difficulty indices for B2, B3 and B5 seem similar

under both timing conditions. In Figures 1, 2, and 3 we illustrate the Test Characteristic Curve (TCC) for both formats for each sample. These figures indicate that, in the G format, subjects generally solve more problems correctly, than they do in the N format for the same level of ability (*Theta θ*).

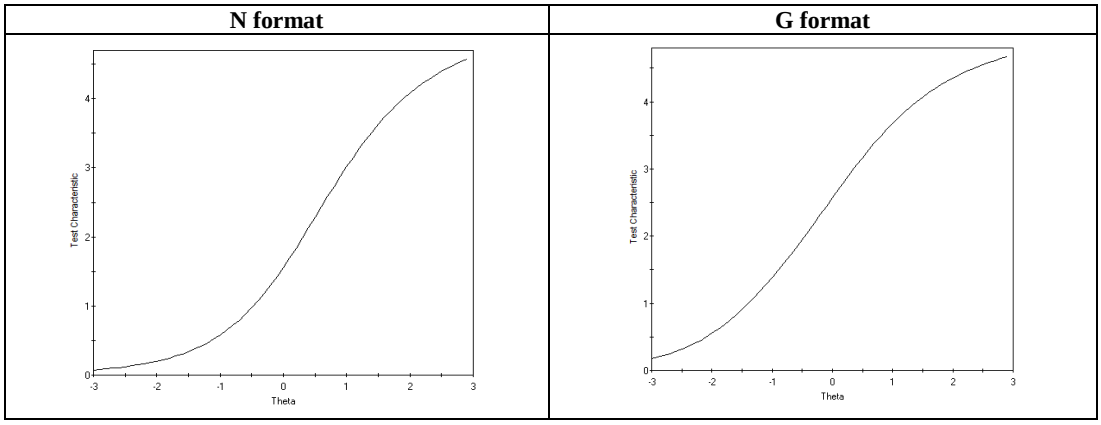


Figure 2. Test Characteristic Curve of the Italian sample without time pressure by format
Note: abscissa axis: Theta parameter; Ordinate axis: Test Characteristic value

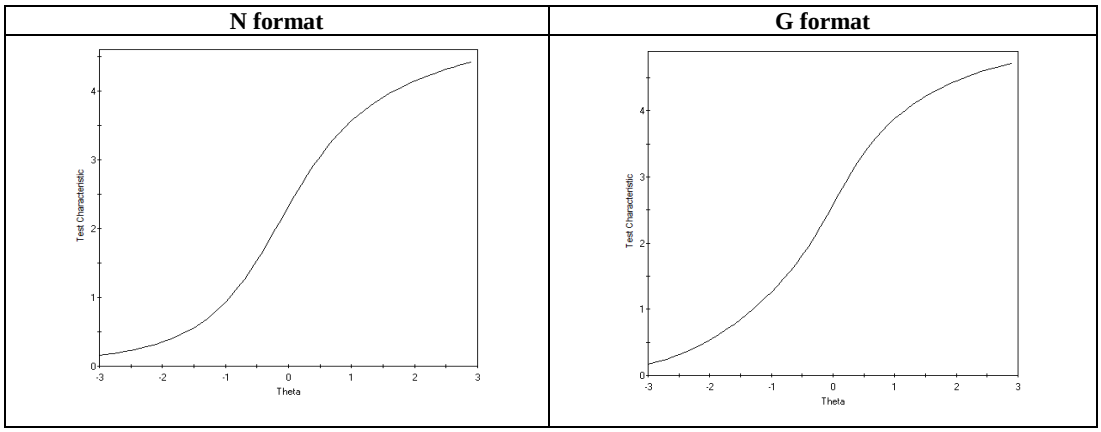


Figure 3. Test Characteristic Curve of the Italian sample with time pressure by format
Note: abscissa axis: Theta parameter; Ordinate axis: Test Characteristic value

GENERAL CONCLUSIONS

Researchers have long discussed the effects of problem presentation format in probabilistic reasoning (e.g., Braithwaite & Goldstone, 2013 a; 2013 b; Brase, 2009; Schonlau & Peters, 2012; Sirota et al., 2014). Notwithstanding, the existing measures used to assess this type of reasoning are only partially appropriate for evaluating the role of problem presentation format. This fact implies some limitations on research about the effect of graphical facilitation (e.g., Moro et al., 2011). Consequently, we developed an instrument to assess the features of this type of reasoning with problems presented in both verbal-numerical and graphical-pictorial formats.

We focused on simple and conditional probabilities expressed as frequencies, consistent with classic studies in the literature (e.g., Brase, 2009; Moro & Bodanza, 2010; Yamagishi, 2003).

The evaluation of the psychometric properties of our problems confirmed that the construction of this assessment tool was demanding. Probabilistic reasoning is a very specific type of thinking, which does not improves at a specific life period, but is related to individual education (e.g., Agnoli & Krantz, 1989; Mandel, 2015). This

specific aspect stimulated us to assess this type of reasoning in a sample of students who have not received statistical education.

The data analysis showed specific patterns that we would illustrate. The CFA highlighted problems with the fit indices, particularly for the G format in the Italian sample without time limits (the RMSEA and CFI produced values over the guidelines) (Schermelleh-Engel et al., 2003). Additionally, the N format problems seem to produce a better fit than the G format problems.

Moreover, a specific concern is related to the α reliability of our scales. Namely, in all samples, these values appeared problematic for both formats. The aim to construct a small and agile instrument containing only five problems in each format could produce practical and empirical advantages but also reduce reliability (Ebel & Frisbie, 1991; Kline, 2000; Sijtsma, 2009).

We provided evidence of a pattern of weak correlations with external constructs, suggestive of concurrent validity. The weak correlations between our scales and external measures (university admission marks and numeric and visuo-spatial PMA scales) might be related to peculiarities of the inquiry (e.g., Klaczynski, 2014; Sirota, & Juanchich, 2011). Indeed, probabilistic reasoning under both presentation formats reasonably presents multifactorial interactions with many processes and aspects that are only partially identified by bivariate linear correlations (Kubiszyn & Borich, 1990).

Therefore, discriminant validity was assessed by estimating the mean differences between undergraduates with different curricula in mathematics. These assessments verified the existence of the supposed differences (Galli et al., 2011). Indeed, we observed that a stronger mathematical education could improve performance in probabilistic reasoning under both formats.

Both the CTT and IRT highlighted higher indices of difficulty, especially under the N format (although with some differences among the specific problems and sub-samples). Furthermore, the values observed for the Italian and Spanish samples without time pressure are similar, although the Spanish exhibited lower difficulty values in the G format. These differences might be related to real differences within the respective populations (e.g., Huggins, 2012; 2014); indeed, the reasoning processes of these undergraduates might be affected by many background variables, reflected as differences in the IRT indices (e.g., Glas, 2001).

The examination of several samples highlighted other interesting differences, especially among Italian students operating with and without time pressure. Larger differences in the index values computed by CTT and IRT were observed for the N format, whilst for the G format, the parameters for discrimination and difficulty were similar. Importantly, three times out of five, the difficulty values were lower for the N format administered under time pressure. These results appear to conflict with the hypothesis that time pressure might impede performance (e.g., Rieskamp & Hoffrage, 2008). The presence of time pressure might support an effective commitment to probabilistic reasoning (e.g., Salehi, Cordero, & Sandi, 2010), particularly in the N format. Moreover, the G format might be less affected by timing in problem administration.

Examining the results of IRT models for both formats, we suppose that the problems in the G format refer to a lower level of ability (*Theta* θ) in probabilistic reasoning than those in the N format. Observing the Test Characteristic Curves (TCC), we noted that the G format appeared simpler than the N format among all samples, especially among Spanish students (consistently, these students exhibited higher visuo-spatial ability scores). These circumstances might also be observed when examining the difficulty indices in the CTT and IRT models. Examining the TCC, we observed that the curves for the Italians students under time pressure seemed better than the corresponding curves of the other samples, because their

performance is superior to the other samples. We might suppose that time pressure improves the functioning of our scales in both formats.

Evaluating the findings of our work, we might suppose that this short instrument of assessment of probabilistic reasoning in N and G formats offers some preliminary evidences of validity, demonstrating its potential applicability and utility in the study of this specific reasoning, also if some aspects could be further enhanced. Indeed, there are a series of implications related to the improvement of the assessment of performance in these problem formats, especially in the investigation of graphical facilitation. Our scales allowed more accurate investigation of probabilistic reasoning with respect to the use of a single problem. Moreover, they permitted the identification of the precise relationships to the individual differences and various outcome variables identified in the literature (e.g., attitudes, emotional aspects, and cognitive styles) (e.g., Lim & Chapman, 2013). To deepen the study of these effects, a promising starting point would be the introduction of a qualitative classification of reasoning. In fact, we recently highlighted that sometimes the N format supports a higher level of reasoning than the G format does (Agus et al., 2014). This aspect requires further investigation, which could be conducted through a multi-method, quantitative and qualitative approach (e.g., Manor, Ben-Zvi, & Aridor, 2014).

However, besides the previous general considerations, some specific criticisms can be identified. For instance, we are aware that the presentation of similar problems in the same work session might be, on the one hand, an advantage (aiming to compare the performance of the same student controlling for the influences of individual differences) but on the other hand, a disadvantage (involving the effect of problem structure learning on performance). In this investigation, we attempted to overcome this limitation by presenting the items in different format orders and problem sequences.

Moreover, the effort to construct a small and agile instrument produces advantages for administration but also reduces reliability and validity evaluations. The construction of a few adequate problems in two formats was complicated by a multiplicity of factors (for example, the effect of wording, problem structure, type of numerical data and type of representation) (e.g., Moro et al., 2011).

Finally, the results of our analyses support the application of this instrument as a sustainable measure of probabilistic reasoning in verbal-numerical and graphical-pictorial formats for undergraduates without statistical expertise. Overall, this measure exhibits acceptable psychometric properties, discriminating across dimensions in two formats and maintaining helpful problems. Although the instrument exhibited some limits from a CTT approach, the preliminary evidence of validity and information furnished by the application of IRT encourage us in the use of these problems. The validation judgement is grounded in a combination of empirical outcomes and theoretical foundations (Messick, 1996). From this perspective, we will systematise future evaluations of our instrument to improve some critical features and assess other aspects of validity. Currently, this instrument offers an improvement over other classical method to assess the effect of graphical facilitation. By evaluating and comparing indices in CTT and IRT, we could identify whether these problems would be suitable to observe these effects and their relationships to other variables identified in the literature, controlling for timing differences. This instrument might be applied in future research, allowing more accurate evaluations of the effects exerted by relevant constructs (cognitive, metacognitive, non-cognitive and contextual dimensions) on graphical facilitation.

The future application of this instrument, extended to subjects with different and specific features, might be useful in determining how students could take advantage of graphical facilitation in probabilistic reasoning.

REFERENCES

Agnoli, F., & Krantz, D. H. (1989). Suppressing natural heuristics by formal instruction: The case of the conjunction fallacy. *Cognitive Psychology*, *21*(4), 515–550. Doi:10.1016/0010-0285(89)90017-0

Agus, M., Peró-Cebollero, M., Guàrdia-Olmos, J., & Penna, M. P. (2013). The measurement of statistical reasoning in verbal-numerical and graphical forms: a pilot study. *Journal of Physics: Conference Series*, *459*(1), 012023. http://doi.org/10.1088/1742-6596/459/1/012023

Agus, M., Peró-Cebollero, M., Penna, M. P., & Guàrdia-Olmos, J. (2014). Towards the development of problems comparing verbal-numerical and graphical formats in statistical reasoning. *Quality & Quantity*, *49*(2), 691–709. http://doi.org/10.1007/s11135-014-0018-7

Agus, M., Peró-Cebollero, M., Penna, M. P., & Guàrdia-Olmos, J. (2015). Comparing Psychology Undergraduates' Performance in Probabilistic Reasoning under Verbal-Numerical and Graphical-Pictorial Problem Presentation Format: What is the Role of Individual and Contextual Dimensions? *Eurasia Journal of Mathematics, Science & Technology Education*, *11*(4), 735–750. http://doi.org/10.12973/eurasia.2015.1382a

Baker, F. B. (2001). *The basics of item response theory*. In ERIC clearinghouse on assessment and evaluation. College Park, MD: University of Maryland. Available http://ericae.net/irt/baker.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *The Behavioral and Brain Sciences, 30*(3), 241–254; http://doi.org/10.1017/S0140525X07001653

Batanero, C., & Sánchez, E. (2005). What is the nature of high school students' conceptions and misconceptions about probability? In G. A. Jones (Ed.), *Exploring probability in school* (pp. 241–266). Springer. Doi:10.1007/0-387-24530-8_11

Ben-Zvi, D., & Garfield, J. (2004). *The challenge of developing statistical literacy, reasoning, and thinking*. Netherlands: Kluwer Academic Pub.

Bentler, P. M. (1995). *EQS structural equations program manual* (Encino, CA., Vol. Multivariate). Multivariate Software.

Bishop, A. (2008). Spatial abilities and mathematics education – A review. In P. Clarkson & N. Presmeg (Eds.), *Critical issues in mathematics education SE - 5* (pp. 71–81). Springer US. Doi:10.1007/978-0-387-09673-5_5

Braithwaite, D. W., & Goldstone, R. L. (2013a). Flexibility in data interpretation: Effects of representational format. *Frontiers in Psychology*, *4*(DEC), 980. Doi:10.3389/fpsyg.2013.00980

Braithwaite, D. W., & Goldstone, R. L. (2013b). Integrating formal and grounded representations in combinatorics learning. *Journal of Educational Psychology*, *105*(3), 666–682. Doi:http://dx.doi.org/10.1037/a0032095

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, *15*(2), 284–289. Doi:10.3758/PBR.15.2.284

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, *23*(3), 369–381. Doi:10.1002/acp.1460

Brase, G. L., Cosmides, L., & Tooby, J. (1998). Individuation, counting, and statistical inference: The role of frequency and whole-object representations in judgment under uncertainty. *Journal of Experimental Psychology: General*, *127*(1), 3. Doi: http://dx.doi.org/10.1037/0096-3445.127.1.3

Brase, G. L., & Hill, W. T. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, *6*, 340. http://doi.org/10.3389/fpsyg.2015.00340

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows [Computer software]*. *Lincolnwood, IL: Scientific Software International* (Scientific.). Lincolnwood, IL.

Castañeda, L. E. G., & Knauff, M. (2013). Individual differences, imagery and the visual impedance effect. In M. Knauff, N. Sebanz, M. Pauen, I. Wachsmuth, (Eds.), *Cooperative minds: Social interaction and group dynamics. Proceedings of the 35th Annual Meeting of the Cognitive Science Society Berlin, Germany, July 31-August 3, 2013* (pp. 2374–9). Austin, TX: Cognitive Science Society. ISBN: 978-0-9768318-9-1

Chan, S. W., & Ismail, Z. (2014). A technology-based statistical reasoning assessment tool in descriptive statistics for secondary school students. *Turkish Online Journal of Educational Technology*, *13*(1), 29–46. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84891655563&partnerID=tZOtx3y1

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3), 1–18.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. Doi:10.1207/S15328007SEM0902_5

Chiesi, F., & Primi, C. (2009). Un modello sul rendimento nelle materie quantitative degli studenti di psicologia. *Giornale Italiano Di Psicologia*, *36*(1), 161–184.

Chiesi, F., Primi, C., & Morsanyi, K. (2011). Developmental changes in probabilistic reasoning: The role of cognitive capacity, instructions, thinking styles, and relevant knowledge. *Thinking & Reasoning*, *17*(3), 315–350. Doi:10.1080/13546783.2011.598401

Clements, M. (2014). Fifty years of thinking about visualization and visualizing in mathematics education: A historical overview. In M. N. Fried & T. Dreyfus (Eds.), *Mathematics & mathematics education: Searching for common ground SE - 11* (pp. 177–192). Springer Netherlands. Doi:10.1007/978-94-007-7473-5_11

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment & Decision Making*, *7*(1). Retrieved from http://journal.sjdm.org/11/11808/jdm11808.html

Corter, J. E., & Zahner, D. C. (2007). Use of external visual representations in probability problem solving. *Statistics Education Research Journal*, *6*(1), 22–50. Retrieved from https://www.stat.auckland.ac.nz/~iase/serj/SERJ6(1)_Corter_Zahner.pdf

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*(1), 1–73. Doi:10.1016/0010-0277(95)00664-8

Del Mas, R. (2004). A Comparison of Mathematical and Statistical Reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking SE - 4* (pp. 79–95). Springer Netherlands. http://doi.org/10.1007/1-4020-2278-6_4

De Ayala, R. J. (2009). *Theory and practice of item response theory*. Guilford Publications.

De Hevia, M. D., Vallar, G., & Girelli, L. (2008). Visualizing numbers in the mind's eye: the role of visuo-spatial processes in numerical abilities. *Neuroscience and Bio-behavioral Reviews*, *32*(8), 1361–72. Doi:10.1016/j.neubiorev.2008.05.015

Díaz, C., & De La Fuente, I. (2006). Assessing psychology students' difficulties with conditional probability and bayesian reasoning. In A. R. & B. Chance (Ed.), *Proceedings of the Seventh International Conference on Teaching Statistics*. Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/17/5E3_DIAZ.pdf

Ebel, R., & Frisbie, D. (1991). *Essentials of educational measurement* (5th ed., p. 383). Englewood Cliffs, NJ. P: Prentice-Hall.

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

Evans, J. S. B. T., Handley, S. J., Neilens, H., & Over, D. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *The Quarterly Journal of Experimental Psychology*, *63*(5), 892–909. Doi:10.1080/17470210903111821

Evans, J. S. B. T., 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. Doi:10.1016/S0010-0277(00)00098-6

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of higher cognition: advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241. http://doi.org/10.1177/1745691612460685

Fan, X. (1998). Item Response Theory and Classical Test Theory: An empirical comparison of their Item/Person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. Doi:10.1177/0013164498058003001

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). Guidelines for assessment and instruction in statistics education (GAISE). Report.

Alexandria: American Statistical Association. Retrieved from http://www.amstat.org/Education/gaise/GAISEPreK-12_Full.pdf

Frey, A., & Seitz, N.N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, *35*(2–3), 89–94. http://doi.org/http://dx.doi.org/10.1016/j.stueduc.2009.10.007

Furlan, S., & Agnoli, F. (2010). The dark side of statistics: Numeracy and luck in the development of probabilistic reasoning. In *ICOTS8 - Data and context in statistics education: Towards an evidence-based society*. Retrieved from http://icots.info/8/cd/pdfs/posters/ICOTS8_P11_FURLAN.pdf

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1–25. http://doi.org/10.1111/j.1751-5823.2002.tb00336.x

Gal, I. (2005). Towards "probability literacy" for all citizens: Building blocks and instructional dilemmas. In G. A. Jones (Ed.), *Exploring probability in school* (pp. 39–63). Springer. Doi: 10.1007/0-387-24530-8_3

Galli, S., Chiesi, F., & Primi, C. (2011). Measuring mathematical ability needed for "non-mathematical" majors: The construction of a scale applying IRT and differential item functioning across educational contexts. *Learning and Individual Differences*, *21*(4), 392–402. Doi:10.1016/j.lindif.2011.04.005

García-Retamero, R., Galesic, M., & Gigerenzer, G. (2011). Cómo favorecer la comprensión y la comunicación de los riesgos sobre la salud. *Psicothema*, 23(4), 599–605.

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, *2*(1), 22–38. Retrieved from http://iase-web.org/documents/SERJ/SERJ2(1).pdf#page=24

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. New York: Springer Verlag.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. New York: Oxford University Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. Doi: http://dx.doi.org/10.1037/0033-295X.102.4.684

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

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, *78*(3), 247–276. Doi:10.1016/S0010-0277(00)00133-5

Girotto, V., & Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition*, *106*(1), 325–344. Doi:10.1016/j.cognition.2007.02.005

Glas, C. A. W. (2001). Differential item functioning depending on general covariates. In *Essays on item response theory* (pp. 131–148). Berlin / Heidelberg: Springer.

González M.A., Campos A., Perez M. J. (1997). Mental imagery and creative thinking. *The Journal of Psychology*, *131*(4), 357–364. Doi:10.1080/00223989709603521

Guàrdia-Olmos, J., Freixa-Blanxart, M., Peró-Cebollero, M., Turbany, J., Cosculluela, A., Barrios, M., & Rifà, X. (2006). Factors related to the academic performance of students in the statistics course in psychology. *Quality & Quantity*, *40*(4), 661–674. Doi:10.1007/s11135-005-2072-7

Hays, R. D., Brown, J., Brown, L. U., Spritzer, K. L., & Crall, J. J. (2006). Classical test theory and item response theory analyses of multi-item scales assessing parents' perceptions of their children's dental care. *Med Care*, *44*, S60–S68. Doi:10.1097/01.mlr.0000245144.90229.d0

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2–3), 57–63. http://doi.org/http://dx.doi.org/10.1016/j.stueduc.2009.10.002

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, *84*(3), 343–352. Doi: 10.1016/S0010-0277(02)00050-1

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. Doi:10.1080/10705519909540118

Huggins, A. C. (2012). *The Effect of Differential Item Functioning on Population Invariance of Item Response Theory True Score Equating*. University of Miami.

Huggins, A. C. (2014). The Effect of Differential Item Functioning in Anchor Items on Population Invariance of Equating. *Educational and Psychological Measurement*, *74* (4), 627–658. Doi:10.1177/0013164413506222

Johnson, E. D., & Tubau, E. (2013). Words, numbers, & numeracy: Diminishing individual differences in Bayesian reasoning. *Learning and Individual Differences*, *28*, 34–40. Doi:10.1016/j.lindif.2013.09.004

Johnson, M. E., Pierce, C. A., Baldwin, K., Harris, A., & Brondmo, A. K. (1996). Presentation Format in Analogue Studies: Effects on Participants Evaluations. *The Journal of Psychology*, *130*(3), 341–349. Doi:10.1080/00223980.1996.9915015

Jones, G. A. (2006). *Exploring probability in school: Challenges for teaching and learning* (Vol. 40). Springer.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124

Kellen, V., Chan, S., & Fang, X. (2006). Individual Differences in Spatial Abilities and the Visualization of Conditional Probabilities. Retrieved from http://www.kellen.net/visualization_wp.htm

Kellen, V., Chan, S., & Fang, X. (2007). Facilitating Conditional Probability Problems with Visuals. In J. Jacko (Ed.), *Human-Computer Interaction. Interaction Platforms and Techniques* (Vol. 4551, pp. 63–71). Berlin / Heidelberg: Springer. Doi:10.1007/978-3-540-73107-8_8

Kellen, V., Chan, S., & Fang, X. (2013). Improving user performance in conditional probability problems with computer-generated diagrams. In *Human-Computer Interaction. Users and Contexts of Use* (pp. 183–192). Springer

Klaczynski, P. A. (2014). Heuristics and biases: interactions among numeracy, ability, and reflectiveness predict normative responding. *Frontiers in Psychology*, *5*, 665. http://doi.org/10.3389/fpsyg.2014.00665

Kline, P. (2000). *Handbook of psychological testing*. Routledge.

Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory & Cognition*, *30*(3), 363–371. Doi:10.3758/BF03194937

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2014). Data seen through different lenses. *Educational Studies in Mathematics*, 1–21. Doi:10.1007/s10649-013-9529-8

Kubiszyn, T., & Borich, G. (1990). *Educational testing and measurement*. Harper Collins Publishers.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, *35*(2–3), 64–70. http://doi.org/http://dx.doi.org/10.1016/j.stueduc.2009.10.003

Lalonde, R. N., & Gardner, R. C. (1993). Statistics as a second language? A model for predicting performance in psychology students. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *25*(1), 108–125. http://doi.org/http://dx.doi.org/10.1037/h0078792

Langrall, C. W., & Mooney, E. S. (2005). Characteristics of Elementary School Students' Probabilistic Reasoning. In G. A. Jones (Ed.), *Exploring Probability in School* (pp. 95–119). Springer. http://doi.org/10.1007/0-387-24530-8_5

Laverdière, O., Morin, A. J. S., & St-Hilaire, F. (2013). Factor structure and measurement invariance of a short measure of the Big Five personality traits. *Personality and Individual Differences*, *55*(7), 739–743. http://doi.org/10.1016/j.paid.2013.06.008

Lean, G., & Clements, M. A. (1981). Spatial ability, visual imagery, and mathematical performance. *Educational Studies in Mathematics*, *12*(3), 267–299. Doi:10.1007/BF00311060

Lim, S., & Chapman, E. (2013). Development of a short form of the attitudes toward mathematics inventory. *Educational Studies in Mathematics*, *82*(1), 145–164. Doi:10.1007/s10649-012-9414-x

Manor, H., Ben-Zvi, D., & Aridor, K. (2014). Student's reasoning about uncertainty while making informal statistical inferences in an "Integrated Pedagogic Approach." In & R. G. Makar, B. de Sousa (Ed.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona,*

USA. Voorburg, The Netherlands: International Statistic. International Statistical Institute.

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Frontiers in Psychology*, 5. http://doi.org/10.3389/fpsyg.2014.01144

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology*. http://doi.org/10.3389/fpsyg.2015.00387

Maule, A. J., Hockey, G. R. J., & Bdzola, L. (2000). Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. *Acta Psychologica*, *104*(3), 283–301. Doi:10.1016/S0001-6918(00)00033-0

Messick, S. (1996). Validity and washback in language testing. *Language Testing* , *13* (3), 241–256. Doi:10.1177/026553229601300302

Moore, D. S. (1990). Uncertainty. *On the Shoulders of Giants: New Approaches to Numeracy*, 95–137.

Moro, R., & Bodanza, G. A. (2010). El debate acerca del efecto facilitador en problemas de probabilidad condicional: ¿Un caso de experimentación crucial? *Interdisciplinaria*, *27*(1), 163–174.

Moro, R., Bodanza, G. A., & Freidin, E. (2011). Sets or frequencies? How to help people solve conditional probability problems. *Journal of Cognitive Psychology*, *23*(7), 843–857. Doi:10.1080/20445911.2011.579072

Onwuegbuzie, A. J., & Seaman, M. A. (1995). The effect of time constraints and statistics test anxiety on test performance in a statistics course. *The Journal of Experimental Education*, *63*(2), 115–124. Doi: 10.1080/00220973.1995.9943816

Paek, I., & Han, K. T. (2013). IRTPRO 2.1 for Windows (Item Response Theory for Patient-Reported Outcomes). *Applied Psychological Measurement, 37*(3), 242–252. Doi:10.1177/0146621612468223

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart & Winston.

Pastore, M., & Lombardi, L. (2014). The impact of faking on Cronbach's alpha for dichotomous and ordered rating scores. *Quality & Quantity*, *48*(3), 1191–1211. http://doi.org/10.1007/s11135-013-9829-1

Penna, M. P., Agus, M., Peró-Cebollero, M., Guàrdia-Olmos, J., & Pessa, E. (2014). The use of imagery in statistical reasoning by university undergraduate students: a preliminary study. *Quality & Quantity*, *48*(1), 173–187. http://doi.org/10.1007/s11135-012-9757-5

Pessa, E., & Penna, M. P. (2000). *Manuale di scienza cognitiva. Intelligenza artificiale classica e psicologia cognitiva*. Roma - Bari: Editori Laterza.

Pollard, B., Dixon, D., Dieppe, P., & Johnston, M. (2009). Measuring the ICF components of impairment, activity limitation and participation restriction: an item analysis using classical test theory and item response theory. *Health and Quality of Life Outcomes*, *7*(1), 41. Doi:10.1186/1477-7525-7-41

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modelling for evaluating questionnaire item and scale properties. *Assessing Quality of Life in Clinical Trial: Methods and Practice*, 55–74. Doi: 10.1007/s11136-007-9198-0

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, *127*(2), 258–276. Doi:10.1016/j.actpsy.2007.05.004

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3), 6–13. Retrieved from http://www.amstat.org/publications/jse/v10n3/rumsey2.html

Salehi, B., Cordero, M. I., & Sandi, C. (2010). Learning under stress: the inverted-U-shape function revisited. *Learning & Memory*, *17*(10), 522–530. Doi:10.1101/lm.1914110

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Schinka, J. A., & Velicer, W. F. (2003). *Handbook of Psychology - vol. 2 - Research methods in psychology*. (I. B. Weiner, Ed.) (p. 738). Wiley.

Schonlau, M., & Peters, E. (2012). Comprehension of graphs and tables depend on the task: empirical evidence from two web-based studies. *Statistics, Politics, and Policy*, *3*(2). Doi:10.1515/2151-7509.1054

Sharps, M. J., Hess, A. B., Price-Sharps, J. L., & Teh, J. (2008). Heuristic and algorithmic processing in English, Mathematics, and Science Education. *The Journal of Psychology*, *142*(1), 71–88. Doi:10.3200/JRLP.142.1.71-88

Sijtsma, K. (2009a). Reliability beyond theory and into practice. *Psychometrika*, *74*(1), 169–173. Doi:10.1007/s11336-008-9103-y

Sijtsma, K. (2009b). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. http://doi.org/10.1007/s11336-008-9101-0

Singh, J. (2004). Tackling measurement problems with Item Response Theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research*, *57*, 184–208. Doi:10.1016/S0148-2963(01)00302-2

Sirota, M., & Juanchich, M. (2011). Role of numeracy and cognitive reflection in Bayesian reasoning with natural frequencies. *Studia Psychologica, 53*(2), 151–161. Retrieved from http://yadda.icm.edu.pl/cejsh/element/bwmeta1.element.defbace5-44e3-3d1b-8464-09dde16918e3

Sirota, M., Kostovičová, L., & Juanchich, M. (2014). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychonomic Bulletin &* Review, 21, 961–968. http://doi.org/10.3758/s13423-013-0555-4

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015). Now you Bayes, now you don't: effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychonomic Bulletin & Review*, 1–9. http://doi.org/10.3758/s13423-015-0810-y

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, *91*(2), 296–309. Doi:10.1016/S0749-5978(03)00021-9

Thurstone, L. L., & Thurstone, T. G. (1981). *PMA - abilità mentali primarie: manuale di istruzioni Livello intermedio (11-17)(Batteria fattoriale delle abilità mentali primarie)*. Firenze: Organizzazioni Speciali.

Thurstone, L. L., & Thurstone, T. G. (1987). *TEA: tests de aptitudes escolares [niveles 1, 2 y 3]Manual*. Madrid: Tea.

Tubau, E. (2008). Enhancing probabilistic reasoning: The role of causal graphs, statistical format and numerical skills. *Learning and Individual Differences*, *18*(2), 187–196. Doi:10.1016/j.lindif.2007.08.006

Tufte, E. R. (2001). *The visual display of quantitative information*. *Visual Explanations* (pp. 194–195). Cheshire, Connecticut: Graphics Press

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. Doi: http://dx.doi.org/10.1037/0033-295X.90.4.293

Wang, M., & Russell, S. S. (2005). Measurement Equivalence of the Job Descriptive Index Across Chinese and American Workers: Results from Confirmatory Factor Analysis and Item Response Theory. *Educational and Psychological Measurement*, *65* (4), 709–732. Doi:10.1177/0013164404272494

Watson, J. M., & Moritz, J. B. (2003). Fairness of dice: A longitudinal study of students' beliefs and strategies for making judgments. *Journal for Research in Mathematics Education*, 270–304.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. http://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? *Experimental Psychology (formerly Zeitschrift Für Experimentelle Psychologie)*, *50*(2), 97–106. Doi:10.1026//1618-3169.50.2.97

Zahner, D., & Corter, J. E. (2010). The process of probability problem solving: Use of external visual representations. *Mathematical Thinking and Learning*, *12*(2), 177–204. Doi:10.1080/10986061003654240

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, *98*(3), 287–308. Doi: http://dx.doi.org/10.1016/j.cognition.2004.12.003

◈◈◈

APPENDIX: VERBAL-NUMERICAL FORMAT PROBLEMS

ITEM A_1

We have a deck of 52 cards with four suits each consisting of 13 cards (hearts, spades, diamonds and clubs). Each suit includes an ace.
What is the probability that a card selected at random from the deck will be a heart or an ace?

a) 13 / 52
b) 16 / 52
c) 17 / 52
d) 4 / 52

ITEM A_2

In a group of 300 students enrolled in a statistics course, 150 completed the coursework, while the remaining 150 completed only half the coursework.
Of the 150 students who completed half of the programme, 15 passed the exam and 135 did not. Of the 150 students who completed the full programme, 145 passed the examination and 5 did not.
What is the probability that a student who has passed the examination completed only half of the coursework?

a) 15 / 300
b) 15 / 160
c) 15 / 135
d) 15 /145

ITEM A_3

Imagine simultaneously throwing two dice each of which has six faces.
What is the probability that the sum is a number less than or equal to six?

a) 15 / 36
b) 16 / 36
c) 18 / 36
d) 12 / 36

ITEM A_4

A factory produces electronic games, but not all the games work well. For every 100 games produced, 20 might have an electrical problem and 80 might work correctly.
The company has developed control systems to identify faulty games; however, these systems do not work properly. In reality, half of the games with electrical problems continue in the production line, where they are considered as well functioning.
If you randomly select a game that has been sent to shops for sale and evaluated as free of defects, what is the probability that it is defective?

a) 10 / 90
b) 10 / 100
c) 10 / 80
d) 20 / 100

ITEM A_5

Some women develop illness X, which seems to be related to high blood pressure (hypertension). To treat this illness, a medication has been developed that might have undesirable side effects. The following table lists occurrence of side effects for this drug, which was administered to both hypertensive and non-hypertensive women. If 103 of 1000 women have side effects with the use of this drug, how many will be hypertensive?

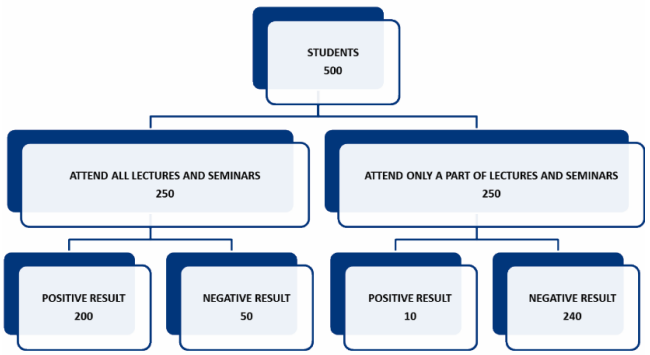
	SIDE EFFECT YES	SIDE EFFECT NO	TOTAL
HYPERTENSIVE WOMEN	80	20	100
NON HYPERTENSIVE WOMEN	950	8950	9900
TOTAL	1030	8970	10000

a) 8 / 103
b) 9.5 / 103
c) 95 / 103
d) 80 / 103

GRAPHICAL-PICTORIAL FORMAT PROBLEMS

ITEM B_1

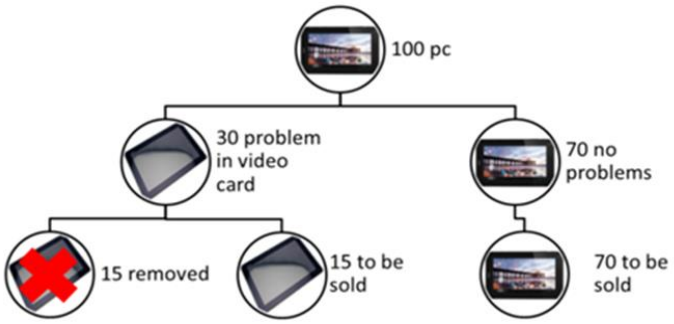
Of 500 students in a biology course, half attended only some lectures and seminars, while the other half attended all lectures and seminars.
Considering the following graph, what is the probability that a student who passed the examination attended only some lectures and seminars?



a) 10 / 250
b) 10 / 240
c) 10 / 210
d) 10 / 500

ITEM B_2

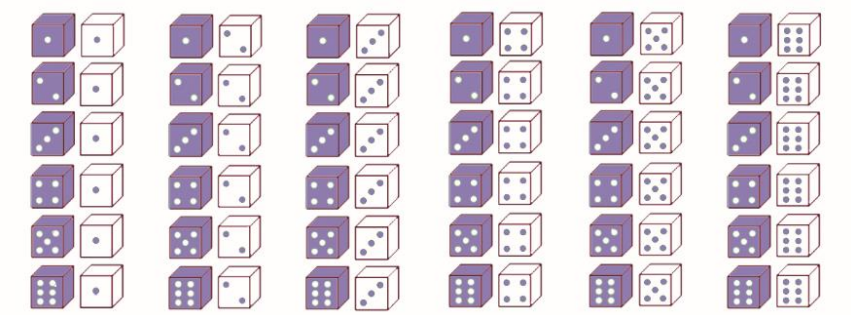
A factory that produces personal computers has problems in its production process. Some of the computers are defective (they have problems with the video card). Such problems are not always identified by quality control and consequently some defective computers are sent along the production line. The graphic below illustrates this process.
What is the probability that a computer sent to shops for sale and evaluated as free of defects is defective?



a) 15 / 100
b) 15 / 70
c) 15 / 85
d) 3 / 10

ITEM B_3

Imagine simultaneously throwing two dice that both have six faces; one die is white and one is purple.
What is the probability that the number on the white face is higher?



a) 15 / 36
b) 6 / 36
c) 18 / 36
d) 12 / 36

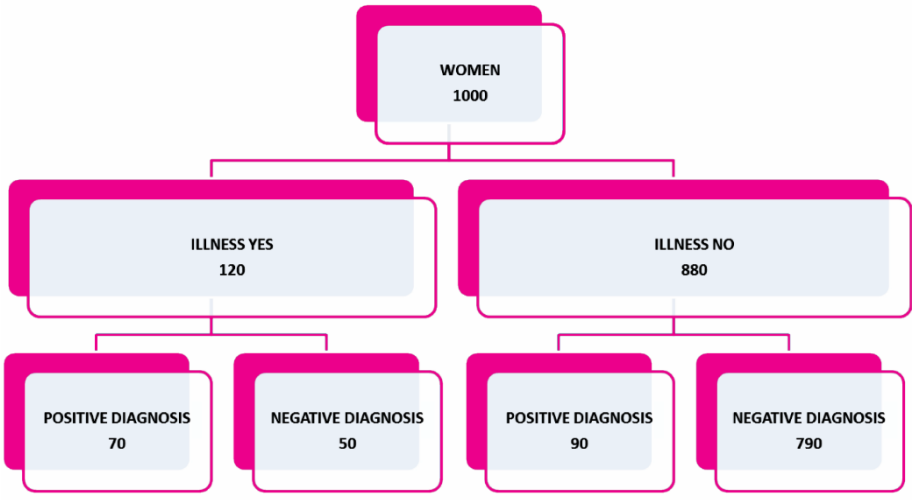
ITEM B_4

Consider the following representation. What is the probability of drawing a ball that is not red from the urn?

a) 8 / 22
b) 10 / 22
c) 4 / 18
d) 18 / 22

ITEM B_5

Some women develop disease Y. To improve diagnosis, a laboratory test has been developed. The following diagram illustrates the results of this test.
If 16 of 100 women receive a positive test result, how many of these 16 women actually have the illness?



a) 7 / 16
b) 9 / 16
c) 5 / 16
d) 12 / 16

