MODESTUM
OPEN ACCESS

EURASIA Journal of Mathematics, Science and Technology Education, 2020, 16(2), em1820
ISSN:1305-8223 (online)
Research Paper https://doi.org/10.29333/ejmste/113111

What's about the Calibration between Confidence and Accuracy? Findings in Probabilistic Problems from Italy and Spain

Mirian Agus [1*], Maribel Peró-Cebollero [2], Joan Guàrdia-Olmos [2], Igor Portoghese [1], Maria Lidia Mascia [1], Maria Pietronilla Penna [1]

[1] University of Cagliari - Faculty of Humanistic Studies, ITALY
[2] Department of Social and Quantitative Psychology, Faculty of Psychology, University of Barcelona, SPAIN

Received 21 June 2019 ▪ Revised 11 October 2019 ▪ Accepted 16 October 2019

ABSTRACT

This paper reports some experiments on probabilistic reasoning designed to investigate the impact of the probabilistic problem presentation format (verbal-numerical and graphical-pictorial) on subjects' confidence in the correctness of their performance, other than the calibration between confidence and accuracy. To understand the potential effect of the format, these dimensions were assessed by monitoring contextual and individual variables: time pressure, numerical and visuospatial abilities, statistical anxiety and attitudes towards statistics. The participants included 257 Psychology students without statistical knowledge, recruited from Italian and Spanish universities, who fulfilled self-report validated measures. The students expressed their retrospective judgments of confidence item-by-item in relation to each probabilistic problem. This approach enabled the computation of two measures of calibration (the Bias Index - the Confidence-Judgment Accuracy Quotient). The results indicated that the problem presentation format did not exert a significant main effect on confidence, with the exception of when the interaction between the format and one subscale of the attitudes towards the statistics test was considered. The Bias Index, however, was significantly related to the interaction between format and time pressure. The study offers a point of reflection in relation to the potential effect exerted by the problem format and time constraint in calibration.

Keywords: probabilistic reasoning, problem format, reasoning confidence, performance accuracy, metacognitive calibration, time pressure

INTRODUCTION

Performance in Probabilistic Reasoning

Probabilistic reasoning is recognized as an essential tool in daily life and the educational lifetime. It is applied by individuals to address uncertainty, collect and analyse information, and understand data (Frosch & Johnson-Laird, 2011). The substantial troubles encountered by students in the solution of these types of problems have been highlighted by many researchers (Brase & Hill, 2015). These difficulties are strong in undergraduates pursuing humanistic studies (e.g., Psychology), who often exhibit negative attitudes towards scientific disciplines, such as statistics and mathematics (Guàrdia-Olmos et al., 2006), evidencing low abilities in these topics, as well as high levels of statistical anxiety (Chiesi, Primi, & Carmona, 2011). The presence of these troubles has been induced to explore the cognitive processes that underlie probabilistic reasoning in individuals without statistical knowledge (Ayal & Beyth-Marom, 2014).

The study of probabilistic problem solving is considered useful to inquire into the relationships between cognitive processes and individual differences, specifically referring to attitudes, abilities and task features.

© 2020 by the authors; licensee Modestum Ltd., UK. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/).

✉ mirian.agus@unica.it **(*Correspondence)** ✉ mpero@ub.edu ✉ jguardia@ub.edu
✉ igor.portoghese@gmail.com ✉ marialidiamascia@gmail.com ✉ penna@unica.it

Contribution of this paper to the literature

- This manuscript involves multiple studies; it describes experiments on probabilistic reasoning designed to investigate the impact of the probabilistic problem presentation format (verbal-numerical and graphical-pictorial) on subjects' confidence in the correctness of their performance, as well as on the calibration between confidence and accuracy. To understand the potential effects of the format, these dimensions were assessed by monitoring contextual and individual variables: time pressure, numerical and visuospatial abilities, statistical anxiety and attitudes towards statistics. This approach furnishes information in relation to the effects of graphical facilitation, which have not been exhaustively defined in the literature.
- An innovative aspect of this work is the evaluation of these relationships in undergraduates in humanities and social sciences, who had not previously studied statistics, in which the effect of graphical facilitation versus graphical impediment could assume specific features. The students expressed their retrospective judgments of confidence item-by-item in relation to each probabilistic problem. This approach enabled the computation of two measures of metacognitive calibration (absolute calibration - the Bias Index, relative calibration - the Confidence-Judgment Accuracy Quotient).
- This investigation stimulates the cross-country comparison between Italian and Spanish undergraduates; they belong to countries in the European Higher Education Area (EHEA) and are characterized by both common and differential aspects. We investigated whether the cultural aspects affect the performances and calibration in probabilistic reasoning.

Particularly, some scholars have underlined the relationships between individual differences and performance in problems that involved inductive, deductive, heuristic and methodological reasoning (Jackson, Kleitman, Howie, & Stankov, 2016; Stanovich & West, 2000).

Hafenbrädl and Hoffrage (2015) analysed multiple aspects of problems that affect the application of probabilistic and Bayesian reasoning, promoting an ecological approach to the study of this type of reasoning. Exploratory analyses highlighted as both qualitative and quantitative dimensions may affect the Bayesian reasoning and the choice of the solution strategy (Hafenbrädl & Hoffrage, 2015).

Within this context, many investigations have concerned the role of different aspects in the enhancement vs detriment of probabilistic reasoning. Among these studies, the most important ones have included the so-called effect of "graphical facilitation"; the latter expression indicates that the use of graphical representations may improve performance in probabilistic problem solving (e.g., Brase & Hill, 2015; Garcia-Retamero & Cokely, 2014; Moro, Bodanza, & Freidin, 2011). However, in the literature, the arguments that may support the occurrence of this effect are controversial. Several scholars have highlighted that the use of strategies could affect the cognitive processes implied in probabilistic reasoning (Gutierrez & Schraw, 2015; Nietfeld & Schraw, 2002). Among these strategies, Gutierrez and Schraw (2015) included the presentation of diagrams, graphs, pictures and tables, which may increase the accuracy, thereby fostering the calibration between accuracy and confidence in problem-solving and learning processes.

Some authors in the literature (Okan, Garcia-Retamero, Cokely, & Maldonado, 2015) have pointed out that there is a relationship among numerical and visuospatial abilities, metacognitive processes and meaningful reasoning about probabilities. These relationships were observed in patients assessing health-relevant numeric data, both in groups with high and low education levels (Garcia-Retamero et al., 2015; Ghazal, Cokely, & Garcia-Retamero, 2014). In a review by Garcia-Retamero and Cokely (2017), psychological and cognitive mechanisms acting in the use of visual aids in this field have been highlighted. These authors observed that graphical representations might enhance judgment and the process of decision-making; indeed, the visual displays seemed to improve the ability of self-assessment and, as well, moderate overconfidence. These effects on metacognitive and behavioural function seem related to a better distribution of attention, which may increase, to a significant level, the comprehension of numerical data. These authors depict a conceptual framework in which graphical representations might serve as "moderators" of the effect of individual differences in numerical and graphical abilities, as may affect final decision-making and behaviour. Additionally, within this framework, there are the mediating effects of cognitive and metacognitive functioning (i.e., accuracy in the uncertainty understanding, trust in information and, finally, behavioural dimensions). Furthermore, graphical representations appear to improve the user's trust in his/her abilities (Garcia-Retamero & Cokely, 2017), promoting a significant relationship among skills, encoding, self-regulation, metacognition, final uncertainty evaluation and decision-making (Cokely & Kelley, 2009; Ghazal et al., 2014). In order to foster the comprehension of probabilistic data, the graphical representations have to be "transparent" (Garcia-Retamero & Cokely, 2017); this means they must be simple, clear, explicitly chosen relative to the aim of the communication and responsive to the user's expertise and needs.

In agreement with these considerations, we designed this study to investigate the specific role of metacognition that supports the probabilistic reasoning performance. In particular, the focus of our interest is on the confidence and the previously described calibration in problems presented in different formats (i.e., verbal-numerical and graphical pictorial formats). In the present research, these aspects have been investigated by monitoring the effects exerted by a specific contextual characteristic (the presence vs absence of time pressure), as well as several cognitive and non-cognitive dimensions (i.e., numerical and visuospatial abilities, statistical anxiety and attitudes).

Metacognition and Confidence in the Correctness of Performance: Their Relationships with Individual and Contextual Dimensions

The confidence in the accuracy of a response has been recognized as a relevant and strong predictor of both achievement (Stankov, 2013; Stankov, Lee, Luo, & Hogan, 2012; Tempelaar, 2009) and probabilistic reasoning (Agus et al., 2015b).

In the literature, it is often indicated that in the probabilistic reasoning, metacognition has a relevant role, which refers to the knowledge of both cognitive processes and cognitive activities (Lin & Zabrucky, 1998). Discussing this construct, Lin and Zabrucky (1998, p. 345) stated that "regulation of cognition (…) refers to the effectiveness with which learners keep track of ongoing cognitive processes and their employment of regulatory strategies in order to solve problems". Metacognition includes the knowledge regarding cognition and the regulation of cognition; these processes control and monitor an individual's decisions and actions (Stankov, 2013). Metacognitive monitoring concerns the link between the performance and the judgment regarding the same performance (Boekaerts & Rozendaal, 2010; Gutierrez, Schraw, Kuch, & Richmond, 2016). Additionally, metacognitive confidence denotes a specific judgment expressed in relation to performance (Dinsmore & Parkinson, 2013; Schraw, 2009). The item-specific confidence is referred to in the literature as comprehension monitoring, metacognition, metamemory and feeling of knowing (Lundeberg, Fox, & Punćochaŕ, 1994).

The theories on self-regulated learning (Boekaerts, 1997; Pintrich, 2000) highlight the role of the processes applied by students to set goals, assess the advancements towards these goals, and modify and regulate their performance on the basis of this monitoring process. Focusing on the relation between the cognitive performance and confidence in their correctness, studies have highlighted two different potential outcomes: overconfidence, when subjects have the illusion of knowing; and underconfidence, when subjects have the illusion of not knowing (Serra & Metcalfe, 2009). In other cases, the subjects properly judge the correctness vs incorrectness of their performance (Gutierrez et al., 2016).

In the literature, many authors have indicated that confidence is dependent on the context (items correct or incorrect) and the specific domain assessed (Glenberg & Epstein, 1987; Lundeberg et al., 1994). Also, other scholars (Dinsmore & Parkinson, 2013) have highlighted that the individual, in the development of confidence judgments, refers to many aspects, including the prior knowledge, the task features and the framework.

Among the dimensions that affect the confidence in the correctness of a response, the previous knowledge and abilities play key roles. Specifically, the effects of numerical abilities have been investigated to account for their effects on our domain of interest, i.e., probabilistic reasoning (Garcia-Retamero & Hoffrage, 2013; Lalonde & Gardner, 1993; Nietfeld & Schraw, 2002). In addition, the effects of visuospatial abilities have been considered, accounting for the skill in understanding and transforming symbolic and non-linguistic data (Gardner, 1993). These abilities exhibited specific relationships with achievement in mathematics and statistics and their corresponding confidence in the correctness of responses (Garcia-Retamero & Cokely, 2013; Kellen, Chan, & Fang, 2013; Maloney, Waechter, Risko, & Fugelsang, 2012). Moreover, attitudes towards statistics have also been investigated. They represent a compound concept, related to the positive versus negative dispositions towards statistical and probabilistic disciplines (Chiesi & Primi, 2009; Gal, Garfield, & Gal, 1997). These attitudes may have a significant effect on the strategies applied by students when they cope with these probabilistic problems. Additionally, the role of statistical anxiety has been inquired; in this regard, many authors have highlighted the relation between this construct and performance in statistical and probabilistic problems (Onwuegbuzie, 1995; Primi & Chiesi, 2016).

As well, among the context features that potentially affect probabilistic reasoning and the relative metacognitive processes, a relevant role of the presence versus absence of time pressure must be recognized (Beilock, Kulp, Holt, & Carr, 2004; Kleiner, 2014). Namely, time constraints may interact with the previously described dimensions, explicitly with statistical anxiety and attitudes (Tobias & Everson, 2009). However, the findings related to the potential influence of time pressure on mathematical and probabilistic problem solving are controversial. Some scholars have highlighted that the presence of time pressure may support performance in these tasks (Hanoch & Vitouch, 2004; Markman, Maddox, & Worthy, 2006). In contrast, other authors have emphasized that time pressure may impede the application of correct solution strategies, for example, by affecting and overloading the working memory (Beilock & Carr, 2005; DeCaro, Thomas, Albert, & Beilock, 2011). Researchers have developed the "Distraction theories" (Beilock & Carr, 2005), which indicate that time pressure produces a disturbing setting that draws attention away from the task. These theories have been challenged by the "Explicit monitoring theories", which suggest that time limits may stimulate more attention to the specific problem, strategies and procedures (Beilock et al., 2004).

The Calibration of Confidence and its Measures

The calibration of confidence is defined as the relationship between performance in a task and confidence in the correctness of this task (Alexander, 2013; Boekaerts & Rozendaal, 2010; Gutierrez et al., 2016; Schraw, 2009). This concept is related to the amount of matching between a subject's judgments on his/her performance and the effective performance gathered (Alexander, 2013). The role of calibration in accounting for performance is important because the subject's judgments regarding the ongoing task affect their ensuing effort and the application of specific solution strategies (Alexander, 2013). According to Dunlosky and Thiede (2013), in general, research regarding calibration focuses on four themes, defined as "cornerstones": judgment bases (the investigation of individuals' construction of metacognitive judgments), judgment accuracy (the analysis of the matching between metacognitive judgments and performance), reliability and stability (the study of the constancy and reliability of the judgments), and control (the investigation of the use of metacognitive judgments to control the developing procedures).

Schraw (2009) agreed that many different variables might affect metacognition and calibration. Among these, we may include the variables that characterise individual differences (e.g., anxiety, attitudes, abilities, knowledge, working memory, executive functions, and cultural context) (Buratti & Allwood, 2015). Dinsmore and Parkinson (2013) remarked that the calibration is affected by the individual's knowledge; specifically, the activation of an unappropriated knowledge may adversely affect confidence judgments and their calibration during performance. In addition, the format and difficulty of the problem may have an impact on the calibration (Schraw, 2009); their influence concerns the type of judgment required and the timing of the judgment that refers to the task. Schraw (2009) introduced different types of metacognitive judgments, which refer to the moment in which the metacognitive judgment is expressed in relation to the performance: prospective (the prediction prior to acting out the assignment), concurrent (the judgment is settled whereas the task is carried out), and retrospective (the judgment is expressed after the performance).

To account for all aspects implied in the calibration, in the literature there are different types of measures that aim to assess the "goodness of fit between a confidence judgment and the corresponding performance" (Schraw, 2009, p. 425). In this study, the choice of the calibration indices to be used has been related to practical considerations on concrete issues related to the features of the research protocol (Rutherford, 2017). Among the measures of calibration, we distinguish between the measures of absolute accuracy and the measures of relative accuracy (Dougherty & Sprenger, 2006; Schraw, 2009).

The measures of absolute accuracy regard the deviation between the confidence and the performance of a task; these measures are useful when a researcher is interested in determining whether a specific format or treatment may enhance the calibration between the performance and confidence (Bol & Hacker, 2001; Bol, Hacker, O'Shea, & Allen, 2005). Among these types of measures, the Bias Index (refer to Formula 1 in the Method section) is computed in such a way as to evaluate the direction and the range of the lack of correlation between performance and confidence. It assesses over-confidence and under-confidence in the judgments (Jackson & Kleitman, 2013; Schraw, 2009). This index ranges from -1 to +1; a value greater than zero indicates overconfidence, whereas a value less than zero indicates underconfidence (Lichtenstein & Fischhoff, 1977). This index indicates whether the student was able to distinguish the tasks that, for him/her, are more difficult or easier, thereby producing judgments that reflect her/his effective level of performance in the tasks (Stankov & Crawford, 1997; Stankov et al., 2012).

Among the other measures of calibration, there are indices of the relative accuracy, which assess the consistency of confidence and performance across a set of tasks. In this regard, the Confidence-Judgment Accuracy Quotient (CAQ Index) (refer to Formula 2 in the Method section) assesses the individual's ability to distinguish between confidence for the correct items and the confidence for the incorrect items (Boekaerts & Rozendaal, 2010; Jackson & Kleitman, 2013; Schraw, 2009; Shaughnessy, 1979). The values of the CAQ may be positive when confidence in the correct answers to some items is higher than in the incorrect answers to other items. The values may be less than zero when confidence in the incorrect answers to some items is higher than in the correct answers to other items (Jackson & Kleitman, 2013).

To measure the calibration, Schraw (2009) recommends the use different indices that assess dissimilar aspects of the metacognitive processes. It has been established that the approach used to compute specific calibration measures may affect the findings of the investigation (Rutherford, 2017). Specifically, the indices of Absolute Accuracy are useful to evaluate the exactness of judgments, whereas, the indices of Relative Accuracy are useful to evaluate the correspondence between performance and the confidence judgments (Alexander, 2013; Jackson & Kleitman, 2013; Schraw, 2009). Consequently, the two types of calibration indices described above (Bias and CAQ Indices – which assess Absolute and Relative accuracy) are both useful. It may be shown that they are statistically independent (an individual could have a discrete absolute accuracy and a low relative accuracy; on the other hand, it is possible to observe the opposite) (Dunlosky & Thiede, 2013; Schraw, 2009).

Aims

The purpose of this research is to investigate the potential differences in the confidence and calibration between confidence judgments and performance for probabilistic problems presented in two formats (verbal-numerical and graphical-pictorial formats, shortly denoted, respectively, as N format and G format) in psychology undergraduates. The participants were asked to rate their confidence in the correctness of their answer to the previously provided task. We are interested in retrospective judgments expressed item-by-item. The domain assessed (the probabilistic reasoning) was unfamiliar to the participants, because they did not have specific knowledge regarding it.

This research aims to assess whether the use of graphical-pictorial representations improves the confidence and the calibration between confidence and accuracy in probabilistic reasoning (Garcia-Retamero & Cokely, 2017). In particular, we aim to assess whether the graphical-pictorial representations affect data encoding, thereby supporting the calibration of confidence and reducing overconfidence or underconfidence (Stankov & Crawford, 1996; Thompson, Prowse Turner, & Pennycook, 2011). Specifically, we may suppose that graphical-pictorial representations are useful to first increase confidence in one's performance and thus the calibration between confidence and accuracy. We also suppose that this finding may be more plausible for students with low numeracy skills (Garcia-Retamero & Cokely, 2013) and high visuospatial abilities (Kellen et al., 2013).

These relationships will be investigated by controlling the effects of dimensions related to individual differences (e.g., numerical and visuospatial abilities, attitudes towards statistics and statistical anxiety) and contextual aspects (e.g., the presence vs. absence of time pressure) (Beilock et al., 2004; Evans, Handley, & Bacon, 2009).

To investigate these aspects, following Schraw's advice (2009), two measures have been computed to assess the calibration between confidence and accuracy: one index of the Absolute Accuracy (the Bias index) (Boekaerts & Rozendaal, 2010; Was, 2014) and one index of the Relative Accuracy (Confidence-Judgment Accuracy Quotient; CAQ) (Jackson & Kleitman, 2013; Shaughnessy, 1979). All indices were computed on the set of items on probabilistic reasoning in the Numerical format and, separately, on the set of items in the Graphical format (Agus et al., 2015b).

Besides, these relationships were assessed disjointedly in two samples of psychology undergraduates without statistical expertise, who belonged to two countries in the European Higher Education Area: Italy and Spain.

To summarize, in the present study, we aim to investigate whether undergraduates may discern the cases in which they know the answer from the cases in which they do not know the answer. The questions of distinctive interest guiding the research are as follows:

- ✓ *Do students differ in confidence for Graphical versus Numerical format problems? Are undergraduates more confident in problems in the Graphical format than in the Numerical format?*
- ✓ *Are students differently calibrated in confidence in Numerical versus Graphical format problems?*
- ✓ *Do individual differences and time pressure affect confidence judgments? Do numerical and visuospatial abilities, statistical anxiety, attitudes towards statistics and time pressure affect the retrospective confidence judgments expressed in the two formats of problem presentation?*
- ✓ *Do individual differences and time pressure affect confidence judgments calibration?*
- ✓ *What is the effect of abilities (numerical and visuospatial), statistical anxiety, attitudes towards statistics and time pressure on the calibration in the two formats of problem presentation?*
- ✓ *Do the Bias Index and the CAQ index differ in assessment across the two formats? Are there differences in the relation to the two indices of calibration computed (the Bias Index and the CAQ Index)? Is the Bias Index significantly different from zero in the two formats (when the value zero indicates a good calibration between confidence and accuracy)? Is the CAQ significantly different from zero in the two formats (when the positive value indicates higher confidence in the correct items and the negative value indicates higher confidence in the incorrect items)?*

All these dimensions were examined separately for Italian and Spanish undergraduates in psychology; indeed, we assume that Italian undergraduates' calibration indices might differ from the Spanish ones in conditions of time pressure. In fact, in the literature it has been suggested that Italian students show better performances in probabilistic reasoning in presence of time pressure, when compared to Spanish students (Agus et al., 2015b).

METHOD

Participants

Two hundred fifty-seven first year Psychology students (28.80% males; age M= 19.76 years, SD = 3.48, age range = 17-52), who were recruited by a non-probabilistic sampling from Universities in Italy (Milan n=82; Rome n=66) and Spain (Barcelona n=109), participated in this study. All subjects voluntarily contributed; they did not have an incentive for participation.

Procedure

The protocol was completed by the undergraduates in paper-and-pencil format. Only the undergraduates who fulfilled all sections of the research protocol were included in the sample.

The administration occurred in quiet rooms in large groups, and the students completed the protocol in one work session. The participants differed in relation to the presence vs absence of time pressure. The session lasted approximately 50 minutes for the participants who worked in the time pressure condition (n=112, 43.6%). The participants who worked without time pressure did not have a time constriction (n=145, 56.4%).

Every participant completed all questionnaires comprising the research protocol and both formats of items (Numerical and Graphical formats). Also, to present the items in different positions in the protocol, the administration was randomly structured, with the application of changed and reversed orders of presentation (NG, at first Numerical then Graphical format, contrasted with GN, in the beginning Numerical then Graphical format) and sequences (1 and 2, the second overturned with respect to the first). For these reasons, in the protocol, there are four modalities of item administration (for the modality NG1 20.2%, NG2 26.5%, GN1 24.5% and GN2 28.8%).

Measures

The questionnaires were administered in a large battery, in which the first part assessed the demographic variables (age, gender, and curricula). The following sections evaluated the relevant dimensions identified in the research design.

Numerical and visuospatial abilities

The two starting sections of the protocol assessed the Numerical and Visuospatial abilities using the Intermediate Form of Primary Mental Abilities (PMA) (Thurstone & Thurstone, 1981, 1987). Specifically, these scales were selected because of their validity in the measurement of numerical abilities, as well as the detection of spatial relationships and spatial patterns in both adolescents and undergraduates (aspects strongly involved in the comprehension of relations embodied in the presentation of graphical devices) (Colom, Contreras, Botella, & Santacreu, 2002; Hegarty & Kozhevnikov, 1999).

Probabilistic reasoning

The succeeding sections assessed the basic probabilistic reasoning (in relation to simple and conditional probabilities) in the Numerical and Graphical formats; all items referred to the simple mathematical achievement reached during high school (Agus et al., 2016). This section incorporated five items; these related to the classical problems examined in the literature; specifically, one problem dealt with classical medical diagnoses (for similar problems see, for example Cosmides & Tooby, 1996; Sloman, Over, Slovak, & Stibel, 2003); one was about decks of cards (Tversky & Kahneman, 1974); one was about the outcomes of university examinations (Girotto & Gonzalez, 2001), one was about the roll of dices (Watson & Moritz, 2003); and one was about production defects in a factory (for an example see **Appendix**; for the questionnaire validation in Italy and Spain refer to (Agus et al., 2016)).

In the Graphical format, we used pictorial-graphical devices (such as tree diagrams and iconic drawings), similar to the devices reported in the classical studies conducted on these subjects (Cosmides & Tooby, 1996; Moro & Bodanza, 2010; Yamagishi, 2003). Each item in both formats included four closed response options (with only one correct option), followed by an open-ended question to explain the reasoning applied in the problem-solving.

We used a numerical and graphical pictorial format of the problems, expressed by frequencies, referring to classic works in the literature on this topic (Brase, 2009; Girotto & Gonzalez, 2001; Moro et al., 2011).

This assessment instrument was developed and validated in the Italian and Spanish versions (then each undergraduate fulfilled the protocol in his/her native language). To assess probabilistic reasoning, we summed the number of correct responses for both Numerical and Graphical scales.

The confidence in the correctness of the response previously provided was assessed using a Likert scale (from 1 = "not confident" to 5 = "completely confident") associated with each item (for an example see the **Appendix**) (Agus et al., 2016).

Attitude towards statistics

The Survey of Attitudes Toward Statistics (SATS-28) (Dauphinee, Schau, & Stevens, 1997) constituted the subsequent section of the protocol, which assessed four scales (Affect - six items; Difficulty - seven items; Cognitive Competence - six items; and Value - nine items) using 7-point Likert scales (from 1 = "strongly disagree", to 7 = "strongly agree"). For each participant, we administered the adapted version in his/her native language. The Cronbach's α were acceptable for both the Italian (.60; .81) and Spanish (.64; .90) versions (Carmona, Primi, & Chiesi, 2008; Chiesi & Primi, 2009).

Statistical anxiety

The final part of the protocol comprised the Statistical Anxiety Scale (SAS) (Chiesi et al., 2011; Vigil-Colet, Lorenzo-Seva, & Condon, 2008). This instrument included 24 items, assessed using a 5-point Likert scale. The scales investigated three dimensions: Examination, Interpretation and Asking for Help. The undergraduates completed the Italian or Spanish versions of the items. The scale exhibited good internal consistency (.85; .90 for the Italian version; .81; .92 for the Spanish version) (Chiesi et al., 2011; Vigil-Colet et al., 2008).

Data Analyses

The statistical data were analysed using R (version 3.6.0) and SPSS (version 22) software.

We independently analysed the samples of Italian and Spanish undergraduates, extracted from different populations. The protocols were assessed in relation to the potential effect of order and sequences, but no significant effect was identified in the Italian and Spanish universities (Agus et al., 2015a, 2015b).

We applied the calibration indices of the Absolute (Bias Index) (see Formula 1) and Relative Accuracy (CAQ) (see Formula 2) as defined in literature (e.g., Boekaerts & Rozendaal, 2010). The computation was carried out referring to the following formulae [(1) and (2)].

Specifically in the Bias Index (Jackson & Kleitman, 2013; Schraw, 2009)

$$Bias\ index = \frac{1}{n}\sum_{i=1}^{n}(c_i - p_i) \tag{1}$$

c_i is the confidence rating for item i, and p_i is the accuracy of the answer to item i (scored 1 in the case of a correct response and 0 in the case of an incorrect response). To compute the Bias Index, for each item, we recoded the Likert Scale confidence ratings in a dummy variable: low confidence = 0 (from 1 – not at all confident – to 3 - moderately confident), high confidence = 1 (from 4 – very confident - to 5 – extremely confident) (Jackson & Kleitman, 2013; Mevel et al., 2014; Stupple, Ball, & Ellis, 2013).

In the CAQ Index (Boekaerts & Rozendaal, 2010; Schneider, 2011; Shaughnessy, 1979)

$$CAQ\ Index = \frac{\left(\sum \frac{c_{i\ correct}}{p} - \frac{\sum c_{i\ incorrect}}{q}\right)}{\sigma} \tag{2}$$

c_i correct is the confidence rating for the correct answer to the i_{th} item, p represents the number of items with correct responses, c_i incorrect is the confidence rating for the incorrect answer to the i_{th} item, q represents the number of items with incorrect responses, and σ is the standard deviation calculated across all confidence ratings (Jackson & Kleitman, 2013).

We computed the descriptive statistical indices related to the mean of the scales, the confidence scores and the calibration indices.

The coefficient of the Bivariate Correlation Pearson's *r* was applied to assess the linear relationships between the inquired dimensions.

To compare the confidence in the answers to the probabilistic problems presented in the Numerical and Graphical formats (assumed as dependent variables), we subsequently performed a Mixed Design Analysis of Covariance, controlling for the effects of other study variables (assumed as covariates - numerical and visuospatial abilities, anxiety and attitudes); the variable "presence vs absence of time pressure" distinguished between two groups of independent observations. The values of partial Eta Squared ($p\eta^2$) were applied to evaluate the effect size of specific dimensions in our dependent variables, partialising the effects of other factors and of interactions (Cohen, 1973; Pierce, Block, & Aguinis, 2004; Richardson, 2011). The rules of thumb suggested by Ferguson (2009) reported

Table 1. Descriptive statistics

	Spain n=109					Italy n=128				
	Min	**Max**	**Mdn**	**Mean**	**sd**	**Min**	**Max**	**Mdn**	**Mean**	**sd**
FN correct responses	1	4	2.000	1.972	1.022	0	5	2.000	2.608	1.353
FG correct responses	1	4	2.000	2.330	1.054	1	5	3.000	3.027	1.142
PMA Visuospatial scale	3	54	26.000	25.834	11.184	0	53	22.000	21.135	10.151
PMA Numerical scale	5	35	16.000	16.752	5.899	7	40	18.000	17.705	5.805
SAS Examination	14	40	33.000	31.981	6.210	3	40	33.000	31.812	7.824
SAS Interpretation	8	32	18.000	18.102	5.432	2	32	17.000	16.210	5.718
SAS Help	8	33	17.000	16.981	7.385	1	40	17.000	16.766	7.309
SATS Affect	11	35	22.000	21.862	4.967	7	31	20.000	19.224	5.578
SATS Competence	15	39	28.000	27.816	4.647	11	36	26.000	25.197	5.248
SATS Value	27	61	48.000	47.642	8.192	21	63	45.000	44.619	8.264
SATS Difficulty	13	43	27.000	26.669	5.622	13	36	26.000	25.115	5.253
FN Confidence	1.50	4.80	3.400	3.302	.753	1.00	4.80	3.200	3.109	.912
FG Confidence	1.33	4.80	3.750	3.550	.820	1.20	4.80	3.500	3.229	.904
Bias FN	-.80	.80	.000	.040	.307	-1.00	.80	-.200	-.156	.341
Bias FG	-.60	.80	.000	.058	.308	-1.00	.40	.000	-.187	.358
CAQ FN	-1.39	4.33	1.118	1.089	1.068	-2.24	4.33	.775	.801	1.140
CAQ FG	-1.83	4.62	1.065	.9519	1.262	-1.22	6.00	1.593	1.258	1.241

Note: FN correct responses = Correct responses in numerical format of problem. FG correct responses = Correct responses in graphical format of problem; PMA= Primary Mental Abilities Questionnaire; SAS = Statistical Anxiety Scale; SATS = Survey of Attitudes toward Statistics; Bias FN = Bias Index for numerical format; Bias FG = Bias Index for graphical format; CAQ FN = Confidence-Judgment Accuracy Quotient in numerical format; CAQ FG = Confidence-Judgment Accuracy Quotient in graphical format; SD = Standard Deviation; Mdn = median

a limit value of .04 for a small effect, .25 for a moderate effect and .64 for a strong effect. The assumptions for the application of these analyses (univariate normality, homogeneity of inter-correlations and homogeneity of variance) were verified and met in the samples (Tabachnick & Fidell, 1996).

We subsequently performed other Mixed Design ANCOVAs to compare the indices of the calibration in the Numerical and Graphical formats (used as dependent variables), controlling for the effects of the dimensions described above (abilities, attitudes and statistical anxiety) and the influence of the time constraint. These analyses were applied separately for the Bias indices and the CAQ Indices. In these analyses, the assumptions were met in the samples (Tabachnick & Fidell, 1996) and the partial Eta Squared was considered in relation to the significant effects (Richardson, 2011).

Finally, to evaluate whether the indices of the calibration (Bias and CAQ Indices) were significantly different from the zero value, Student's t tests were computed in relation to both samples, separately for the N and G formats of problem presentation and distinctively for the subjects working with and without time pressure. In these comparisons the Cohen's d was computed to evaluate the effect sizes of significant differences highlighted (Cohen, 1977). These analyses were useful to determine the presence of underconfidence or overconfidence in the set of problems in Numerical and Graphical formats (concerning the Bias Index). Additionally, they were performed to assess whether there is higher confidence in the correct vs incorrect performances (referring to the CAQ Index).

RESULTS

In order to examine the distributions of the variables in Italian and Spanish undergraduates, the descriptive statistics were computed separately for two countries (minimum, maximum, median, mean and standard deviation). **Table 1** presents these statistics for the variables included in this study (number of correct responses in probabilistic reasoning in two formats, confidence in the correctness of these responses, PMA dimensions, SAS and SATS scales, Bias and CAQ indices for both formats).

To assess the linear relationships between the calibration indices, the Pearson's r coefficient was computed on the values concerning all conditions (in the presence and absence of time pressure) (**Table 2**). The pattern of the correlations in the Italian and Spanish undergraduates showed the same trend for the Bias Index. In both samples, we identified a positive and significant correlation between the Bias Index computed for the N and G formats (r= .598, p<.01 for Italians, r=.387, p<.01 for Spanish) (**Table 2**). In contrast, we determined that the CAQ indices for the N and G formats did not exhibit significant correlations. Also, there is a negative linear relationship between the CAQ in the G format and the Bias Index for the N format (r=-.195, p<.05) only for the Italians.

Table 2. Pearson's r linear correlations between indices of calibration (above the diagonal, the values for Spanish undergraduates; below the diagonal, the values for Italians)

		1	2	3	4
1	Bias FN	1	.387**	-.049	.016
2	Bias FG	.598**	1	-.033	.147
3	CAQ FN	-.054	.032	1	-.076
4	CAQ FG	-.195*	-.061	.111	1

Note: p<.05*; p<.01**; Bias FN = Bias Index for numerical format; Bias FG = Bias Index for graphical format; CAQ FN = Confidence-Judgment Accuracy Quotient in numerical format; CAQ FG = Confidence-Judgment Accuracy Quotient in graphical format

Table 3. Results of mixed design Ancova (dependent variable – Confidence)

Sample	Source	Wilks' Lambda	F	p	Partial η^2
Spanish sample	Format	.994		.460	
	Format * SAS Examination	.998		.629	
	Format * SAS Interpretation	.993		.417	
	Format * SAS Help	.999		.739	
	Format * SATS Affect	.999		.743	
	Format * SATS Competence	.990		.324	
	Format * SATS Value	.994		.467	
	Format * SATS Difficulty	1.000		.957	
	Format * PMA Visuo-spatial scale	.990		.325	
	Format * PMA Numeric scale	.998		.646	
	Format* Time pressure	.980		.169	
	Significant Between-Subjects Effects				
	SATS Affect		6.054 (df=1;96)	.016*	.059
	PMA Numeric scale		4.147 (df=1;96)	.044*	.041
Italian sample	Format	.988		.230	
	Format * SAS Examination	.998		.635	
	Format * SAS Interpretation	1.000		.998	
	Format * SAS Help	.996		.487	
	Format * SATS Affect	.964	4.388 (df=1;118)	.038*	.036
	Format * SATS Competence	1.000		.951	
	Format * SATS Value	.951	6.105 (df=1;118)	.015*	.049
	Format * SATS Difficulty	.983		.159	
	Format * PMA Visuo-spatial scale	.999		.803	
	Format * PMA Numeric scale	1.000		.964	
	Format * Time pressure	.970		.059	
	Significant Between-Subjects Effects				
	SATS Affect		8.901 (df=1;118)	.003**	.070
	SATS Competence		8.419 (df=1;118)	.004**	.067
	PMA Numeric scale		6.185 (df=1;118)	.014*	.050

Note: **p < .01; *p < .05; Partial η2 = partial Eta Squared for significant differences; PMA= Primary Mental Abilities Questionnaire; SAS = Statistical Anxiety Scale; SATS = Survey of Attitudes toward Statistics

This finding indicates that in both samples, higher under confidence in the Numerical format is correlated with higher under confidence in the Graphical format. Moreover, higher over confidence in the Numerical format is correlated with higher over confidence in the Graphical format. In contrast, the CAQ indices for the Numerical and Graphical formats did not highlight significant linear correlations.

Do Students Differ in Confidence for Graphical versus Numerical Format Problems? Do Individual Differences and Time Pressure Affect Confidence Judgment?

To assess the effect of the problem format (Numerical and Graphical) on the confidence in probabilistic reasoning, an Analysis of Covariance with a Mixed Design was conducted (separately for the Italian and Spanish undergraduates), in which the means of the confidence in the correctness of the problems in the Numerical and Graphical formats were used as repeated measures. The presence vs absence of time pressure was used as a between factor. The scales of the numerical and visuospatial abilities, statistical anxiety and attitudes towards statistics were assumed as covariates (**Table 3**).

In the Spanish undergraduates, there was no main effect of the repeated measures (Wilks' Lambda = .994, p= .460). Moreover, none of the covariates exerted a significant effect on the mean confidence in the correctness of the

Table 4. Results of mixed design Ancova (dependent variable – Bias Index)

Sample	Source	Wilks' Lambda	F	p	Partial η^2
Spanish sample	Format	.992		.384	
	Format * SAS Examination	.982		.189	
	Format * SAS Interpretation	.993		.397	
	Format * SAS Help	1.000		.932	
	Format * SATS Affect	.995		.469	
	Format * SATS Competence	1.000		.894	
	Format * SATS Value	.992		.393	
	Format * SATS Difficulty	1.000		.908	
	Format * PMA Visuo-spatial scale	.989		.305	
	Format * PMA Numeric scale	.986		.241	
	Format* Time pressure	.999		.751	
	Significant Between-Subjects Effects				
	SATS Affect		4.302 (df=1;96)	.041*	.043
Italian sample	Format	.998		.649	
	Format * SAS Examination	.998		.622	
	Format * SAS Interpretation	.994		.396	
	Format * SAS Help	.999		.746	
	Format * SATS Affect	.976		.089	
	Format * SATS Competence	.996		.509	
	Format * SATS Value	.974		.080	
	Format * SATS Difficulty	.997		.528	
	Format * PMA Visuo-spatial scale	.990		.286	
	Format * PMA Numeric scale	.985		.184	
	Format* Time pressure	.942		.008**	.050
	Significant Between-Subjects Effects				
	Time pressure		7.263 (df=1;118)	.008**	.058

Note: **p < .01; *p < .05; Partial η^2 = partial Eta Squared for significant differences; PMA= Primary Mental Abilities Questionnaire; SAS = Statistical Anxiety Scale; SATS = Survey of Attitudes toward Statistics

responses in the N and G formats (**Table 3**). We could observe small significant values of the Between Subjects Effect for the Sats Affect ($F_{(1;96)}$=6.054, p=.016, partial η^2=.059), and PMA Numerical scale ($F_{(1;96)}$=4.147, p=.044, partial η_2=.041).

In the Italian sample, there was no main effect of the within factor (Wilks' Lambda = .988, p= .230). However, there were small significant effects of the interaction Format*Sats Affect (Wilks' Lambda = .964; $F_{(1;118)}$ = 4.388; p= .038; partial η^2=.036) and Format*Sats Value (Wilks' Lambda = .951; $F_{(1;118)}$ =6.105; p= .015; partial η^2= .049); the other effects were not significant. We identified small significant values of the Between Subjects Effect for the Sats Affect ($F_{(1;118)}$=8.901, p=.003, partial η^2=.070), Sats Competence ($F_{(1;118)}$=8.419, p=.004, partial η^2=.067), and PMA Numerical ($F_{(1;118)}$=6.185, p=.014, partial η^2=.050).

These data indicated that there are no significant differences in the confidence expressed in the answers to problems in the Numerical and Graphical formats for both samples of Spanish and Italian undergraduates; however, there is a small effect of attitudes towards statistics and numerical abilities.

Are Students Differently Calibrated in Confidence in Numerical versus Graphical Format Problems? Do Individual Differences and Time Pressure Affect Calibration?

To further investigate these aspects, we assessed whether there were differences in the calibration in the Numerical and Graphical formats. Mixed Design ANCOVAs were conducted separately for the Italian and Spanish students.

The first Mixed ANCOVA analysis was performed using the Bias Index in the Numerical and Graphical formats as the repeated measures; the presence vs absence of time pressure was used as a between factor; the scales of the numerical and visuospatial abilities, statistical anxiety and attitudes towards statistics were assumed as covariates (**Table 4**). The second Mixed ANCOVA was conducted using the CAQ Index in the N and G formats as the repeated measures; the presence vs absence of time pressure was used as a between factor; the scales of the numerical and visuospatial abilities, statistical anxiety and attitudes towards statistics were assumed as covariates (**Table 5**).

Table 5. Results of the mixed design Ancova (dependent variable – CAQ)

Sample	Source	Wilks' Lambda	F	p	Partial η^2
Spanish sample	Format	.999		.763	
	Format * SAS Examination	.997		.717	
	Format * SAS Interpretation	.997		.577	
	Format * SAS Help	.996		.556	
	Format * SATS Affect	.999		.726	
	Format * SATS Competence	.999		.783	
	Format * SATS Value	.998		.652	
	Format * SATS Difficulty	.990		.338	
	Format * PMA Visuo-spatial	.997		.609	
	Format * PMA Numeric	1.000		.887	
	Format* Time pressure	.965		.067	
	Significant Between-Subjects Effects				
	Time pressure		8.321 (df=1;96)	.005**	.080
Italian sample	Format	.991		.401	
	Format * SAS Examination	.993		.473	
	Format * SAS Interpretation	.999		.800	
	Format * SAS Help	.963		.095	
	Format * SATS Affect	.998		.701	
	Format * SATS Competence	1.000		.957	
	Format * SATS Value	.999		.750	
	Format * SATS Difficulty	.991		.403	
	Format * PMA Visuo-spatial	.999		.803	
	Format * PMA Numeric	1.000		.945	
	Format* Time pressure	.998		.671	

Note: **p < .01; *p < .05; Partial η2 = partial Eta Squared for significant differences; PMA= Primary Mental Abilities Questionnaire; SAS = Statistical Anxiety Scale; SATS = Survey of Attitudes toward Statistics

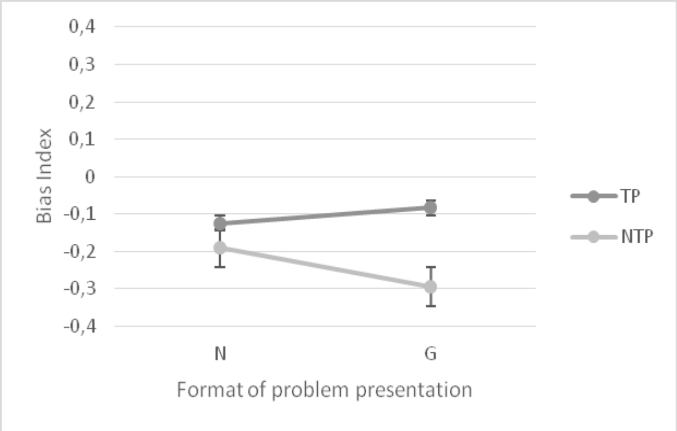


Figure 1. Estimated marginal means for the Bias Index in relation to the significant effect of Format*Time pressure in Italian undergraduates

Note: TP Time pressure; NTP No Time Pressure; N Verbal numerical format; G Graphical pictorial format

Regarding the Bias Index, in the Spanish undergraduates, there was no main effect of the repeated measures (Wilks' Lambda = .992, p= .384) (**Table 4**). Also, the covariates did not exert a significant effect on the mean confidence in the correctness of the responses in the Numerical and Graphical formats (**Table 4**). The scale of SATS Affect has a significant effect for Between Subject test ($F_{(1;96)}$ = 4.302, p= .041; partial η^2= .043).

In the Italian sample, there was no main effect of the within factor (Wilks' Lambda = .998, p= .649) or significant effects of the interaction among the within factor and the covariates. It is highlighted that there is a small significant effect of interaction of Format * Time pressure (Wilks' Lambda = .942, p= .008; partial η^2= .050) (see **Figure 1**). Time pressure also has a significant effect for Between Subject test ($F_{(1;118)}$ = 7.263, p= .008; partial η^2= .058).

In relation to the application of the same analyses for the CAQ Index, in the Spanish sample, there was no significant effect for the within factor test (Wilks' Lambda = .999, p= .763) or the covariates (**Table 5**). The time pressure exerts a significant effect in the Between Subjects test ($F_{(1;96)}$ = 8.321; p= .005; partial η^2 = .080). In the Italian

Table 6. Student's t Test to assess the difference in the calibration indices from zero

	BIAS FN						**BIAS FG**					
	No TP mean ± sd	**t (df) p Cohen's d**	**TP mean ± sd**	**t (df) p Cohen's d**	**Total mean ± sd**	**t (df) p Cohen's d**	**No TP mean ± sd**	**t (df) p Cohen's d**	**TP mean ± sd**	**t (df) p Cohen's d**	**Total mean ± sd**	**t (df) p Cohen's d**
ES	.063 ± .339	1.489 (62) p =.142	.009 ± .260	.227 (45) p =.821	.040 ± .308	1.370 (108) p=.174	.073 ± .305	1.897 (62) p=.062	.039 ± .314	.844 (45) p=.403	.059 ± .308	1.988 (108) p=.049* .190
IT	-.192 ±.376	-4.6094 (81) p=.001** .602	-.112 ± .290	-3.146 (65) p=.002* .387	-.156 ± .341	-5.559 (147) p=.0001** .306	-.312 ± .355	-7.948 (81) p=.0001* .769	-.033 ± .299	-.906 (65) p=.368	-.187 ± .358	-6.373 (147) p=.0001** .530
	CAQ FN						**CAQ FG**					
ES	.779 ±.884	6.992 (62) p=.001** .881	1.515 ± 1.158	8.871 (45) p=.001** 1.308	1.089 ± 1.068	10.648 (108) p=.001** 1.020	.880 ± 1.071	6.517 (62) p=001** .821	1.051 ± 1.493	4.774 (45) p=.001** .703	.952 ± 1.263	7.870 (108) p=.001** .754
IT	.702 ±1.139	4.445 (81) p=.0001** .405	.880 ± 1.144	6.247 (65) p=.001** .769	.805 ± 1.140	7.634 (147) p=.0001** .708	1.374 ± 1.258	8.390 (81) p=.0001** .850	1.154 ± 1.227	7.638 (65) p=.001** .940	1.258 ± 1.241	11.325 (147) p=.0001** 1.084

Note: p<.05* p<.01**; No TP= absence of time pressure; TP=presence of time pressure; IT = Italian ; ES = Spanish; sd=standard deviation; t= Student's t test; df= degrees of freedom; Bias FN = Bias Index for numerical format; Bias FG = Bias Index for graphical format; CAQ FN = Confidence-Judgment Accuracy Quotient in numerical format; CAQ FG = Confidence-Judgment Accuracy Quotient in graphical format; Cohen's d = Cohen's d Effect size for significant differences

undergraduates for the CAQ Indices, we note that there are no significant effects of Format; no further significant effects were identified.

In summary, for the Italian students, the absolute index of Bias is affected by the interaction concerning the presence of time pressure and the format. However, the data suggested that for the Spanish undergraduates, there were no significant differences in the Bias expressed in the N and G formats. Thus, we speculate that there are no differences in the evaluation of the trend or the range of the gap between performance and confidence (i.e., overconfidence versus underconfidence in the judgments) for the Spanish students. In contrast, the Italian undergraduates exhibited a dissimilar ability to distinguish between simple and complex tasks in the N and G formats (see **Figure 1**). Specifically, in the N format, the Bias index is similar both in the presence and absence of time pressure (together characterized by underconfidence). For the G format, in the absence of time pressure, we may observe a stronger underconfidence (lower level of bias index) than in the presence of time constraints. For the G format in time pressure, the Italians clearly showed a Bias index close to zero, which indicates a better calibration between performance and confidence.

In relation to the CAQ Index, defined as the assessment of individual ability to distinguish between the confidence for the correct items and the confidence for the incorrect items, the data suggested that there are no significant differences in the indices in the N and G formats for both samples. All undergraduates exhibited higher levels of confidence in relation to the problems solved correctly (CAQ > 0).

Do the Bias Index and the CAQ Index Differ in Assessment across the Two Formats? Are there Differences in the Relation to the Two Indices of Calibration Computed?

To assess the goodness of the calibration between confidence and accuracy in probabilistic reasoning in two formats (N and G), we applied Student's t tests for each sample (Italian and Spanish undergraduates). This assessment was conducted for the calibration indices of the Bias and CAQ in both formats, as well as distinguishing between the indices concerning the administrations in conditions of the presence versus absence of time pressure. These analyses enabled us to account for the significant effect of the time pressure previously identified in this study.

In **Table 6**, it is highlighted that for the Spanish undergraduates, there is a good calibration in the N format, whereas in the G format, there is significant overconfidence for the total sample (working with and without time pressure). In contrast, for the Italian students, there is significant underconfidence in both formats of problem presentation.

It is highlighted that students (both Italian and Spanish) report CAQ values significantly higher than zero, which indicates greater confidence in the correct items, both for the Numerical and Graphical formats in the presence and absence of time pressure (**Table 6**). They appear to exhibit good levels of discrimination between incorrect and correct judgments, evidencing an appropriate adjustment of the confidence between the problems solved incorrectly and correctly. In relation to the effect of the time pressure identified for the Italians, at the descriptive level, it is interesting to observe that the CAQ Index in the Numerical format is higher in the presence of time pressure, whereas the CAQ Index in the Graphical format is higher in the absence of time pressure.

DISCUSSION AND CONCLUSIONS

The purpose of this study was to investigate whether undergraduates are distinctively confident and calibrated concerning their performance in probabilistic problem presentation in relation to different formats of problem presentation (verbal-numerical and graphical-pictorial).

Specifically, we posed research questions that focused on the level of confidence and two different indices of calibration between confidence and accuracy: one index of the Absolute Accuracy, the Bias Index, and one index of the Relative Accuracy, the CAQ. We compared these variables computed distinctively in relation to a set of items in Numerical and Graphical formats, controlling for the effects of individual dimensions (e.g., attitudes towards statistics, visuospatial and numerical abilities, and statistical anxiety) and a contextual dimension (e.g., time pressure). The study was conducted with Italian and Spanish Psychology undergraduates in relation to a domain (the probabilistic reasoning) in which the participants did not have explicit knowledge.

The relevance of the confidence and calibration in problem-solving has often been investigated in the literature (Jackson & Kleitman, 2014; Morony, Kleitman, Lee, & Stankov, 2013). Examining if students differ in confidence for Graphical versus Numerical format problems, our findings suggest that the level of confidence may be similar in relation to a specific type of problem (in this case, a probabilistic problem), regardless of the problem format (Numerical and Graphical). It is fascinating to observe that this confidence may be affected by the attitudes towards statistics and the numerical ability in both Italian and Spanish undergraduates. These data may support the hypothesis that the formats of problem presentation and time pressure did not exert an influence on the level of confidence of the correctness of the responses, which, to some extent, is related to the attitudes towards statistics and numerical abilities.

These outcomes are plausible in relation to the statements of several authors (Jackson & Kleitman, 2014), which affirm that each subject tends to apply a stable style in metacognitive confidence judgments and their calibration.

Likewise, these results may be understood in relation to the features of our psychology undergraduates, who did not have statistical and probabilistic expertise. The literature indicates the influence of previous knowledge in relation to the confidence and the calibration (Gutierrez & Schraw, 2015). Concerning this aspect, we note that in our samples, the absence of statistical and probabilistic knowledge has been used as an inclusion criterion in the study, defining a characteristic of our population. Besides, it was highlighted that problem-solving and confidence in solutions are based, in addition to previous knowledge, on solution strategies applied by the same subject in the past (Iannello, Perucca, Riva, Antonietti, & Pravettoni, 2015). The application of the previous solution strategies was conducted regardless of the potential discrepancies between the actual problem and the previous experiences. This finding may indicate that earlier experience, although not completely consistent with the proposed problems, may guide and affect probabilistic problem solving and confidence (Iannello et al., 2015; Riva, Monti, & Antonietti, 2011). It is possible to speculate that because our undergraduates have no specific education in probabilistic reasoning, they tend to apply the same response pattern, the same strategies of solution, in both formats of problem presentation (N and G). Moreover, these aspects are consistent with the finding that confidence is a good predictor of accuracy in problem-solving in different formats of problem presentation (Jackson & Kleitman, 2014; Stankov et al., 2012).

Regarding the question if the students are differently calibrated in confidence in Numerical versus Graphical format problems, if individual differences and time pressure affect confidence judgment and calibration, it was highlighted that the Italians and Spanish did not exhibit significant effects of the format and other inquired dimensions for the CAQ index. The CAQs in the N and G formats have values significantly higher than zero, which indicates higher confidence in the correct items in the presence and absence of time pressure. The undergraduates displayed well-meaning levels of discrimination, which demonstrates a notable adjustment of the confidence for the problems answered incorrectly and correctly.

Nevertheless, concerning the assessment of the Bias Index, we note several specificities. In particular, consistent with a previous work (Agus et al., 2015a, 2015b) in which it was highlighted that the performance in probabilistic reasoning was improved in Italian undergraduates working in time pressure, we speculate that the Bias of calibration improves in the presence of time pressure, specifically in the problems in the Graphical format. This effect is highlighted only for Italian undergraduates and not for Spanish undergraduates. The Italians exhibited strong responsiveness to the assessment situations in which there were time limits (Kleiner, 2014).

These findings may be related to many interacting dimensions. For example, regarding PISA mathematics achievement, Chiu and Xihua (2008) analysed the differences across 41 countries. The responses of thousands of students were assessed via multilevel statistical analyses, which highlighted fascinating differences in many countries, including Italy and Spain. The authors assessed family and motivation effects on mathematics achievement and indicated that these two countries have different scores in relation to the dimensions of "individualism" (higher for Italians) and "uncertainty avoidance" (higher for Spanish). The analyses also indicated other remarkable aspects that differentiated Italy and Spain. Specifically, the mathematical achievement for Italians is positively and significantly affected by the interest in mathematics, self-efficacy and self-concept. In contrast, for Spanish students, the positive and significant effect of the foreign language at home, the cultural communication, the effort and the perseverance are highlighted; besides, there is a negative effect on the math achievement exerted by grandparents and the number of siblings (Chiu & Xihua, 2008). Moreover, Lee (2009) compared the levels of math self-concept, math self-efficacy and math anxiety and identified underlying differences across the 41 countries that participated in the PISA assessment. In particular, the Spanish students exhibited lower math anxiety than the Italians. We speculate that these differences may affect student performance in probabilistic problems, accounting for several differences between these two countries in their responsiveness to the presence of time limits.

In relation to our data, we suppose that the probabilistic reasoning and calibration of Italians may be significantly and positively affected by the presence of time pressure, which appears to improve the performance and enhance the confidence calibration (reducing the values of the Bias Index).

These discoveries may be consistent with the Distraction Theory (Beilock & Carr, 2005) and the research conducted by Markman et al. (2006). These authors indicated that a performance decline in mathematical problems under a time pressure condition (referred to as "choking") may be related to interference in the application of explicit solution strategies in problem-solving. Markman et al. (2006) emphasized that a time pressure condition could induce a decline in the performance related to the application of a learned rule; alternatively, the time constraint could enhance the performance in the solution of problems that require the application of a holistic information integration strategy. The authors related these findings to working memory, which is overloaded in a time pressure condition and induces the individual to apply an information-integration strategy rather than hypothesis testing strategies (Markman et al., 2006). Additionally, consistent with these aspects, the research conducted by Maddox et al. (Maddox, Ashby, Ing, & Pickering, 2004) focused on the different categorization processes applied in learning and problem-solving situations. They highlighted that the explicit hypothesis-testing system is presumed to control the learning of rule-based category learning tasks, whereas the implicit procedural learning system controls the learning of information-integration category learning tasks. These different processes may be differentially affected by the presence of time pressure, in relation to the different roles of working memory and the dissimilar application of strategies (Markman et al., 2006).

At this point, it may be useful to consider that in our research, the undergraduates did not have statistical knowledge; they did not master the correct strategies to solve the probabilistic problems presented. This aspect may be considered crucial to understanding the specificity of these findings. In problem-solving, they must refer to the basic strategies learned in the study of math in high school. For this reason, we speculate that in the probabilistic problem-solving, they had to apply a strategy only partially appropriate to the problem. Following this approach, we may observe that in the Graphical format, under time pressure, the subject may be supported in these processes and thus obtain better performances and exhibit better calibration between confidence and performance (Bias Indices close to zero).

Gimmig et al. (Gimmig, Huguet, & Caverni, 2006) suggested that time pressure may overload the working memory and reduce the fluid reasoning abilities. This effect may be related to the subjective meaning given to the assessment and the task, thereby demonstrating a strong variability in relation to the cultural contexts, the domains and the personal relevance of the problem (Gimmig et al., 2006). Beyond the common identified sources of time pressure (both in the familial and working contexts), in the literature, many other dimensions have been identified that may powerfully affect the individual perception of time pressure (i.e., emotional and cognitive aspects, socio-economic status, cultural and social meanings given to an activity) (Kleiner, 2014).

Based on our data, we speculate that there is a difference in the perception and the meaning given to the time pressure in Spain and Italy. Our findings highlighted a specific sensitivity of Italian undergraduates to the presence of time pressure, which the Spanish students did not exhibit in a similar manner. These findings confirm the potential existing differences in Italian and Spanish students, yet sustained in literature (Agasisti & Cordero-Ferrera, 2013; Agasisti & Pérez-Esparrells, 2010; Agus et al., 2019).

These outcomes may also be underscored in relation to the classical models of metacognitive monitoring processes, which embrace the relevant roles of the interactions between social aspects, affect, external conditions and attitudes (Efklides, 2008).

The dimension of time pressure is considered relevant in the Italian context. Namely, a previous study identified a "graphical facilitation effect" only in the time pressure condition for this population (Agus et al., 2015b). Specifically, it was observed that the presence of time pressure might enhance the engagement on a task and the application of functional solution strategies, thereby reducing the application of dysfunctional strategies, particularly in the Graphical format. Many authors in the literature highlighted the so-called "effect of graphical facilitation" in relation to the accuracy in probabilistic reasoning (Brase, 2009; Brase & Hill, 2015). The specificities of this effect require further investigations to understand the aspects implied. Additionally, these outcomes appear to highlight a new perspective in the description and comparison of the features of probabilistic reasoning in verbal-numerical and graphical pictorial formats.

Nevertheless, our findings require supplementary investigations to further assess these aspects and overcome several limitations. The high number of variables included in the analyses could reduce the power of the statistical results. It may be useful to conduct the same analyses with larger samples of students, to confirm the relationships identified between the variables for Italian and Spanish undergraduates. Moreover, it may be interesting to assess the interaction effect between country and time pressure, to deepen the differences exhibited by these undergraduates in probabilistic reasoning. Still, it may be interesting to deepen the role of time pressure by administering probabilistic problems in other real-world circumstances, in which the time constraint is applied in different ways. It is also necessary to highlight that the present study could not consider the effects of other psychological dimensions, which may affect confidence and calibration (for example, the cognitive styles). Another limitation is related to the generalizability of these findings, because of the features of the examined subjects (Psychology undergraduates). The current findings deserve further investigation in future studies to clarify these features.

The influence of the previously described social and cultural aspects may be observed in relation to the values of the Bias Index in the two countries. We speculate that Italians, who are more sensitive to the presence of time pressure, appear to enhance their accuracy and reduce the bias between accuracy and confidence (Bias of calibration) in the presence of time constraints. This improvement appears specifically in the graphical pictorial format, when the solution may likely be related to the application of different solution strategies of probabilistic problems with respect to the Numerical format.

In summary, in this work it was highlighted that:

- the levels of confidence in the correctness of response are analogous in N and G formats, both for Italians and Spaniards; furthermore, confidence is partially affected by numerical abilities and attitudes;
- in Spanish undergraduates significant differences in the calibration (Bias index) are not observed in N and G formats; in Italians, the calibration between correctness and accuracy (Bias index) is significantly but weakly affected by the interaction *Format * Time pressure*; specifically, the calibration is similar in the N format in presence/absence of time pressure (when there is a significant underconfidence), but for the G format in time pressure there is a better calibration between performance and confidence. Regarding the CAQ index, for both countries, there are not significant differences in N and G formats, in which the students showed always higher confidence in the problems solved appropriately;
- for the Spanish students, there is a virtuous calibration in the N format; in the G format, there is significant overconfidence; in contrast, for Italians there is significant underconfidence in both formats of problem presentation.

Our data raise several inspiring questions for upcoming research. This study extends findings in relation to the key roles of confidence and calibration in probabilistic problem solving, which have to be assessed in relation to the presence vs. absence of time pressure. The findings provide a point of reflection in relation to the effect exerted by time constraint in probabilistic problem solving and the monitoring of cognitive processes. Likewise, considering the limitations of our work, these results may provide useful suggestions for individuals interested in designing better learning strategies suited to the domain of probabilistic problems.

REFERENCES

Agasisti, T., & Cordero-Ferrera, J. M. (2013). Educational disparities across regions: A multilevel analysis for Italy and Spain. *Journal of Policy Modeling*, *35*(6), 1079-1102. https://doi.org/10.1016/j.jpolmod.2013.07.002

Agasisti, T., & Pérez-Esparrells, C. (2010). Comparing efficiency in a cross-country perspective: the case of Italian and Spanish state universities. *Higher Education*, *59*(1), 85-103. https://doi.org/10.1007/s10734-009-9235-8

Agus, M., Penna, M. P., Peró-Cebollero, M., & Guàrdia-Olmos, J. (2016). Assessing Probabilistic Reasoning in Verbal-Numerical and Graphical-Pictorial Formats: An Evaluation of the Psychometric Properties of an Instrument. *Eurasia Journal of Mathematics, Science & Technology Education*, *12*(8), 2013–2038. https://doi.org/10.12973/eurasia.2016.1265a

Agus, M., Peró-Cebollero, M., Guàrdia-Olmos, J., Pessa, E., Figus, R., & Penna, M. (2019). A Comparison of Probabilistic Reasoning in Psychology Undergraduates in Italy and Spain: Seeking Cross-national Evidence. *Eurasia Journal of Mathematics, Science and Technology Education*, *15*(10). https://doi.org/10.29333/ejmste/106232

Agus, M., Peró-Cebollero, M., Penna, M. P., & Guàrdia-Olmos, J. (2015a). Towards the development of problems comparing verbal-numerical and graphical formats in statistical reasoning. *Quality and Quantity*, *49*(2), 691–709. https://doi.org/10.1007/s11135-014-0018-7

Agus, M., Peró-Cebollero, M., Penna, M. P., & Guàrdia-Olmos, J. (2015b). 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*(5), 735–750. https://doi.org/10.12973/eurasia.2015.1382a

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction*, *24*(1), 1-3. https://doi.org/10.1016/j.learninstruc.2012.10.003

Ayal, S., & Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgment and Decision Making*, *9*(3), 226-242.

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "Choking under pressure" in math. *Psychological Science*, *16*(2), 101-105. https://doi.org/10.1111/j.0956-7976.2005.00789.x

Beilock, S. L., Kulp, C. A., Holt, L. E., & Carr, T. H. (2004). More on the Fragility of Performance: Choking Under Pressure in Mathematical Problem Solving. *Journal of Experimental Psychology: General*, *133*(4), 584-600. https://doi.org/10.1037/0096-3445.133.4.584

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, *7*(2), 161-186. https://doi.org/10.1016/S0959-4752(96)00015-1

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction*, *20*(5), 372-382. https://doi.org/10.1016/j.learninstruc.2009.03.002

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *Journal of Experimental Education*, *69*(2), 133–151. https://doi.org/10.1080/00220970109600653

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The Influence of Overt Practice, Achievement Level, and Explanatory Style on Calibration Accuracy and Performance. *Journal of Experimental Education*, *73*(4), 269-290. https://doi.org/10.3200/JEXE.73.4.269-290

Brase, G. L. (2009). How different types of participant payments alter task performance. *Judgment and Decision Making*, *4*(5), 419-428. https://doi.org/10.1111/j.1559-1816.1997.tb00644.x

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

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. https://doi.org/10.3389/fpsyg.2015.00340

Buratti, S., & Allwood, C. M. (2015). Metacognition: Fundaments, Applications, and Trends. In A. Peña-Ayala (Ed.), *Intelligent Systems Reference Library*. https://doi.org/10.1007/978-3-319-11062-2

Carmona, J., Primi, C., & Chiesi, F. (2008). Testing for measurement invariance of the Survey of Attitudes Toward Statistics: A comparison of Italian and Spanish students. *III European Congress of Methodology, Oviedo, Spain*.

Chiesi, F., & Primi, C. (2009). Assessing statistics attitudes among college students: Psychometric properties of the Italian version of the Survey of Attitudes toward Statistics (SATS). *Learning and Individual Differences*, *19*(2), 309-313. https://doi.org/10.1016/j.lindif.2008.10.008

Chiesi, F., Primi, C., & Carmona, J. (2011). Measuring Statistics Anxiety. Cross-Country Validity of the Statistical Anxiety Scale (SAS). *Journal of Psychoeducational Assessment*, *29*(6), 559-569. https://doi.org/10.1177/0734282911404985

Chiu, M. M., & Xihua, Z. (2008). Family and motivation effects on mathematics achievement: Analyses of students in 41 countries. *Learning and Instruction*, *18*(4), 321-336. https://doi.org/10.1016/j.learninstruc.2007.06.003

Cohen, J. (1973). Eta-Squared and Partial Eta-Squared in Fixed Factor Anova Designs. *Educational and Psychological Measurement*, *33*(1), 107-112. https://doi.org/10.1177/001316447303300111

Cohen, J. (1977). Statistical power analysis for the behavioral sciences, Rev. ed. In *Statistical power analysis for the behavioral sciences, Rev. ed.* Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Cokely, E., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, *4*(1), 20-33.

Colom, R., Contreras, M. J., Botella, J., & Santacreu, J. (2002). Vehicles of spatial ability. *Personality and Individual Differences*, *32*(5), 903-912. https://doi.org/10.1016/S0191-8869(01)00095-2

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. https://doi.org/10.1016/0010-0277(95)00664-8

Dauphinee, T. L., Schau, C., & Stevens, J. J. (1997). Survey of attitudes toward statistics: Factor structure and factorial invariance for women and men. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 129-141. https://doi.org/10.1080/10705519709540066

DeCaro, M. S., Thomas, R. D., Albert, N. B., & Beilock, S. L. (2011). Choking under pressure: Multiple routes to skill failure. *Journal of Experimental Psychology: General*, *140*(3), 390-406. https://doi.org/10.1037/a0023466

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction*, *24*(1), 4-14. https://doi.org/10.1016/j.learninstruc.2012.06.001

Dougherty, M. R., & Sprenger, A. (2006). The influence of improper sets of information on judgment: how irrelevant information can bias judged probability. *Journal of Experimental Psychology: General*, *135*(2), 262-281. https://doi.org/10.1037/0096-3445.135.2.262

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, *24*(1), 58-61. https://doi.org/10.1016/j.learninstruc.2012.05.002

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist*, *13*(4), 277-287. https://doi.org/10.1027/1016-9040.13.4.277

Evans, J. S. B. T., Handley, S. J., & Bacon, A. M. (2009). Reasoning Under Time Pressure. *Experimental Psychology*, *56*(2), 77-83. https://doi.org/10.1027/1618-3169.56.2.77

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532-538. https://doi.org/10.1037/a0015808

Frosch, C. A., & Johnson-Laird, P. N. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica*, *137*(3), 280-291. https://doi.org/10.1016/j.actpsy.2011.01.015

Gal, I., Garfield, J., & Gal, Y. (1997). *The assessment challenge in statistics education* (Vol. 12). IOS Press.

Garcia-Retamero, R., & Cokely, E. (2013). Communicating Health Risks With Visual Aids. *Current Directions in Psychological Science*, *22*(5), 392-399. https://doi.org/10.1177/0963721413491570

Garcia-Retamero, R., & Cokely, E. (2014). The Influence of Skills, Message Frame, and Visual Aids on Prevention of Sexually Transmitted Diseases. *Journal of Behavioral Decision Making*, *27*(2), 179-189. https://doi.org/10.1002/bdm.1797

Garcia-Retamero, R., & Cokely, E. (2017). Designing Visual Aids That Promote Risk Literacy: A Systematic Review of Health Research and Evidence-Based Design Heuristics. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(4), 582-627. https://doi.org/10.1177/0018720817690634

Garcia-Retamero, R., Cokely, E., & Hoffrage, U. (2015). Visual aids improve diagnostic inferences and metacognitive judgment calibration. *Frontiers in Psychology*, *6*, 932. https://doi.org/10.3389/fpsyg.2015.00932

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine (1982)*, *83*, 27-33. https://doi.org/10.1016/j.socscimed.2013.01.034

Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York: Basic books.

Ghazal, S., Cokely, E., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Making*, *9*(1), 15-34. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84893087076&partnerID=tZOtx3y1

Gimmig, D., Huguet, P., & Caverni, J.-P. (2006). Choking under pressure and working memory capacity: When performance pressure reduces fluid intelligence. *Psychonomic Bulletin & Review*, *13*(6), 1005-1010.

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

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition*, *15*(1), 84-93. https://doi.org/10.3758/BF03197714

Guàrdia-Olmos, J., Freixa, M., Peró, 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. https://doi.org/10.1007/s11135-005-2072-7

Gutierrez, A. P., & Schraw, G. (2015). Effects of Strategy Training and Incentives on Students' Performance, Confidence, and Calibration. *Journal of Experimental Education*, *83*(3), 386-404. https://doi.org/10.1080/00220973.2014.907230

Gutierrez, A. P., Schraw, G., Kuch, F., & Richmond, A. S. (2016). A two-process model of metacognitive monitoring: Evidence for general accuracy and error factors. *Learning and Instruction*, *44*, 1-10. https://doi.org/10.1016/j.learninstruc.2016.02.006

Hafenbrädl, S., & Hoffrage, U. (2015). Toward an ecological analysis of Bayesian inferences: how task characteristics influence responses. *Frontiers in Psychology*, *6*, 939. Retrieved from http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00939

Hanoch, Y., & Vitouch, O. (2004). When less is more information, emotional arousal and the ecological reframing of the Yerkes-Dodson law. *Theory & Psychology*, *14*(4), 427-452. https://doi.org/10.1177/0959354304044918

Hegarty, M., & Kozhevnikov, M. (1999). Types of visual–spatial representations and mathematical problem solving. *Journal of Educational Psychology*, *91*(4), 684-689. https://doi.org/10.1037/0022-0663.91.4.684

Iannello, P., Perucca, V., Riva, S., Antonietti, A., & Pravettoni, G. (2015). What do physicians believe about the way decisions are made? A pilot study on metacognitive knowledge in the medical context. *Europe's Journal of Psychology*, *11*(4), 691-706. https://doi.org/10.5964/ejop.v11i4.979

Jackson, S., & Kleitman, S. (2013). Individual differences in decision-making and confidence: Capturing decision tendencies in a fictitious medical test. *Metacognition and Learning*, *9*(1), 25-49. https://doi.org/10.1007/s11409-013-9110-y

Jackson, S., & Kleitman, S. (2014). Individual differences in metacognitive feelings of confidence: The generality and predictive validity of judgement confidence and its calibration in a medical decision-making task. *Personality and Individual Differences*, *60*(2014), S32. https://doi.org/10.1016/j.paid.2013.07.065

Jackson, S., Kleitman, S., Howie, P., & Stankov, L. (2016). Cognitive abilities, monitoring confidence, and control thresholds. Explain individual differences in heuristics and biases. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01559

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). New York: Springer.

Kleiner, S. (2014). Subjective time pressure: general or domain specific? *Social Science Research*, *47*, 108-120. https://doi.org/10.1016/j.ssresearch.2014.03.013

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*, *25*(1), 108-125. https://doi.org/10.1037/h0078792

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences*, *19*(3), 355-365. https://doi.org/10.1016/j.lindif.2008.10.009

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159-183. https://doi.org/10.1016/0030-5073(77)90001-0

Lin, L.-M., & Zabrucky, K. M. (1998). Calibration of Comprehension: Research and Implications for Education and Instruction. *Contemporary Educational Psychology*, *23*(4), 345-391. https://doi.org/10.1006/ceps.1998.0972

Lundeberg, M. A., Fox, P. W., & Punćochaŕ, J. (1994). Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgments. *Journal of Educational Psychology*, *86*(1), 114-121. https://doi.org/10.1037/0022-0663.86.1.114

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, *32*(4), 582-591. https://doi.org/10.3758/BF03195849

Maloney, E. A., Waechter, S., Risko, E. F., & Fugelsang, J. A. (2012). Reducing the sex difference in math anxiety: The role of spatial processing ability. *Learning and Individual Differences*, *22*(3), 380-384. https://doi.org/10.1016/j.lindif.2012.01.001

Markman, A. B., Maddox, W. T., & Worthy, D. A. (2006). Choking and excelling under pressure. *Psychological Science*, *17*(11), 944-948. https://doi.org/10.1111/j.1467-9280.2006.01809.x

Mevel, K., Poirel, N. N., Rossi, S., Cassotti, M., Simon, G. G., Houdé, O., & De Neys, W. (2014). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*(2), 227-237. https://doi.org/10.1080/20445911.2014.986487

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. Retrieved from http://www.scielo.org.ar/pdf/interd/v27n1/v27n1a11.pdf

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. https://doi.org/10.1080/20445911.2011.579072

Morony, S., Kleitman, S., Lee, Y. P., & Stankov, L. (2013). Predicting achievement: Confidence vs self-efficacy, anxiety, and self-concept in Confucian and European countries. *International Journal of Educational Research*, *58*, 79-96. https://doi.org/10.1016/j.ijer.2012.11.002

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research*, *95*(3), 131-142. https://doi.org/10.1080/00220670209596583

Okan, Y., Garcia-Retamero, R., Cokely, E., & Maldonado, A. (2015). Improving Risk Understanding Across Ability Levels: Encouraging Active Processing With Dynamic Icon Arrays. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000045

Onwuegbuzie, A. J. (1995). Statistics test anxiety and female students. *Psychology of Women Quarterly*, *19*(3), 413-418. https://doi.org/10.1111/j.1471-6402.1995.tb00083.x

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary Note on Reporting Eta-Squared Values from Multifactor ANOVA Designs. *Educational and Psychological Measurement*, *64*(6), 916-924. https://doi.org/10.1177/0013164404264848

Pintrich, P. R. (2000). *The role of goal orientation in self-regulated learning.* Academic Press.

Primi, C., & Chiesi, F. (2016). Statistics anxiety: A mediator in learning probability. *13th International Congress on Mathematical Education*, 1-7. Hamburg, July 24-31, 2016.

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135-147. https://doi.org/10.1016/j.edurev.2010.12.001

Riva, S., Monti, M., & Antonietti, A. (2011). Simple heuristics in over-the-counter drug choices: a new hint for medical education and practice. *Advances in Medical Education and Practice*, *2*, 59-70. https://doi.org/10.2147/AMEP.S13004

Rutherford, T. (2017). The measurement of calibration in real contexts. *Learning and Instruction*, *47*, 33-42. https://doi.org/10.1016/j.learninstruc.2016.10.006

Schneider, W. R. (2011). The Relationship Between Statistics Self-Efficacy , Statistics Anxiety , and Performance in an Introductory Graduate Statistics Course. *University of South Florida Scholar Commons*, 65. https://doi.org/3450237

Schraw, G. (2009). Measuring Metacognitive Judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 415-429). Routledge.

Serra, M. J., & Metcalfe, J. (2009). Effective Implementation of Metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 278-298). Routledge.

Shaughnessy, J. J. M. (1979). Confidence-judgment accuracy as a predictor of test performance. *Journal of Research in Personality*, *13*(4), 505-514. https://doi.org/10.1016/0092-6566(79)90012-6

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

Stankov, L. (2013). Noncognitive predictors of intelligence and academic achievement: An important role of confidence. *Personality and Individual Differences*, *55*(7), 727-732. https://doi.org/10.1016/j.paid.2013.07.006

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, *21*(6), 971-986. https://doi.org/10.1016/S0191-8869(96)00130-4

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence*, *25*(2), 93-109. https://doi.org/10.1016/S0160-2896(97)90047-7

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences*, *22*(6), 747-758. https://doi.org/10.1016/j.lindif.2012.05.013

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *The Behavioral and Brain Sciences*, *23*(5), 645-665. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11301544

Stupple, E. J. N., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking & Reasoning*, *19*(1), 54-77. https://doi.org/10.1080/13546783.2012.735622

Tabachnick, B. G., & Fidell, L. S. (1996). *Using Multivariate Statistics* (3rd ed.). New York: HarperCollins.

Tempelaar, D. T. (2009). The Role of Self-theories of Intelligence and Self-perceived Metacognitive Knowledge, Skills, and Attitudes, in Learning Statistics. *Fifth Global SELF International Biennial Conference. Enabling Human Potential*, 13-15. Retrieved from http://www.self.ox.ac.uk/documents/Tempelaar.pdf

Thompson, V., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107-140. https://doi.org/10.1016/j.cogpsych.2011.06.001

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

Thurstone, L. L., & Thurstone, T. G. (1987). *TEA - tests de aptitudes escolares : manual* (Vol. 5a). Madrid: Tea.

Tobias, S., & Everson, H. T. (2009). The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education. In D. L. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Handbook of metacognition in education* (pp. 107-127). New York: Routledge.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema*, *20*(1), 174-180.

Was, C. A. (2014). Discrimination in measures of knowledge monitoring accuracy. *Advances in Cognitive Psychology*, *10*(3), 104-112. https://doi.org/10.5709/acp-0161-y

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.

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? *Experimental Psychology*, *50*(2), 97-106. https://doi.org/10.1026//1618-3169.50.2.97

APPENDIX (Agus et al., 2016)

Item Example in Verbal-Numerical Format

A factory produces electronic games, but not all of them work well. Of every 100 game products: 20 may have an electrical problem, 80 can 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 etract a game that has been sent to shops for ommercialisation and evaluated as free of defects, what is the propability that it is defective?

a) 10/90

b) 10/100

c) 10/80

d) 20/100

What reasoning did you apply to solve this problem?

____

How confident are you in the correctness of your response?

Not at all confident	Slightly confident	A moderate amount confident	Moderately confident	Extremely confident
1	2	3	4	5

Item Example in Graphical-Pictorial Format

A factory that produces personal computers has problems in the production process. Some of the computers are defective (problems with the video card). Such problems are not always identified by the quality control and consequently some defective computers are sent forward in the production line. The graphic below shows this process.

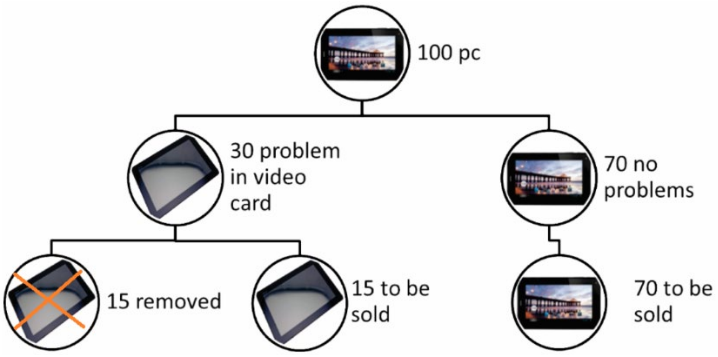
What is the probability that a computer sent to shops for commercialization and evaluated as free of defects, is defective?

a) 15/100

b) 15/70

c) 15/85

d) 3/10



What reasoning did you apply to solve this problem?

____

How confident are you in the correctness of your response?

Not at all confident	Slightly confident	A moderate amount confident	Moderately confident	Extremely confident
1	2	3	4	5