Journal of Cognitive Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pecp21

Sets or frequencies? How to help people solve conditional probability problems

Rodrigo Moro [a b], Gustavo A. Bodanza [a b] & Esteban Freidin [b c]

[a] Universidad Nacional del Sur, Bahía Blanca, Argentina

[b] Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina

[c] Centro de Recursos Naturales Renovables de la Zona Semiárida, Bahía Blanca, Argentina

Available online: 05 Oct 2011

To cite this article: Rodrigo Moro, Gustavo A. Bodanza & Esteban Freidin (2011): Sets or frequencies? How to help people solve conditional probability problems, Journal of Cognitive Psychology, DOI:10.1080/20445911.2011.579072

To link to this article: http://dx.doi.org/10.1080/20445911.2011.579072

Psychology Press
Taylor & Francis Group

Sets or frequencies? How to help people solve conditional probability problems

Rodrigo Moro[1,2], Gustavo A. Bodanza[1,2], and Esteban Freidin[2,3]

[1]Universidad Nacional del Sur, Bahía Blanca, Argentina
[2]Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
[3]Centro de Recursos Naturales Renovables de la Zona Semiárida, Bahía Blanca, Argentina

Since the 1970s, the Heuristics and Biases Program in Cognitive Psychology has shown that people do not reason correctly about conditional probability problems. In the 1990s, however, evolutionary psychologists discovered that if the same problems are presented in a different way, people's performance greatly improves. Two explanations have been offered to account for this facilitation effect: the *natural frequency hypothesis* and the *nested-set hypothesis*. The empirical evidence on this debate is mixed. We review the literature pointing out some methodological issues that we take into account in our own present experiments. We interpret our results as suggesting that when the mentioned methodological problems are tackled, the evidence seems to favour the natural frequency hypothesis and to go against the nested-set hypothesis.

Keywords: Base-rate neglect; Bayesian inference; Conditional probability problems; Evolutionary Psychology Programme; Frequency effect; Heuristic and Biases Programme; Natural frequency hypothesis; Nested-set hypothesis.

Probability problems are hard. Even people with some training struggle to solve probability problems correctly. In addition to intuitions or personal experience from the classroom, there is empirical evidence to support these claims. In the early 1970s Amos Tversky and Daniel Kahnemann founded a research programme in Cognitive Psychology called Heuristics and Biases (HBP) which partly consisted of documenting people's judgement errors on diverse circumstances. Since its beginnings, researchers in the HBP have found extensive evidence of people tending to commit reasoning errors when dealing with probability problems (see Gilovich, Griffin, & Kahneman, 2002, for a review). A particular case involves people's tendency to fail when reasoning about *conditional* probability problems. Here is the most famous example of this type of problem, the medical diagnosis problem:

Standard version of the medical diagnosis problem:

> The chance that an American selected at random has the disease X is 1/1000. A test has been developed to detect such a disease. An individual who does not have the disease has a 50/1000 chance of testing positive. An individual

Correspondence should be addressed to Rodrigo Moro, Departamento de Humanidades, Universidad Nacional del Sur, 12 de Octubre 1198, 5to piso, 8000 Bahía Blanca, Argentina. E-mail: rmoro@uns.edu.ar

This research was financially supported by Agencia Nacional de Promoción Científica y Tecnológica (PICT 693 and PICT 106) and by Universidad Nacional del Sur (PGI 24/ZI29 and PGI 24/ZI38). We thank Linden Ball, Gary Brase, W. Trey Hill, and an anonymous reviewer for helpful comments on a previous version of this paper.

http://www.psypress.com/ecp http://dx.doi.org/10.1080/20445911.2011.579072

> who does have the disease will definitely test positive. Suppose that we select an American by lottery and we know nothing about the person's symptoms or signs. Assume also that this person is tested for disease X and found to have a positive result. What is the chance that this individual actually has the disease?

Most people (even physicians!) tend to respond "95%". The correct answer according to the probability calculus is around 2%.[1] Studies show that typically less than 20% of participants get the correct answer (Casscells, Schoenberger, & Grayboys, 1978; Cosmides & Tooby, 1996; Eddy, 1982; Evans, Handley, Perham, Over, & Thompson, 2000; Gigerenzer & Hoffrage, 1995; Sloman, Over, Slovak, & Stivel, 2003). Thus, since the discovery of the phenomenon in the late 1970s, it has seemed clear that most people (without proper training) are unable to solve this type of problem.

In the mid 1990s, Gerd Gigerenzer and other evolutionary psychologists came along and gave an important turn to the state of the art. Evolutionary psychologists began noticing that a problem like the one just presented has at least two features that are worth considering: (1) The information is presented in a probability format; for example, notice the information about the base rate of the disease: "The chance that an American selected at random has the disease X is 1/1000"; and (2) the crucial question is about a single-event probability, namely, that a given person has the disease. Evolutionary Psychologists showed that if the same problem is framed differently, people's performance greatly improves. More specifically, if the problem presents the information under a specific format called "natural frequency format", around 50% of participants get the correct answer, in contrast with the 20% success in the probability format (Gigerenzer & Hoffrage, 1995). In the case of the medical diagnosis problem, the natural frequency condition would read as follows:

Natural frequency version of the medical diagnosis problem:

> One out of every 1000 American has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 999 people who are perfectly healthy, 50 of them test positive for the disease. Imagine we have arranged a random sample of 1000 Americans. They were selected by lottery. Those who conducted the lottery had no information about the health status of any of these people. Given the information above, on average, how many people who test positive for disease will *actually* have the disease? ______ out of ______

Notice that the information is now presented in a frequentist format (e.g., 1 out of every 1000 American has disease X). But this is not just a frequentist format but, more specifically, a *natural* frequentist format. Such formats have an essential feature, namely, that the statistical information is presented as a partition or as if it were obtained by *natural sampling*, that is, by updating event frequencies through sequentially obtained information (Kleiter, 1994). It turns out that people perform significantly better with natural frequentist versions (see Gigerenzer & Hoffrage, 1995, for a systematic study). This is a robust finding. For more than a decade, though, there has been a heated debate on how to explain this facilitation effect.

Despite the controversy, there seems to be agreement about the following point. The natural frequentist version is computationally simpler than the standard probabilistic version. Gigerenzer and Hoffrage (1995) spell out this point by showing that the equation needed to solve the former version is simpler (i.e., it contains fewer elements and operations, and uses whole numbers instead of fractions or decimals) than the equation needed to solve the latter version. Nonetheless, researchers in the area usually go further than this computational point when trying to account for the facilitation effect at stake.

There are two main proposals, one by members of the Evolutionary Psychology Program (EPP) and the other by members of the HBP. The natural frequency hypothesis supported by EPP basically says that the *natural frequency format* is the responsible factor for the improvement in people's performance (Brase, Cosmides, & Tooby, 1998). This format requires both the information and the question of the problem to be given in terms of natural frequencies (rather than in terms

[1] This can been seen as the result of a direct application of Bayes rule: P (disease/positive) $= P$ (disease and positive)/P (positive) $= .001 \times 1/.001 \times 1 + .999 \times .05 \ \approx .02$.

of probabilities). The advocates of EPP even speculate with an evolutionary scenario where our ancestors repeatedly faced uncertain situations and the acquisition and use of frequentist information (e.g., ''we were successful five times out of 20 when we hunted in the north canyon'') improved their survival. But regardless of the evolutionary origin of the phenomenon, the EPP standpoint stresses the natural frequentist phrasing as the key feature responsible for the facilitation effect.

Members of the HBP, in turn, have proposed the so-called ''nested-set hypothesis'' to explain the facilitation effect (Sloman et al., 2003; Tversky & Kahneman, 1983). The basic idea is that natural frequency versions tend to make clear the relevant subset relations of the problem. In the medical diagnosis problem, the natural frequency version would make clear that the set of people who test positive includes all sick people and some healthy people as well (see Figure 1). When people see clearly the set relations involved in this kind of problem, they tend to use base rates correctly and thus, their performance improves. According to this view, the success of the frequency effect does not have to do with natural frequency formats per se. In turn, HBP supporters predict that any format whatsoever that makes the relevant set relations clear will aid people's probabilistic judgements.

The empirical evidence on this debate is mixed. Some studies seem to support the natural frequency hypothesis (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Krämer & Gigerenzer, 2005; Zhu & Gigerenzer, 2006); others seem to support the nested-set hypothesis (Evans et al., 2000; Girotto & Gonzalez, 2001; Macchi, 2000; Mellers & McGraw, 1999; Sloman et al., 2003; Yamagishi, 2003). The resolution of this dispute is important because, besides increasing the understanding of this phenomenon, it may yield practical consequences regarding how to teach probability theory.

The present paper is structured as follows. We first review the literature pointing out some methodological issues to take into account in experimental research about probability judgement. We then report the results of three experiments where we either avoided or corrected the previously mentioned issues. And finally, based on the methodological issues reviewed and our own results, we argue that the evidence seems to favour the natural frequency hypothesis and to go against the nested-set hypothesis.

REVIEW OF THE LITERATURE

In order to decide between our rival hypotheses, the main strategy is to create *probability* versions where the set structure is indeed clarified and see whether such versions elicit the facilitation effect. This is without a doubt the right strategy to follow since HBP and EPP predict opposite results. HBP predicts that such versions will elicit a performance comparable to natural frequency versions of the problem, whereas EPP predicts no facilitation at all.

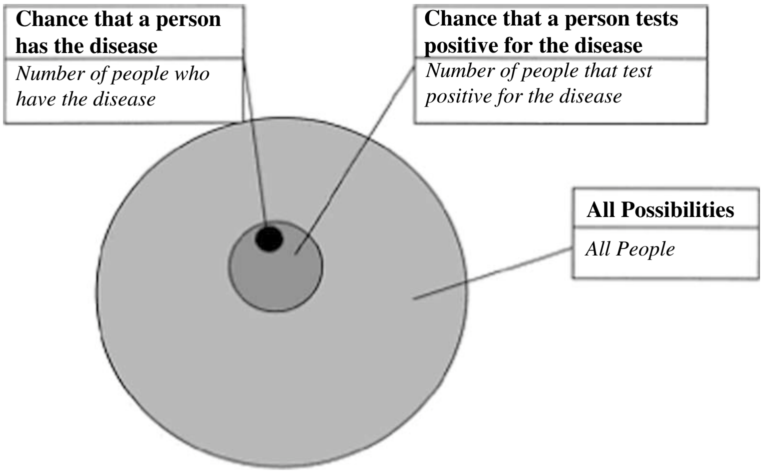
Chance that a person has the disease
Number of people who have the disease
Chance that a person tests positive for the disease
Number of people that test positive for the disease
All Possibilities
All People

Figure 1. Nested-set relations of the medical diagnosis problem. Bold text appeared in probability conditions. Italicized text in frequency conditions. Reprinted from *Organizational behavior and human decision processes*, Vol. 91, Issue 2, S. Sloman, Frequency illusions and other fallacies, p. 296, Copyright (2003), with permission from Elsevier.

The reader may think that there is another possible strategy, namely, to use natural frequency versions where the set relations are *not* clarified. If people perform well under such frameworks, the natural frequency hypothesis would receive empirical support. If not, the nested-set hypothesis would do so. Unfortunately, this possibility is ruled out from the outset. The reason is that one of the main features of *natural* frequency versions is that they present the information in a partitive way, leaving, thus, the set structure of the problem explicitly revealed.

The strategy of using probability versions with clarified set structures, on the other hand, raises an important methodological problem: how to distinguish probability versions from natural frequency versions. The problem with this distinction is that the concept of probability can be interpreted as suggesting relative frequencies. In fact, the frequentist interpretation is among the most plausible interpretations of probability (von Mises, 1957). Furthermore, the same numerical expressions—percentages, fractions, and even whole numbers—can be legitimately used for both types of versions. In other words, the distinction is vague. Nonetheless, there is some agreement in the literature on classifying some wording as typically frequentist (e.g., 3 out of every 10 cases of A are also cases of B) and some wording as genuinely probabilistic or nonfrequentist (e.g., the chance or probability that the single event A is a case of B is 30%). Unfortunately, some researchers are not particularly careful about this point, thus leading to disputes about how to interpret their results (see, for example, Hoffrage, Gigerenzer, Krauss, & Martignon, 2002). Taking this consideration into account, we next comment on three techniques that apply the general strategy of using probability versions with a clarified set structure. These three techniques are so far the most successful ones in showing important improvements in performance, apparently providing, thus, support for the nested-set hypothesis.

Technique 1: Using a natural chance format

One of the most effective techniques is the one used by Girotto and Gonzalez (2001). These authors have found a very clever way to express single event probabilities that emulates natural frequency setups. Here is an example of such a chance format:

> The applicants for admission to a prestigious university have to pass an entrance examination which involves an oral test and a written test. Here is the information about the results of last year examination.
>
> An applicant had 5 chances out of 100 of being accepted. 3 of the 5 chances of being accepted were associated with success in the oral test. 7 of the remaining 95 chances of being rejected were associated with success on the oral test. Imagine that Jean is an applicant to the entrance examination. Out of 100 chances, Jean has ___ chances of passing the oral test, ___ of which will be associated with being accepted. (Girotto and Gonzalez, 2001, pp. 272–273)

Under this condition, the partitive structure makes transparent the set structure of the problem. This chance format is shown to elicit a similar facilitation effect as the natural frequency version. This result seems to provide empirical support for the nested-set hypothesis.

What is the response from the advocates of the natural frequency hypothesis? The main objection is the suspicion that Girotto and Gonzalez's (2001) chances are not but "frequencies in disguise" (Hoffrage et al., 2002, p. 350). In fact, the structure of natural chance versions emulates the structure of natural frequency versions. Furthermore, the introduction of the problem mentions the *results* of last year's examination. That may induce people to think in frequentist terms. Additionally, phrases like "3 out of 5 chances of event A are associated with event B" invite many interpretations, one of which may be frequentist. These observations call into question the interpretation of the results. The facilitation effect may be due to a clarification of set relations or to cues leading to think of the problem in a frequentist way.

Actually, there seems to be some empirical evidence to support the suspicion of a frequentist interpretation. This comes from a study by Brase (2008). After giving participants the university admission problem, he made them choose among the following options:

- I thought about the information as a single application with some possibility of having been successful on the oral test and some possibility of having been accepted. [*probability interpretation*]

- I thought about the information as a large number of applications, some of which were successful on the oral test, and some of which were accepted. [*frequency interpretation*]
- Other: I thought about the information as ________. (Brase, 2008, p. 285)

Brase reports that an important percentage of participants (around 30%) selected the frequentist interpretation as their own. Thus, even if most people selected the probabilistic interpretation (around 60%), it is clear that the format is somewhat ambiguous. More important, the group that selected the frequentist interpretation performed significantly better than the group that selected the probabilistic interpretation. Thus, the reported facilitation seems to depend partially on the ambiguity of the format at stake. In other words, the main problem with the chance format is whether it can be taken as a genuine probability format or it is rather a frequency format in disguise. Until this point is clarified, the reported evidence does not seem to provide conclusive evidence in favour of the nested-set hypothesis. Let us consider, then, the next technique that seems to discriminate between our rival hypotheses.

Technique 2: Using improved wording

Sloman et al. (2003) used different versions of the medical diagnosis problem, some of which have an improved wording that reflects the set structure of the problem. Here is one of these versions:

Improved probability format version with transparent nested-sets relations:

> The probability that an average American has disease X is 1/1000. A test has been developed to detect if a person has disease X. *If the test is given and the person has the disease, the test comes out positive. But the test can come out positive even if the person is completely healthy.* Specifically, the chance is 50/1000 that someone who is perfectly healthy would test positive for the disease. Consider an average American. Assume you know nothing about the health status of this person. What is the probability that if this person is tested and found to have a positive result, the person would actually have the disease? (Sloman et al., 2003, p. 303, emphasis added)

Notice that the technique consists of two points: (1) making the problem about an average person, so the statistical information becomes relevant; and (2) stressing the possibility of positive tests being associated with both sick and healthy people. The problem is that Sloman and colleagues did not find consistent results. They tested several versions of probability formats *without* clarified set relations and several probability versions *with* the improved wording (and also a frequentist version). In some experiments, they found a big improvement in performance (from 20% to 48% of correct answers). However, they tested several versions because they recognised that some of these versions contained ambiguities (e.g., probability versions that may be interpreted as frequentist versions). Now, there is a comparison that Sloman et al. (2003) did not make. This is the comparison of the versions where ambiguities are highly reduced or eliminated. This is the most relevant comparison. The versions that contain ambiguities are not reliable because these ambiguities might be the source of errors. So, the comparison that really counts is the one with no, or at least less, ambiguous material. According to their own criteria, when facing the best probability version *without* clear set relations, 39% of participants give correct answers. And when facing the best probability version *with* clear set relations, 40% of people got the correct answer. These percentages are almost identical! So, for the best version of each type, the clarification of the nested-set relation did not seem to bring any improvement.

Given the inconsistency of results with this technique, we cannot confidently pass judgement on the issue at stake. However, we consider this technique as perfectly legitimate to distinguish among our rival hypotheses, so we decided to test its efficacy. But before we report our empirical data, let us move on to evidence of the last technique that shows apparently decisive results.

Technique 3: Using graphical representations

There is an additional way to create probability versions that reveal the set structure of the problem: to include a graphical representation that shows such a structure. This technique was used by Cosmides and Tooby (1996), Sloman et al. (2003), and Yamagishi (2003). We think this is actually one of the best ways to reveal the set structure of a problem. However, as in previous conditions, the interpretation of these results is not a straightforward matter. Again, precautions should be taken to avoid suggesting a frequentist reading. This is exactly the problem with Cosmides and Tooby's study because they made participants draw a graphical representation where there was one square per represented individual, clearly suggesting a frequentist reading.

Sloman et al. (2003) and Yamagishi (2003), on the other hand, seemed to avoid such problem. Sloman and colleagues (Exp. 2) used the following version of the medical diagnosis problem:

> Consider a test to detect a disease that a given American has a 1/1000 chance of getting. An individual that does not have the disease has a 5% chance of testing positive. An individual who does have the disease will definitely test positive. What is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? _____ % (Sloman et al., 2003, p. 300)

This version seems genuinely probabilistic. Under the control condition, they gave participants this version alone, and under the key experimental condition, they also included a graphical representation of the situation (Figure 1). Sloman and colleagues reported a significant improvement, from 20% of correct answers *without* the diagram to 48% *with* the diagram. This improvement was almost the same as the one elicited by the frequentist version (51%).

Yamagishi (2003), in turn, reports a similar result. He used different versions of the following problem:

> A factory manufactures artificial gemstones. Each gemstone has a 1/3 chance that it is blurred, a 1/3 chance that it is cracked, and a 1/3 chance that it contains neither. An inspection machine removes all cracked gemstones, and retains all clear gemstones. However, the machine removes 1/2 of blurred gemstones. What is the chance that a gemstone is blurred after the inspection? (Yamagishi, 2003. p. 99)

In the key experimental condition, Yamagishi (2003) included a graphical representation as in Figure 2. Across four experiments, he reported an improvement from an average of 15% of correct answers *without* the diagram to an average of 75% *with* the diagram. In this case, the diagram effect was actually stronger than the natural frequency effect (49%).

These two results are the clearest evidence in favour of the nested-set hypothesis and against the natural frequency hypothesis. However, a couple of objections can raise doubts about the validity of these results. The first objection has to do with the graphical representations in Sloman's and Yamagishi's studies (Figures 1 and 2, respectively). In our opinion, the graphical representations used in both studies give more information than the mere set structure of the problem. The figures show (Yamagishi) or suggest (Sloman) the *relative proportions* of the sets involved. Notice that this additional feature is *not* required by the nested-set hypothesis as it is usually stated. One can perfectly present the set structure of the problem without suggesting the relative size of each set. Showing the relative proportions, in turn, may suggest a frequentist interpretation of the problem since the differences in proportions can be viewed as differences in the amount of individuals in each set.[2] Alternatively, the suggestion of relative proportions may clarify the very goal of the

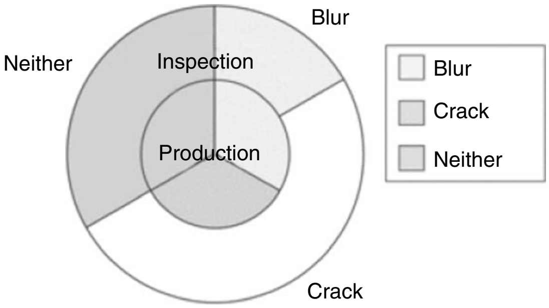

Figure 2. Graphical representation of the gemstone problem. Reproduced with permission from Experimental Psychology Vol. 50, (2), 2003, pp. 97–106. © 2003 Hogrefe & Huber Publishers, Cambridge, MA, Toronto, Göttingen, Bern.

[2] An anonymous reviewer asked for evidence of participants' association between relative sizes of sets and amount of individuals in those sets. Unfortunately, we were unable to find any studies that bear on the matter as research on set understanding is practically nonexistent. However, we plan to study experimentally the issue in further work.

task at hand, which is, after all, to obtain a determined proportion. Thus, it may happen that what produces the facilitation effect is not the clarification of the set structure but rather the clue about proportions. One can see this challenge as an opportunity to prove the *robustness* of the nested-set hypothesis. The point is, again, that such a hypothesis does not say anything about relative sizes of sets. So if we find that it predicts well even in conditions where relative sizes of sets are not suggested or shown, this result would support the hypothesis in conditions that have not been tested yet. On the contrary, if the result is negative, an advocate of such a hypothesis may try to reformulate it in such a way that it would involve information about the relative size of sets.

Furthermore, another objection can be raised regarding the problems used in both studies. Neither of these problems seems completely adequate. In the first place, the version of the medical diagnosis problem used by Sloman and colleagues is very problematic as they themselves admit (2003, p. 301). A first problem has to do with the fact that the information is given of an individual "getting" a disease, without specifying a time period over which the disease might be "gotten", so that whether the base rate information applies to the event at hand is questionable. A second problem has to do with the fact that the version of the problem at stake does not mention that the person was selected at random. So, participants may assume a different prior probability for the hypothesis of the disease. The gemstone problem used by Yamagishi (2003), in turn, can be also questioned since the presentation of the statistical information is not carefully phrased as genuinely probabilistic but it is rather frequentist in nature (e.g., the inspection machine removes half of the blurred gemstones). Thus, it is an open question whether the facilitation effect still occurs with improved versions of the same problems, that is, with versions that avoid the mentioned methodological weaknesses. This is exactly what we explore in our Experiment 2.

To sum up, although the evidence provided so far seems to favour the nested-set hypothesis, more empirical work is needed to provide a conclusive case for such a hypothesis. Next, we present a series of experiments aimed at: (1) tackling some of the methodological issues of the previous summarised studies; (2) testing whether the facilitation effect on probability judgements persists after that; and (3) finding evidence to further assess the relative merits of the nested-set and natural frequency hypotheses.

EXPERIMENT 1: TESTING SLOMAN AND COLLEAGUES' IMPROVED WORDING TECHNIQUE

In Experiment 1 we test whether Sloman and colleagues' (2003) improved wording technique enhances participants' performance on probability judgements relative to the standard probability format and also compared against a frequentist format.

Method

Participants were 63 undergraduate students (21 per condition) from Business, Economics and Philosophy at Universidad Nacional del Sur, Argentina. They were fulfilling course requirements in the second semester of 2007 and had no previous training in probability theory. The experiment was run in sessions of 15–30 students. We gave each participant a sheet that contained the medical diagnosis problem in one of three versions described later. The version each participant received was randomly determined. We asked them to take as much time as they liked to work on the problem.

In this experiment we tested three versions of the medical diagnosis problem: a probability version *without* clarified set relations (probability condition 1), a probability version *with* clarified set relations (through improved wording; probability condition 2), and a natural frequentist version. Given the inconsistency of results mentioned earlier, our main goal was to test Sloman and colleagues' (2003) improved wording technique. If the nested-set hypothesis is correct, the version with clarified set relations through improved wording should elicit significantly more correct answers than the probability version without such improved wording. We used materials that avoided ambiguities mentioned in the previous section (see a description of the materials in the Appendix).

Results and discussion

As Sloman and colleagues (2003) did, we scored responses between 1.8 and 2.2% for probabilistic versions (written in any format) as correct (we discuss this criterion in the introduction of Experiment 2). For the natural frequentist version, we were stricter and only accepted the answer 1 out of 51 as correct since it has been shown that the answer 1 out of 50 may hide conceptual errors (Evans et al., 2000). Results are shown in Table 1.

The difference between probability conditions 1 and 2 was the transparency of nested-set relations (absent in condition 1 and present in condition 2—following Sloman and colleagues', 2003, technique). Given that none of these conditions elicited a single correct answer, the transparency of nested-set relations caused no improvement at all. This result clearly goes against the nested-set hypothesis. However, the advocates of the nested-set hypothesis can interpret the result as suggesting that Sloman and colleagues' technique does not help revealing the set structure of the problem. Alternatively, they may argue that the enhancing effect of this technique was merely hidden by a floor effect for the probability versions of the problems, since the probability versions were just "too hard" to solve.

At this point, one may wonder about the substantial difference between present results and those high percentages of correct responses found by Sloman and colleagues (2003). We do not have a satisfactory explanation for this difference. But our low percentages of correct responses are closer to the literature than the ones reported by Sloman and colleagues. Actually, their 39% of correct answers in a standard probability format is the highest in the literature. Yet, there is an aspect of their data that we did replicate: There is no facilitation effect in versions that are clear and free of ambiguities.

In addition, we also found the facilitation effect, as shown in Table 1. The natural frequency condition elicited significantly more correct responses than each of the probability conditions (Fisher's exact test, both *p*s < .01).

TABLE 1
Percentages (and frequencies) of correct answers (Experiment 1)

Probability without nested-set wording	*Probability with nested-set wording*	*Natural frequency*
0.0	0.0	(0/21)
(0/21)	33.3	(7/21)

EXPERIMENT 2: TESTING THE USE OF THE GRAPHICAL REPRESENTATION TECHNIQUE

The main goal of this experiment was to test the graphical representation technique discussed previously. In addition, we incorporated several methodological improvements:

1. It has been correctly argued (Evans et al., 2000) that the response "2%" in the probability version of the medical diagnosis problem is ambiguous in a crucial sense. On the one hand, it may be that the participant rounds up the right answer (1.96%), which seems perfectly reasonable. On the other hand, it may be that the participant incorrectly takes the false positive rate as applying to the whole population rather than to the healthy population. This last interpretation does not apply to the results from Experiment 1, since nobody was even close to the right answer. Still, it may be a source of misinterpretations. In order to tackle such a problem, we changed the ratios of the two problems used so that a mistake as the one just mentioned would yield an error of 5% or more in the final answer, and hence it could not be confounded with a simple rounding of the final answer.
2. Under the probability versions, we also gave more alternatives in the presentation of statistical information (besides fractions, we also presented percentages and decimals in the first datum), so that participants could make calculations under their preferred format. Additionally, we did not stipulate that the answer should be given as a percentage, accepting decimals or fractions as well.
3. Given that many participants in Experiment 1 gave the base rate of the disease as an answer, we decided to emphasise the fact that the selected person had received a positive result.
4. We used two problems. Besides the medical diagnosis problem, we incorporated Yamagishi's gemstone problem.

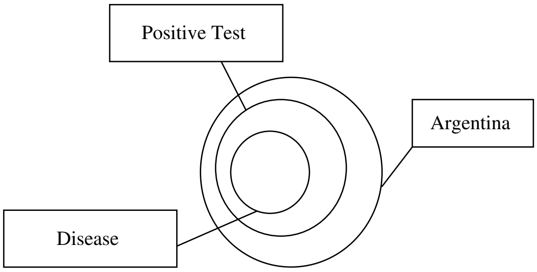

Figure 3. Graphical representation of the medical diagnosis problem used in Experiments 2 and 3.

5. We rewrote the probability version of the gemstone problem so that it did not contain frequentist wording.
6. In graphical representation conditions, we used graphics that did not suggest the relative proportions of each class, avoiding, thus, a possible frequentist reading of the problem (see Figures 3 and 4). Curly brackets (i.e., "{", "}") as appeared in Figure 4 are frequently used (in all levels of education) to present a *set of elements a given term involves.* Thus, we expected participants to read the production part of the graphic as a set containing cracked, blurred, and clear stones and the retention part as a set containing all clear gemstones and some blurred ones.[3]

Method

We used a 2 × 2 design by taking problem format (probability vs. frequency) and graphical representation (presence vs. absence) as the factors to study. Thus, each problem could be presented under four different formats (probability with graphic, probability without graphic, frequency with graphic, and frequency without graphic). The materials are described in the Appendix.

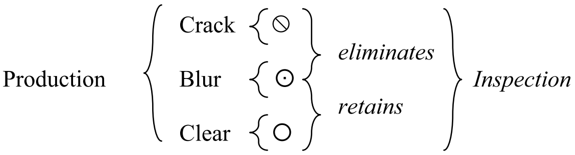

Figure 4. Graphical representation of the gemstone problem used in Experiments 2 and 3.

Participants were 80 undergraduate students from Accounting at Universidad Nacional del Sur, Argentina, who were fulfilling course requirements during the first semester of 2009 and had no formal training in probability theory. Each participant was given a booklet containing two pages, each with a different problem. Thus, each participant worked on both problems (i.e., Sloman et al.'s, 2003, and Yamagishi's, 2003), but under different formats. The overall condition of each participant was randomly determined. The order of presentation of problems and the order of versions (i.e., graphics/no graphics, frequencies/probabilities) were crossed resulting in 24 different types of booklet. Participants were required to work on each problem at a time and not to come back to the first problem after solving the second one. There was no time restriction and the session lasted around 40 minutes.

Results and discussion

Summarised results are presented in Table 2. First and without discriminating graphic from non-graphic conditions, it could be observed that frequentist formats improved performance relative to probabilistic ones supporting the classic facilitation effect (for both the medical diagnosis problem and the gemstones problem, Fisher's exact test, $p < .001$). Second and focusing on the probability conditions of both problems, it could be read from Table 2 that these probability problems (with and without graphics) elicited very few correct answers. In the probability version of the medical diagnosis problem, the graphical representation did not seem to help at all, since no participant correctly solved the

TABLE 2
Percentages (and frequencies) of correct answers by problem, format, and the presence or absence of graphics (Experiment 2)

	Medical diagnosis problem		*Gemstone problem*	
	No graphic	*Graphic*	*No graphic*	*Graphic*
Probability	5.3 (1/19)	0.0 (0/20)	4.3 (1/23)	4.8 (1/21)
Frequency	33.3 (7/21)	50.0 (10/20)	88.9 (16/18)	72.2 (13/18)

[3] An anonymous reviewer argues that Figure 4 fails to show the hierarchical structure of the key subset–superset relationships. We will take into account this reasonable criticism in materials of additional experiments to complement our research.

problem in the graphic condition. In the probability version of the gemstone problem, the graphical representation improved performance for less than one percentage point, making this difference nonsignificant (Fisher's exact test, $p = .73$). Third, we focus on the frequency versions of both problems. Though we found that the percentages of correct answers were high in both problems (at least, relative to their probability versions), the graphical representation seemed to have a positive effect in the medical diagnosis problem, and a negative effect in the gemstone problem. However, neither of these graphic-induced differences was even near reaching statistical significance (for the medical diagnosis problem, Fisher's exact test, $p = .22$; for the gemstone problem, Fisher's exact test, $p = .20$). Last, considering that all subjects responded to two successive problems, it was important to test for potential carryover or transfer effects from the first to the second format presented. We found no transfer effect from the initial experience of either graphics (after graphics in the first problem, 36% of participants responded correctly on the second problem, whereas without graphics in the first, 29% of participants presented a correct performance in the second: Fisher's exact test, $p = .42$) or the frequentist wording (after a frequentist wording in the first problem, 27% of participants responded correctly on the second problem, whereas after a probabilistic wording in the first, 39% of participants performed well in the second problem: Fisher's exact test, $p = .32$). In other words, having a graph or not in the first problem had no apparent influence on the performance in the second problem, and the same was true for participants' performance after an initial problem in a frequentist format or a probabilistic one.

All in all, these results show the typical frequency effect where performance greatly improves when the problem is stated in a natural frequentist format as opposed to a probabilistic one. In addition, and contrary to Sloman and colleagues' (2003) and Yamagishi's (2003) results, we did not find a facilitatory effect of having a graphic representation. Recall that the main prediction of the nested-set hypothesis was that any probability condition that revealed the set structure of the problem would elicit a facilitation effect. We tried to reveal the set structure of each problem by using graphical representations that avoid previous methodological weaknesses. We found no significant improvement in performance in graphic-probability conditions. Thus, if our graphical representations succeeded in revealing the set structure of the problems, the nested-set factor does not seem to account for the facilitation effect at stake.

EXPERIMENT 3: TESTING PARTICIPANTS' UNDERSTANDING OF SET RELATIONS

A possible explanation of our failure to replicate the facilitation effect in the last experiment could be that the graphical representations used in Experiment 2 might not clearly convey the relevant set relations of the problem. Indeed, if our graphical representations were inefficient to expose the key set relations, we would expect exactly what we found in Experiment 2: the graphical conditions eliciting no substantial improvement and the natural frequency conditions eliciting the facilitation effect since the set relations were verbally exposed.

The goal of Experiment 3 was to test participants' understanding of set relations on the material of Experiment 2 in order to check whether there is some correlation between understanding of set relations and performance on conditional probability problems. On the one hand, if the nested-set hypothesis were correct, we would expect a pattern of set understanding very similar to the pattern of correct answers in Experiment 2. On the other hand, if the natural frequency hypothesis were correct, we would not expect any resemblance between performance in Experiment 2 and set-relation understanding in Experiment 3 across conditions. It is worth noting that the understanding of set structures has not been empirically tested before in the literature of probability judgement. It has been just assumed that the natural frequency formats and graphical representations can improve such understanding but no empirical ground has been provided for such a claim.

Method

A completely different group of 92 undergraduate students from Accounting at Universidad Nacional del Sur, Argentina, participated in Experiment 3 while fulfilling course requirements during the first semester of 2009. The experiment was run in two sessions of about 45 participants each. No time restrictions were imposed. Each session lasted about 30 minutes.

We used exactly the same design and materials as in Experiment 2, but instead of asking for probability judgements, we asked several questions about the relevant set relations of each problem (e.g., true/false questions such as "The group of people who tested positive is completely included in the group of sick people"; see also Figure 5. For a description of the materials, see the Appendix). Participants' responses allowed us to obtain an individual score consisting of the percentage of correct responses. This score could be regarded as a continuous variable, the distribution of which was checked for normality and homogeneity of variance before performing the corresponding ANOVAs with format (frequentist vs. probabilistic) and graphic (presence vs. absence) as between-subject factors.

Results and discussion

The main results of Experiment 3 are presented in Table 3. The most striking aspect is that we found very high percentages of correct answers in all conditions (a mean (±1 *SEM*) of 81% (±0.01) of correct answers for the medical diagnosis problem and of 89% (±0.01) for the gemstones problem). Participants seemed to understand very well the set relations involved in the problems, even in the probability conditions (a mean (±1 *SEM*) of 82% (±0.01)) where almost no participant gave a correct answer in Experiment 2. In addition, the presence of a graph significantly improved participants' performance in the probability format of the medical diagnosis problem (see Table 3). This was confirmed by the ANOVA of the percentage of correct responses which showed a significant Format × Graph interaction, $F(1,\ 88) = 6.49$, $p < .05$, partial $\text{eta}^2 = 7\%$, and by LSD post hoc contrasts that showed that the source of the significant interaction was the performance improvement in the probability condition with graphs (mean ± 95% CI: 80.8 ± 5) relative to the same condition without graphs (72.2 ± 6); $p = .014$, whereas participants did not differ in the frequency format condition with and without graphs (84.1 ± 4, and 0.87.8 ± 4, respectively), $p = .28$. That ANOVA also showed a main effect of format, $F(1,\ 88) = 15.44$, $p < .001$, partial $\text{eta}^2 = 15\%$, and no effect of graph, $F(1,\ 88) = 1.05$, $p = .31$. Moreover, the 2 × 2 ANOVA with participants' correct responses on the gemstones problem showed no reliable differences among conditions: format, $F(1,\ 88) = 0.07$; graph, $F(1,\ 88) = 1.03$; Format × Graph interaction, $F(1,\ 88) = 0.02$ (see also the right-hand side of Table 3).

In the medical diagnosis problem, the graphical representation elicited a significant improvement in set-relation understanding (from 72% to 81% of correct answers), but this graph had no impact in probability judgement performance in Experiment 2. In the gemstone problem in Experiment 3, there was an improvement in set-

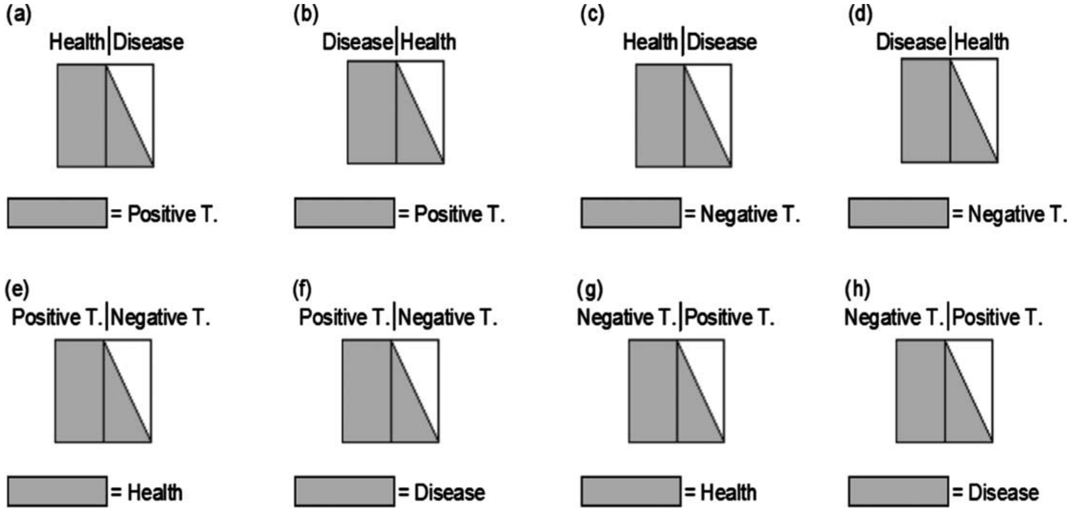

Figure 5. Graphical representation exercise presented in Experiment 3. Participants were asked to circle all the graphics that correctly represent the relations between health, disease, and test results.

TABLE 3
Mean proportion of correct answers on set-relation understanding as a function of problem format and the presence/absence of graphics (Experiment 3)

	Medical diagnosis problem		*Gemstone problem*	
	No graphic	*Graphic*	*No graphic*	*Graphic*
Probability	72.2	80.8	86.0	89.4
Frequency	87.8	84.1	89.2	90.8

relation understanding (from 86% to 89% of correct answers) but it fell short of statistical significance. Under frequency conditions in Experiment 3, the percentages of correct answers were very similar and high in both conditions with and without graphic representations, thus suggesting proper set-relation understanding. There was no evidence of graph facilitation on that understanding.

Two points of results from Experiment 3 are particularly clarifying in terms of present goals, especially when considered in relation to results from Experiment 2. First, participants showed high understanding of set relations in Experiment 3. Second, the medical diagnosis graph used to aid set-relation understanding had indeed an improving effect in participants' answers about set relations, at least, for the probability format of such problem. Hence, if set-relation understanding were the key to solve conditional probability problems, we could derive two retrodictions for the results of Experiment 2: (1) Most people across conditions should have correctly solved the problems; and (2) focusing on the probability conditions of the medical diagnosis scenario, participants under the graphical condition should have performed significantly better than participants under the nongraphical condition. Clearly, each of these retrodictions from the nested-set perspective failed.

All in all, although the graphical representations used only elicited improvement in one format of one of the problems, there was, in fact, not much room to improve in the other conditions because participants did not seem to have much trouble in understanding the relevant set relations. However, the crucial message here is that a clear understanding of set relations does not seem to translate into good performance on conditional probability problems as the joint results of Experiments 2 and 3 suggest. Thus, these results provide strong evidence against the nested-set hypothesis.

A final remark is worth mentioning about the composition of our samples. Our samples, composed mostly by undergraduates from Economics and Business in Experiment 1 and from Accounting in Experiments 2 and 3, differ from the typical sample of other studies on the same topic, usually composed only by psychology participants. A particular issue of concern is that Chapman and Liu (2009) showed that frequency effects only appeared for those participants high in *numeracy* (i.e., ability to process basic numerical concepts). One then may argue that accounting and economics students may be, on average, higher in numeracy than the average psychology student. This would raise concerns about the comparability with results from other studies. However, our results on the frequency effect do not differ from similar studies in the literature, so comparability does not seem to be a problem in this context.

GENERAL DISCUSSION

Conditional probability problems are hard. Even in the most favourable conditions, people struggle to get the right answer. However, it has been shown that when problems are presented under a natural frequency format, people's performance improves. For more than a decade, researchers have engaged in an important debate on how to explain such a facilitation effect. One reasonable response is that under those formats, participants clearly see the set relations of the problem at stake. We reviewed several studies that provide empirical support for such an account of the facilitation effect. However, we also pointed out some methodological problems that blur the interpretation of these studies. Finally, we reported three original experiments that show that when these issues are considered and the problems avoided, a clear understanding of the relevant set relations does not seem to be the cause of the facilitation effect on conditional probability judgements.

One may be tempted, then, to proclaim the natural frequency formats as the cause of the facilitation effect. However, we prefer to be cautious. The reason is that both natural frequency problems in our study are computationally simpler than our probability problems. Thus, it is an open question whether computational complexity is actually the key factor to account for the facilitation phenomenon at stake. Fortunately, there are some studies that seem to shed light on the matter. Brase (2008) shows that, for computationally equivalent problems, the frequentist-wording condition elicits a significantly better performance than the probabilistic-wording condition. Brase (2009), in turn, shows that a frequentist pictorial representation elicits a significantly better performance than a nested-set pictorial representation, again for problems that are computationally equivalent. If computational complexity were the key for the facilitation effect at stake, one would predict opposite results. Unfortunately, as Brase (2009) recognises, the use of only one reasoning problem makes necessary more empirical work to establish the generalisability of his results. Again, our own study is completely silent about it, since natural frequency formats and computational simplicity are confounded factors in our study. Thus, more research will be needed to settle the matter of the real source of the facilitation effect. From a practical point of view, however, the recommendation is clear. If you want to help people to solve conditional probability problems, you should definitely use natural frequency formats.

Original manuscript received October 2010
Revised manuscript received March 2011
First published online October 2011

REFERENCES

Brase, G., Cosmides, L., & Tooby, J. (1998). Individuation, counting, and statistical inference: The role of frequency and whole object representations in judgment under uncertainty. *Journal of Experimental Psychology: General*, *127*, 3–21.

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin and Review*, *15*, 284–289.

Brase, G. L. (2009). Pictorial representations and numerical representations in Bayesian reasoning. *Applied Cognitive Psychology*, *23*, 369–381.

Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, *299*, 999–1000.

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decisión Making*, *4*(1), 34–40.

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

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovik, & A. Tverky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.

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

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

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

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, *78*, 247–276.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, *84*, 343–352.

Kleiter, G. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer & D. Lang (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 375–388). New York, NY: Springer-Verlag.

Krämer, W., & Gigerenzer, G. (2005). How to confuse with statistics or the use and misuse of conditional probabilities. *Statistical Science*, *20*, 223–230.

Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes*, *82*, 217–236.

Mellers, B., & McGraw, P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review*, *106*, 417–424.

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

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Von Mises, R. (1957). *Probability, statistics and truth*. London, UK: Allen & Unwin.

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? *Experimental Psychology*, *50*, 97–106.

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, *98*, 287–308.

APPENDIX 1: MATERIALS USED IN EXPERIMENT 1 (ALL MATERIALS ARE TRANSLATED FROM SPANISH)

Version 1: Probability version without transparent nested-set relation

> The chance that an Argentinean selected at random has the disease X is 1/1000. A test has been developed to detect such a disease. An individual who does not have the disease has a 50/1000 chance of testing positive. An individual who does have the disease will definitely test positive. Suppose that we select an Argentinean by lottery and we know nothing about the person's symptoms or signs. Assume also that this person is tested for disease X and found to have a positive result. What is the chance that this individual actually has the disease? _____ %

Version 2: Probability version transparent nested-set relation

> The chance that an Argentinean selected at random has disease X is 1/1000. A test has been developed to detect such a disease. If the test is given and the person has the disease, the test comes out positive. But the test can come out positive even if the person is completely healthy. Specifically, the chance is 50/1000 that someone who is perfectly healthy would test positive for the disease. Suppose that we select an Argentinean by lottery and we know nothing about the person's symptoms or signs. Assume also that this person is tested for disease X and found to have a positive result. What is the chance that this individual actually has the disease? _____ %

Finally, the last version was the natural frequentist version shown earlier.

APPENDIX 2: MATERIALS USED IN EXPERIMENT 2

Probability version of the medical diagnosis problem

> There is an epidemic in our country. The probability that an Argentinean selected at random has the disease X is 7/10 (or 70% or 0.7). A test has been developed to detect such a disease. An individual who does *not* have the disease has a probability of 1/10 of testing positive. An individual who does have the disease will definitely test positive. Suppose that we select an Argentinean by lottery and we know nothing about the person's symptoms or signs. Assume also that this person is tested for disease X and found to have a positive result. Given this positive result, what is the chance that this individual actually has the disease? _____ %

The frequency version of the medical diagnosis problem is the same as in Experiment 1, except for the numbers (e.g., 70 of every 100 people have the disease) and the addendum "There is an epidemic in our country". The graphical representation that came with the graphical version is given in Figure 3.

Probability version of the gemstone problem

> A factory manufactures artificial gemstones. Each gemstone has a ¼ chance (or 25% or 0.25) of being blurred, a ¼ chance of being cracked, and a ½ chance of being clear. An inspection machine has been created to distinguish them and retain only the clear ones but it still does not work very well. If a gemstone is cracked, the machine will remove it; if the stone is clear, it will retain it. However, if a stone is blurred, there is a ½ chance of it being retained. Let us suppose that we select at random a gemstone that has been retained by the inspection machine. Given that this gemstone has been retained, what is the chance that it is blurred?

Frequentist version of the gemstone problem

> A factory manufactures 1200 artificial gemstones daily. Among the 1200, 300 gemstones are blurred, 300 are cracked, and 600 contain neither. An inspection machine removes all cracked gemstones, and retains all clear gemstones. However, the machine removes half of blurred gemstones. How many gemstones pass

the inspection, and how many among them are blurred?

The graphical representation for the graphical conditions can be seen in Figure 4.

APPENDIX 3: MATERIALS USED IN EXPERIMENT 3

The materials of Experiment 3 are basically based on those of Experiment 2. The information given in each problem is exactly the same as that used in Experiment 2 except for the text that says that either a person or a group was selected and the final probability judgement question. For example, the probability version of the medical diagnosis problem reads:

> There is an epidemic in our country. The probability that an Argentinean selected at random has the disease X is 7/10 (or 70% or 0.7). A test has been developed to detect such a disease. An individual who does *not* have the disease has a probability of 1/10 of testing positive. An individual who does have the disease will definitely test positive.

We use three set of questions to test the understanding of set relations: The first set contained true/false questions in the form of conditional statements. For example, some of the questions of the medical diagnosis problem were:

(a) If a person has the disease, the test will come out either positive or negative.
(b) If the test comes out positive, the person for sure has the disease.

The second set of questions involved eight graphical representations. We asked participants to circle the ones that correctly represent the relations between health, disease, and test results pointing out that the amount of correct representations could go from 0 to 8. Figure 5 shows the graphics used for such exercise for the medical diagnosis problem.

The third and last exercise involved a true/false exercise that contained set-theory statements. For example, some statements for the medical diagnosis problems were as follows:

(a) The group of people who tested positive involves all healthy people and some sick people.
(b) The group of people who tested positive is completely included in the group of sick people.