Judgment and Decision Making, Vol. 3, No. 2, February 2008, pp. 140–152

Frequency formats, probability formats, or problem structure? A test of the nested-sets hypothesis in an extensional reasoning task

William P. Neace*, Steven Michaud, Lauren Bolling, Kate Deer, and Ljiljana Zecevic
Department of Psychology
University of Hartford

Abstract

Five experiments addressed a controversy in the probability judgment literature that centers on the efficacy of framing probabilities as frequencies. The natural frequency view predicts that frequency formats attenuate errors, while the nested-sets view predicts that highlighting the set-subset structure of the problem reduces error, regardless of problem format. This study tested these predictions using a conjunction task. Previous studies reporting that frequency formats reduced conjunction errors confounded reference class with problem format. After controlling this confound, the present study's findings show that conjunction errors can be reduced using either a probability or a frequency format, that frequency effects depend upon the presence of a reference class, and that frequency formats do not promote better statistical reasoning than probability formats.

Key words: probability judgment, nested-sets, conjunction fallacy, frequency format, probability format.

1 Introduction

Evidence suggests that information presented in frequency formats rather than probability formats attenuates many of the cognitive biases found in probabilistic reasoning (e.g., Kahneman, Slovic, & Tversky, 1982). There is evidence for a frequency advantage in Bayesian reasoning (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995), overcoming the overconfidence bias (Gigerenzer, Hoffrage, & Kleinbölting, 1991), and in reducing conjunction errors in extensional reasoning (Hetwig & Gigerenzer, 1999; Tversky & Kahneman, 1983). Although specific explanations for the frequency effect vary by task, the general conclusion reached by proponents of the natural frequency perspective is that presenting frequencies promotes intuitive statistical reasoning because such formats are compatible with an evolutionary-based computational algorithm (Brase, Cosmides, & Tooby, 1998; Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). Alternative explanations, as well as contradictory evidence, for the frequency effect have been offered (Evans, Handley, Perham, Over, & Thompson, 2000; Griffin & Buehler, 1999; Macchi, 2000; Mellers & McGraw, 1999; Sloman, Over, Slovak, & Stibel, 2003; Yamagishi, 2003). In particular, the nested-sets hypothesis (e.g., Sloman et al., 2003) suggests that frequency effects may be an indirect consequence of inducing a set-inclusion problem representation, which contributes to making the problem's logical structure transparent, and thus easily solvable.

According to the nested-sets hypothesis, presenting information in a way that allows people to extract subsets relative to supersets in the problem structure is the key to facilitating reasoning. Such facilitation can occur whether or not information is presented as frequencies or as probabilities, so long as the critical set-subset structure is made salient. For example, Sloman et al. (2003) tested three versions of a medical diagnosis problem first posed by Casscells, Schoenberger, and Grayboys (1978). The probability version of the problem was stated as follows:

> 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 50/1000 chance of testing positive. An individual that 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 nothing about the person's symptoms or signs? _____%

A frequency version of the problem, adapted from Cosmides & Tooby (1996) was stated as follows:

*We thank the anonymous reviewers for their valuable comments on earlier versions of this paper. We also thank Scott Standish-Parkin, Katherine Bragoni, and Clint Kuban for their assistance in collecting data. Correspondence regarding this article should be addressed to William P. Neace, Department of Psychology, University of Hartford, 200 Bloomfield Avenue, West Hartford, CT 06117. Email: Neace@hartford.edu.

> One out of every 1000 Americans 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 1000 people who are perfectly healthy, 50 of them test positive for the disease.
>
> Imagine we have assembled 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 the disease actually have the disease? _____out of_____

The third problem version was a nest-sets probability version that highlighted the set-subset structure of the problem. It was stated as follows:

> The prevalence of disease X among Americans is 1/1000. 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, the chance is 50/1000 that someone who is perfectly healthy would test positive for the disease.
>
> Imagine we have just given the test to a random sample of Americans. They were selected by lottery. Those who conducted the lottery had no information about the health status of any of these people.
>
> What is the chance that a person found to have a positive result actually has the disease? _____%

Note that, in all three problem versions, a specific reference class (i.e., 1000 Americans) is provided. What is different between problem versions, however, is that both the frequency version and the nest-sets probability version highlight the set-subset structure among the critical categories of those who test positive whether or not they have the disease (50/1000), and those who test positive and have the disease. (1/50).

The natural frequency hypothesis predicts that only the frequency problem version will facilitate Bayesian reasoning and result in an approximately correct solution (1/51 or ˜ 2%). The nested-sets hypothesis predicts that both the frequency version and the nested-sets probability version will facilitate correct responding because both versions make the set-subset relationships among the critical categories transparent. Sloman et al.'s (2003) findings support the nested-sets hypothesis. They found that the probability version resulted in significantly fewer correct answers than did either the frequency version or the nested-sets probability version. The latter two problem versions did not significantly differ in the number of correct solutions they elicited. Evans et al. (2000) ran a similar study but included a "frequency hard" condition in which the false positive rate of 5% was stated as 1/20 instead of 50/1000, thus obscuring the problem's nested-sets structure. In further support of the nested-sets hypothesis, Evans et al. (2000) found that both the probability and the frequency hard problem versions resulted in significantly fewer correct responses than did the frequency version that highlighted the set-subset structure by stating the false positive rate as 50/1000. These results are also consistent with findings reported by Macchi (2000) and Mellers and McGraw (1999), who offered similar interpretations.

Perhaps the strongest evidence in favor of the nested-sets hypothesis is provided by Yamagishi (2003), who presented a Bayesian reasoning problem in frequency and probability formats, crossed with the presence or absence of a diagrammatical representation of problem structure (a roulette wheel whose areas reflect the relative proportion of hits, misses, and false positives). He found that, in the absence of the diagram, there was a frequency effect. In the presence of the diagram, however, there was no significant difference in proportion of correct responses between frequency and probability problem formats. Such evidence suggests that Bayesian reasoning is facilitated by proper problem representation, and that frequency formats offer no additional advantage when the problem structure of the task is clarified.

Aside from Bayesian reasoning, almost no research examining the nested-sets hypothesis has been conducted on other judgment biases in which frequency effects have also been reported. It is theoretically meaningful to extend the growing evidence favoring the nested-sets hypothesis to tasks that reveal failures of extensional reasoning, such as the conjunction fallacy (Tversky & Kahneman, 1983). To date, only two published accounts have examined frequency effects in the classic Linda problem (Fiedler, 1988; Hertwig & Gigerenzer, 1999), and only one study provided a direct test of the nested-sets hypothesis for that problem (Sloman et al., 2003, Experiment 5). Both Fiedler (1988) and Hertwig and Gigerenzer (1999) reported that frequency formats led to fewer proportions of participants committing conjunction errors in comparison to probability formats. There was a potential confound in those studies, however. A reference class

https://doi.org/10.1017/S1930297500001480 Published online by Cambridge University Press

was given in the frequency version of the problem that was absent in the probability version. It is unclear from those studies whether the reduction in conjunction errors is the direct result of manipulating problem format or the result of presenting information in a manner that facilitates problem representation. To illustrate, consider the two versions of the Linda problem presented by Fiedler (1988). Both versions presented the following information:

> Linda is 31 years old, single, outspoken, and very bright. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear war demonstrations.

In the probability format, participants were asked to rank order a list of 8 statements about Linda according to their probability. Among the statements were two constituent categories ("Linda is a bank teller; "Linda is active in the feminist movement") and their conjunction ("Linda is a bank teller and is active in the feminist movement"). In the frequency format, participants were given the same information and statements to judge but were asked "To how many of 100 women who are like Linda do the following statements apply?"

Note that the frequency format asks for a numeric frequency estimate *and* provides a specific reference class out of which to make that estimate. The probability format asks for ranks and *does not* provide a reference class for the judgment. A similar confound existed in the studies reported by Hertwig & Gigerenzer (1999). The confound between response mode (numeric estimate vs. rank) and reference class (present vs. absent) prohibits a clear interpretation of the findings. Indeed, Hertwig and Chase (1998) reported that significantly more conjunction errors are committed in a ranking probability response mode than in a numeric probability estimation response mode. The frequency effect found in both the Fiedler (1988) and the Hertwig & Gigerenzer (1999) studies could be the result of confounding response mode with problem format (see Hertwig & Gigerenzer, 1999 for an extensive discussion of this issue). The facilitating effect of a frequency format could also be a secondary consequence of providing participants with a reference class and focusing their attention on sub-categories within that class, rather than be a direct result of framing the problem as a frequency judgment. Sloman et al. (2003) reported finding no significant differences between frequency and probability formats in the Linda problem when the set-inclusion relationships between critical categories were made opaque by separating constituent categories from their conjunction with seven "filler" statements. Their finding suggests that, when the set-subset structure of the problem is obscured, the frequency effect disappears. The finding is consistent with the nested-sets hypothesis.

The purpose of the present study is to examine whether frequency effects remain after controlling for the reference class confound noted in earlier research on conjunction errors. In addition, the study extends the research on nested-sets beyond problems of Bayesian inference to problems in extensional reasoning, and in doing so, tests alternative predictions derived from the natural frequency and nested-sets hypotheses.

2 Experiment 1

The first experiment sought to test whether framing the Linda problem (Tversky & Kahneman, 1983) as a frequency judgment facilitates extensional reasoning over probability formats after controlling the reference class confound discussed earlier. The natural frequency hypothesis predicts that fewer participants will commit conjunction errors when asked for frequencies than when asked to judge probabilities of the constituent categories and their conjunction. The nested-sets hypothesis predicts that participants will be no more likely to commit conjunction errors when asked for either frequencies or probabilities, as long as the problem is presented in a manner that facilitates a proper set-subset representation of the categories. The experiment manipulated problem format (ranking vs. numeric probability estimation vs. frequency) using a between-subjects design with participants being randomly assigned to the three conditions. While neither hypothesis makes a specific prediction regarding the particular type of response requested for the two probability judgment formats (i.e., rankings vs. numeric estimates), some evidence suggests that ranking probabilities produces more conjunction errors than estimating numeric probabilities (Hertwig & Chase, 1998). Thus, the ranking probability condition was included as a comparison for the frequency and numeric probability problem formats.

2.1 Method

2.1.1 Materials

The conjunction problem used was similar to the original Linda problem found in Tversky and Kahneman (1983), except that a reference class was added to the question stem at the end of the description of Linda, and only the constituent categories and their conjunction were provided for participants to judge. Three versions of the Linda problem were created by manipulating the format in which participants were asked to respond. All of the problem versions provided the following information:

https://doi.org/10.1017/S1930297500001480 Published online by Cambridge University Press

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and participated in demonstrations against capital punishment.

In the ranking probability format, the last sentence of the paragraph read:

> Please rank the probability of each statement on a scale of 1 (most probable) to 3 (least probable) as they apply to 100 women who are like Linda.

In the numeric probability format, the last sentence read:

> Please estimate the probability of each statement on a scale from 0% to 100% as they apply to 100 women who are like Linda.

In the frequency format, the last sentence read:

> To how many out of 100 women who are like Linda do the following statements apply?

The constituent categories ("Linda is a bank teller," Linda is active in the feminist movement") and their conjunction ("Linda is a bank teller and is active in the feminist movement") appeared directly beneath the description in the order just mentioned. Care was taken to insure that the statements appeared exactly the same across problem versions so that the only difference between the them was the type of response requested. The manner in which the reference class and response categories were provided is consistent with Fiedler's (1988) frequency version of the Linda problem. Mention of a concrete reference class was held constant across problem versions to control for any potential confounding effects noted in previous studies reporting a frequency effect for the Linda problem.

2.1.2 Participants

Participants were 100 introductory psychology students who were fulfilling a course requirement.

2.1.3 Procedure

Participants were run in small groups in a classroom setting. They were seated at least one chair apart to insure independence of responses between conditions. Booklets containing a randomly selected version of the Linda problem and some filler items were distributed to participants after they were seated and had signed an informed consent form. The booklets consisted of a brief demographic questionnaire that included a question asking participants if they had ever had a statistics course. The demographic questionnaire was followed by instructions informing the participants that they will be provided with some background information about a person or an event, and that they will be making some judgments based on the information. All participants received the Linda problem first, followed by the filler items. Participants worked through the materials at their own pace.

Table 1: Percentages of participants who committed conjunction errors as a function of problem format in Experiment 1.

Problem format	%	N
Ranking probability format	87.9	33
Numeric probability format	56.3	32
Frequency Format	57.1	35

N is the total number of participants in each condition.

2.2 Results and discussion

Table 1 provides data on the percentage of participants who committed conjunction errors in each of the three problem formats. Fully 88% of participants committed conjunction errors in the ranking probability format, compared to 56% of participants in the probability estimation format and 57% in the frequency format.

Overall, conjunction errors depended on problem format ($\chi^2_2 = 9.71$, p = .019). Subsequent analysis confirms that significantly more participants committed conjunction errors in the ranking probability format than in either the numeric probability format ($\chi^2_2 = 8.12$, p = .004) or the frequency format ($\chi^2_1 = 7.97$, p = .005). There was no significant difference in the number of participants who committed conjunction errors between the numeric probability format and the frequency format ($\chi^2_1 = .005$, p = .937). None of the demographic characteristics were correlated significantly with conjunction errors and so will not be further discussed.

The results suggest that frequency formats offer no advantage over probability formats when the problem format provides some cue as to how the to-be-judged categories are related. The results also suggest that the ranking probability format obscures any benefit offered by providing a concrete reference class to help clarify how the constituent categories are related to their conjunction. According to Hertwig and Chase's (1998) explanation for the response mode effect, ranking probabilities promotes

https://doi.org/10.1017/S1930297500001480 Published online by Cambridge University Press

a cue-wise strategy in which Linda's attributes are evaluated with respect to pairs of categories (e.g., bank teller vs. feminist bank teller), and ranks reflect the degree of evidential support resulting from the pair-wise comparisons. Estimating numeric probabilities promotes an integration strategy in which the categories are evaluated independently with respect to the evidential support provided by Linda's attributes, with the resulting probability estimates reflecting the degree of support separately for each constituent category and their conjunction. Sloman et al. (2003) offer a simpler explanation for the response mode effect in which rankings force a choice among categories. Since only one rank can be assigned to each statement about Linda, conjunction errors may result from restricting the fuller range of responses provided by estimating a numeric probability.

The present experiment was not designed to distinguish between the two explanations, however. It can only be concluded that ranking produced significantly more conjunction errors than either the numeric probability format or the frequency format. The results suggest that initial findings in favor of a frequency effect in conjunction problems (Fiedler, 1998; Hertwig & Gigerenzer, 1999) may have been the indirect consequence of providing a reference class in frequency formats that was absent in probability formats. When the reference class confound was controlled in this experiment, no evidence of a frequency advantage was found. The results suggest that providing a reference class might influence how the problem is structured, and that such facilitation can occur regardless of problem format.

3 Experiment 2

Experiment 2 was conducted to further investigate whether the lack of a frequency effect found in the first experiment was indeed due to the facilitating effect of framing the problem in a way that makes salient the set-subset structure of the relationship among the categories. That is, it sought to more closely examine participants' strategies for making their judgments. It was anticipated that the second experiment would replicate the findings of the previous experiment, and provide evidence that reasoning strategies were more closely aligned with extensional reasoning than with non-extensional reasoning in the presence of a reference class.

3.1 Method

3.1.1 Materials

The materials used for Experiment 1 were adapted for use in this experiment. Two changes were made to the original stimuli. First, in the frequency format, participants were asked to provide their frequency estimates in a manner consistent with the studies reported by Hertwig and Gigerenzer (1999): The last sentence of the frequency version read as follows:

> To how many out of 100 women who are like Linda do the following statements apply?
>
> _____ out of _____ are bank tellers.
>
> _____ out of _____ are active in the feminist movement.
>
> _____ out of _____ are bank tellers and are active in the feminist movement.

The other change was that participants were also asked to provide a written explanation for how they made their judgments after providing their responses.

3.1.2 Participants

Participants were 108 introductory psychology students fulfilling a course requirement.

3.1.3 Procedure

The procedures for this experiment mirrored those used in Experiment 1. An additional set of instructions that asked participants to elaborate on their judgment strategies, particularly on how they represented the given information (Linda's attributes) and the relationship among the categories for which they provided their judgments, was included on the last page of the booklets. The instructions suggested to participants that they could use a diagram to illustrate how they represented the relationships among the information and the categories. Participants were told that they could refer back to the problems when thinking about their explanations but were asked not to change their initial judgments. Inspection of the completed booklets provided no evidence that participants changed their judgments after providing their written explanations.

3.2 Results and discussion

Table 2 provides the data on the percentage of participants who committed conjunction errors in each problem format condition of Experiment 2. The ranking probability format produced the highest percentage of conjunction errors (82%), followed by the numeric probability format (71%) and the frequency format (33%). As in Experiment 1, conjunction errors depended on problem format (χ^2_1 = 20.68, p < .001).

Subsequent analysis indicates that significantly fewer participants committed conjunction errors in the frequency format compared to the numeric probability format (χ^2_1 = 10.72, p = .001) and the ranking probability

https://doi.org/10.1017/S1930297500001480 Published online by Cambridge University Press

Table 2: Percentages of participants who committed conjunction errors as a function of problem format in Experiment 2.

Problem format	%	N
Ranking probability format	82.4	34
Numeric probability format	71.4	35
Frequency format	33.3	39

N is the total number of participants in each condition.

Table 3: Percentages of participants exhibiting one of three reasoning strategies in the ranking probability, numeric probability, and frequency problem formats in Experiment 2.

Reasoning Strategy	Rank Probability	Frequency	Total
Representativeness	15.7	13.9	9.3	38.9
Extensional	5.6	6.5	1.9	13.9
Other	11.1	11.2	25.0	47.2

format (χ^2_1 = 7.73, p < .001). The difference in the number of participants who committed conjunction errors between the ranking and the numeric probability formats was not significant (χ^2_1 = 1.16, p = .279). Surprisingly, the pattern of results in Experiment 2 did not replicate the pattern of results from the first experiment. The findings show a frequency effect, and failed to show a significant difference between the ranking probability format and the numeric probability format. This inconsistency is explored further in Experiment 3.

Judgment strategies were also collected in this experiment. Participants' written explanations for how they made their judgments were evaluated by two independent raters blind to the purpose of the study. Participants' strategies were coded into three categories: Those that showed evidence of extensional reasoning, those that showed evidence of reasoning using the representativeness heuristic, and an other category. Inter-rater reliability was quite acceptable (.98), and disagreements were resolved by discussion. The percentage of participants falling into the three strategy categories within each problem format is presented in Table 3.

Two findings are of interest in the pattern of strategy use. The first is that most participants either used a representativeness strategy (39%), in which they based their judgments on the similarity between Linda's description and a prototypical category exemplar, or some strategy other than an extensional one (47%). The other noteworthy finding is that, of the 13.9% (15 participants) who reported using some form of extensional reasoning strategy, only 1.9% (2 participants) were in the frequency format condition.

Also of interest is that none of the participants represented their reasoning strategies diagrammatically, indicating that merely suggesting use of a diagram did not affect participants' post-hoc descriptions of their judgment strategies.

Overall, 74% of participants who reported using a representativeness-based reasoning strategy committed conjunction errors, compared to only 40% who reported using an extensional strategy and 57% whose strategies fell into the "other" category. The differences were significant (χ^2_1 = 6.05, p = .049), indicating that correct reasoning resulted in fewer errors. Of more interest is that problem format also influenced strategy selection (χ^2_4 = 12.77, p = .012), with significantly more participants in the frequency format condition using a strategy other than one based on either representativeness or extensional reasoning (χ^2_2 = 8.94, p = .011). Even though participants in the frequency format were less likely to use a representativeness strategy, they did not compensate for this by replacing it with a more appropriate extensional reasoning strategy. Of the 15 participants who reported using some form of extensional reasoning, 9 of them did not commit a conjunction error. Of those 9, 4 were in the ranking probability format, 3 were in the numeric probability estimation format, and 2 were in the frequency format.

Together, the findings indicate that participants who used the correct reasoning strategy were less likely to commit a conjunction error; however, there is no compelling evidence to suggest that more people in the frequency format used an appropriate strategy compared to the other two problem formats. Since all participants were given a reference class, it is not possible to determine whether their strategies might be different in the absence of a reference class. This issue is explored further in Experiment 4.

4 Experiment 3

Experiment 3 was designed to investigate the discrepancy in the findings between Experiments 1 and 2. One reason for the difference in the pattern of findings is that the manner in which the frequency response was solicited differed between the two experiments. To test this possibility, four versions of the Linda problem were used: Ranking probability, numeric probability estimation, and two versions of the frequency estimation format. One version of the frequency format from Experiment 1 was used, and the other version of the frequency format from Experiment 2 was used. For clarity, the frequency format from Experiment 1 will be referred to as the "frequency

https://doi.org/10.1017/S1930297500001480 Published online by Cambridge University Press

Table 4: Percentages of participants who committed conjunction errors as a function of problem format in Experiment 3.

Problem format	%	N
Ranking probability format	79.0	62
Numeric probability format	60.0	60
Frequency-hard format	56.4	55
Frequency-easy format	35.6	59

N is the total number of participants in each condition.

hard" format, and the frequency format from Experiment 2 will be referred to as the "frequency easy" format.

4.1 Method

4.1.1 Materials

The materials used in the previous two experiments were used here.

4.1.2 Participants

Participants were 236 introductory psychology students fulfilling a course requirement.

4.1.3 Procedure

The procedures for this experiment are the same as those used in the Experiment 1. Participants were randomly assigned to receive the Linda problem in one of four problem formats (rank, numeric probability estimate, frequency hard format, and frequency easy format).

4.2 Results and discussion

The percentage of participants who committed conjunction errors in each of the four problem formats is presented below in Table 4. As in Experiments 1 and 2, conjunction errors depended on problem format (χ^2_3 = 23.59, p < .001). Subsequent analyses indicated that more participants committed conjunction errors in the ranking format than in the numeric probability or frequency-hard formats (χ^2_2 = 7.86, p = .02), which did not significantly differ from each other. This pattern of results is consistent with the findings from Experiment 1.

Subsequent analysis also indicated that there were fewer participants who committed conjunction errors in the frequency-easy format than in the ranking or numeric probability formats (χ^2_2 = 23.58, p < .001), consistent with findings from Experiment 2. Further, more participants committed conjunction errors in the frequency hard format than in the frequency easy format (χ^2_1 = 4.95, p = .026).

Taken together, the findings from Experiment 3 indicate that, once the reference class confound was controlled, participants were no more likely to commit conjunction errors when the problem was formatted in terms of probabilities or frequencies, unless the particular frequency format used serves to focus attention fully on the logical constraint imposed by the conjunction rule. Asking participants "to how many out of 100 women like Linda" the categories bank teller, feminist, and feminist bank teller apply is evidently different than asking them to estimate a frequency for these categories in the form "_____ out of _____ (given 100 women like Linda)." Though merely providing a reference class might benefit participants in either a probability or a frequency format, the findings suggest that the frequency-easy format provides an additional benefit by making concrete the fact that participants are reasoning about categories that have a set-subset structure.

5 Experiment 4

The results from Experiments 1 and 3 provide consistent evidence that including a reference class reduces the likelihood of committing conjunction errors whether or not participants are asked for a probability judgment or a frequency judgment. The results of Experiment 3 also offer an explanation for the discrepant findings between Experiments 1 and 2 — the manner in which a frequency estimate is requested impacts the likelihood of committing an error. The frequency-hard version of the Linda problem may be similar to the numeric probability version of the problem in the way participants represent it when they are given a reference class. On the other hand, the frequency-easy version of the problem not only provides participants with a reference class but also focuses their attention on structuring a response that makes highly salient the set-subset structure of the problem (e.g., how many .... _____ out of _____). The difference in conjunction errors in the two frequency conditions of Experiment 3 suggests that the frequency-easy format produces a stronger effect on problem representation than does the frequency-hard format. In essence, the frequency-easy format structures participants' responses for them, making salient the logical constraint imposed by the conjunction rule.

Two experiments provide consistent evidence that participants are no more likely to commit conjunction errors in a probability format than they are in a frequency format when a concrete reference class is provided. The findings suggest that frequency effects may well be a secondary consequence of the manner in which the problem

https://doi.org/10.1017/S1930297500001480 Published online by Cambridge University Press

is structured. Stronger evidence would come from examining whether fewer conjunction errors occur when a reference class is provided compared to when it is absent, and whether reasoning strategies also differ as a function of the presence or absence of a reference class. Experiment 4 tested these hypotheses using a 2 (reference class: present vs. absent) by 3 (problem format: rank vs. numeric probability vs. frequency) between-subjects factorial design.

5.1 Method

5.1.1 Materials

Six versions of the Linda problem were created. There were 3 versions of the problem that contained a reference class of "100 women who are like Linda" for each problem format (ranking probability, numeric probability estimation, frequency estimation), and three versions of the problem that did not contain a concrete reference class. In the reference class present condition, participants were given the usual description of Linda, with the question stem asking them to "imagine 100 women who are like Linda." The only difference between problem formats was in the type of response requested. The last sentence in the ranking probability format read:

> Please rank order the following according to the probability that out of 100 women who are like Linda ....

The last sentence in the numeric probability format read:

> Please estimate the probability that out of 100 women who are like Linda ....

In both the ranking and numeric probability formats, the constituent categories and their conjunction were presented as single-event probabilities (e.g., please rank [estimate] the probability that out of 100 women who are like Linda, a randomly chosen woman (italics added for emphasis) is a [bank teller, feminist, bank teller and feminist]). The last sentence of the frequency estimation format read:

> How many out of 100 women who are like Linda are ....

The question stem was followed by the constituent categories (bank tellers, feminists) and their conjunction (bank tellers and feminists). Note that the frequency version used here did not focus attention specifically on responding with "____ out of _____" to maintain consistency in the frequency formats used for both reference class present and reference class absent conditions.

In the reference class absent condition, a specific reference class was not provided. Instead, the question stems following Linda's description asked participants to:

> Rank order the following according to their probability (ranking condition) ....
>
> Estimate the probability that: (numeric estimation condition) ....

Since the frequency estimation condition called for a frequency estimate, the last sentence of that version read as follows:

> Imagine that there are other women like Linda. How many of those women are: ....

The question stems were followed by the constituent categories and their conjunction. Notice that the only difference between reference class conditions is that either a specific, concrete reference class was provided in the problem description or it was not provided. Care was taken to insure that the problem format conditions remained highly consistent with each other in all respects save for whether a reference class was or was not provided.

5.1.2 Participants

Participants were 193 introductory psychology students fulfilling a course requirement.

5.1.3 Procedure

The procedures for this experiment are similar to those used in the previous experiments. Participants were randomly assigned to conditions. They received their booklets with the Linda problem and other filler items, provided their judgments in accord with the condition to which they had been assigned, and were then asked for their judgment strategies.

5.2 Results and discussion

Table 5 reports the percentage of participants who committed conjunction errors as a function of reference class and problem format. Separate χ^2 analyses were conducted to examine the effects of reference class and problem format, and to investigate whether the effect of problem format depended upon the presence of a reference class (i.e., to assess the "interaction" of these factors). Looking first at the effect of reference class on conjunction errors, the analysis indicates that significantly more participants committed conjunction errors when a reference class was absent than when a reference class was

https://doi.org/10.1017/S1930297500001480 Published online by Cambridge University Press

Table 5: Percentages of participants who committed conjunction errors as a function of reference class and problem format.

	Problem format		
Reference class	Ranking	Probability	Frequency
Present	74	33	31
Absent	71	68	60

present (χ^2_1 = 7.58, p = .006). In the absence of a reference class, 66% of participants committed conjunction errors as compared to only 46% when a reference class was present.

Problem format also produced a significant effect on conjunction errors (χ^2_2 = 11.76, p = .003), with 73% of participants committing conjunction errors in the ranking format, compared to 51% in the probability format and 46% in the frequency format. Subsequent analysis indicated that the ranking format significantly differed from both the probability format (χ^2_1 = 6.79, p = .009) and the frequency format (χ^2_1 = 10.60, p = .001), with no significant difference in the number of participants who committed conjunction errors between the probability and frequency formats. The latter finding must be interpreted in light of the interaction between reference class and problem format, however.

As can be seen in Table 5, there was little difference in percentages of participants who committed conjunction errors in each problem format condition when a reference class was absent (bottom row), but there were larger differences in those percentages when a reference class was present (top row). That is, it looks like problem format interacted with reference class such that the problem format effect depended upon whether or not a reference class was present. To examine this possibility, separate χ^2 analyses were conducted for problem format in the reference class present condition and in the reference class absent condition. When a reference class was absent, there were no significant differences in the number of participants who committed conjunction errors in each problem format condition. When a reference class was present, however, problem format did have a significant effect on conjunction errors (χ^2_1 = 16.12, p < .001). Looking at the top row of Table 5, 74% of participants committed conjunction errors in the ranking condition, compared to only 33% in the probability condition and 31% in the frequency condition. There was no significant difference between the numeric probability and frequency problem formats in that condition. These results replicate those from Experiments 1 and 3, and provide further evidence that the frequency effect depends upon the presence of a reference class, and that the numeric probability and frequency formats do not significantly differ when the reference class confound is controlled.

Table 6: Percentage of participants exhibiting one of four reasoning strategies in the reference class present versus reference class absent conditions in Experiment 4.

	Reference class		
Reasoning strategy	Present	Absent	Total
Reference class used	14.8	1.0	15.8
Conjunction rule used	8.2	7.1	15.3
Representativeness used	18.9	24.5	43.4
Other	9.7	15.8	25.5

Participants' reasoning strategies were also examined as a function of both problem format and presence versus absence of a reference class. Reasoning strategies were evaluated by two rates blind to the experimental conditions, and inter-rater reliability was acceptable (.96). Differences were resolved by discussion. Four categories of reasoning strategy were derived from participants' written responses. Participants who mentioned using a reference class as part of their reasoning strategy were coded 1, those who mentioned using the conjunction rule were coded 2, those whose strategies were largely based on representativeness were coded 3, and those whose responses did not provide sufficient information to classify their strategies were coded 4. The percentage of participants using one of these reasoning strategies in the reference class present and reference class absent conditions is presented in Table 6. An analysis of the relationship between reasoning strategy and conjunction errors indicated that 71.4% of participants whose strategies involved using the representativeness heuristic committed conjunction errors, compared to only 35.5% whose strategies involved using the given reference class, and 31.1% whose strategies were based on the conjunction rule.

Thus, participants' performance was consistent with their stated problem strategies. More interesting for the purposes of the present study is to examine reasoning strategy as related to both reference class condition and problem format. Reasoning strategy depended upon whether or not a reference class was provided (χ^2_3 = 27.79, p < .001). The data suggest that the participants do not spontaneously structure the Linda problem by using a reference class when it is not provided for them. In that case, they are more likely to use representativeness or some other strategy.

The percentage of participants who used one of the four reasoning strategies in each problem format condition is presented in Table 7. Reasoning strategy also depended on problem format (χ^2_6 = 28.90, p < .001).

https://doi.org/10.1017/S1930297500001480 Published online by Cambridge University Press

Table 7: Percentages of participants exhibiting one of four reasoning strategies in the ranking probability, numeric probability, and frequency problem formats in Experiment 4.

Reasoning strategy	Rank	Probability	Frequency	Total
Reference class	2.0	3.6	10.2	15.8
Extensional	8.7	5.1	1.5	15.3
Representativeness	16.3	17.3	9.7	43.4
Other	8.2	8.2	9.2	25.5

The findings suggest that more participants in the frequency format mentioned using the reference class as part of their reasoning strategy, and were less likely to use a strategy based on representativeness compared to the two other problem formats. The findings show no evidence that frequency formats promote better statistical reasoning than do probability formats, however, since use of the conjunction rule was mentioned by more participants in the two probability problem formats. Thus, consistent with findings from Experiment 2, participants in the frequency format were less likely to use a representativeness reasoning strategy but were not more likely to replace it with an appropriate reasoning strategy (i.e., the conjunction rule).

Taken together, the findings from Experiment 4 support the conclusion that frequency effects are a consequence of confounding reference class with problem format, and when that confound is controlled, as in this case by systemically manipulating reference class, there is no longer evidence to suggest that frequency formats reduce conjunction errors relative to numeric probability estimation formats. The findings do suggest that participants were using the given reference class information to help them structure their problem representations. The findings indicate that it is the representation of the problem, not problem format, that reduces judgment error.

6 Experiment 5

A reference class size of 100 was used in the previous four experiments. There is the possibility that using 100 might somehow give an undue advantage to the numeric probability estimation condition, and that the reason there were no significant differences between probability and frequency problem formats is because using a reference class size of 100 somehow psychologically equates the frequency and the numeric probability formats. Frequency effects may well re-emerge when different reference class sizes are used. Experiment 5 was conducted to test this hypothesis. A 2 (problem format: numeric probability vs. frequency estimation) by 3 (reference class size: 50 vs. 100 vs. 279) factorial design was used to investigate whether frequency effects depend upon the size of the reference class. Given the insensitivity of the ranking probability format to reference class (as evidenced in Experiments 1 through 4), it was excluded in order to focus on the two problem formats of theoretical interest here.

Table 8: Percentages of participants who committed conjunction errors as a function of reference class size and problem format.

	Reference class size		
Problem format	Size = 50	Size = 100	Size = 279
Frequency	40	41	46
Probability	44	35	43

6.1 Method

6.1.1 Materials

Six versions of the Linda problem were used. The numeric probability estimation and frequency estimation problem formats from Experiment 4 were used with the only difference being the size of the reference class that was specified. Reference class sizes of 50, 100, and 279 were specified for each problem format.

6.1.2 Participants

Participants were 182 introductory psychology students fulfilling a course requirement.

6.1.3 Procedure

The procedures for this experiment are similar to those used in the previous experiments. Participants were randomly assigned to conditions in a 2 (problem format) by 3 (reference class size) between-subjects factorial design. They received their booklets with the Linda problem and other filler items, and provided their judgments in accord with the condition to which they had been assigned.

6.2 Results and discussion

Table 8 reports the percentage of participants who committed conjunction errors within each condition of Experiment 5. Overall, 42% of participants committed conjunction errors in the frequency format, and 41% committed conjunction errors in the probability format. The difference was not statistically significant (χ^2_1 = .57, p =

https://doi.org/10.1017/S1930297500001480 Published online by Cambridge University Press

.85). Overall, 42% of participants committed conjunction errors for a reference class size of 50, 38% committed conjunction errors for a reference class size of 100, and 46% committed conjunction errors for a reference class size of 279. These differences were not statistically significant either (χ^2_2 = .57, p = .75).

Subsequent analyses were conducted to examine differences in conjunction errors between probability and frequency formats within each reference class size condition, with no significant results obtaining (for all χ^2 analyses, p > .10). These results fail to provide any evidence that using a reference class size of 100 gives an unfair advantage to the probability format over the frequency format. There is also no evidence to suggest that frequency effects re-emerge when reference class sizes other than 100 are used. Thus, the findings indicate that, when a reference class is present, participants are no more likely to commit conjunction errors in frequency formats than they are in probability formats, regardless of the size of the reference class.

7 General Discussion

A series of five experiments were conducted to investigate whether frequency formats facilitate extensional reasoning over probability formats using the classic Linda conjunction problem (Tversky & Kahneman, 1983). Previous studies that have reported a frequency effect for that problem (Fiedler, 1988; Hertwig & Gigerenzer, 1999) inadvertently confounded problem format with whether or not a concrete reference class was provided, and also with the type of response that was requested. A reference class (e.g., "imagine there are 100 women like Linda") was provided in frequency versions of the Linda problem but not in probability versions. Moreover, participants in frequency versions were asked to estimate a number (frequency) while participants in the probability versions were asked to provide probability rankings. Thus, while those studies reported that fewer conjunction errors were committed in frequency formats than in probability formats, it is not possible to unambiguously conclude that frequency formats facilitate probabilistic reasoning in this task. According to the natural frequency view (Brase, Cosmides, & Tooby, 1998; Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002), information from the environment comes in the form of frequencies that are experienced directly, and it is this mode of information acquisition to which the human mind has become attuned through evolution. That is, a computational algorithm evolved to respond to event frequency rather than probability. Moreover, the term "probability" is itself a vague term because it carries more than one interpretation, for example, as "degree of belief," or "plausibility of an assertion" (Fiedler, 1988; see also Gigerenzer, 1994 and Hertwig & Gigerenzer, 1999 for a detailed discussion of this issue). Accordingly, asking for probability judgments results in error by introducing semantic ambiguity. It follows from the natural frequency hypothesis that judgments based on information provided as frequencies will exhibit less error than judgments based on the same information presented as probabilities.

A plausible alternative interpretation for the frequency effect is that it is a secondary consequence of presenting information in a way that clarifies the set-subset relation between the constituent categories (bank teller, feminist) and their conjunction (bank teller and feminist). This explanation is consistent with the nested-sets hypothesis (e.g., Sloman et al., 2003), which is the general claim that making the logical structure of a probability problem transparent facilitates coherent judgments. Such transparency may be the result of the manner in which the problem is represented, for example, as one that involves making judgments about multiple instances of given categories. Frequency formats necessarily engender such a representation because they provide a concrete reference class as part of the problem statement. The nested-sets hypothesis predicts that any manipulation that facilitates problem representation will reduce errors in probability judgments regardless of whether a probability format or a frequency format is used.

The majority of the findings reported in this study provide support for the nested-sets hypothesis. After controlling for the reference class confound noted in previous studies of the Linda problem, we found that participants were no more likely to commit conjunction errors when the problem was formatted as probabilities than they were when a frequency format was used (Experiments 1, 3, 4, and 5). When reference class was manipulated in Experiment 4, we found that significantly more participants committed conjunction errors in the absence of a reference class than committed the error when a reference class was provided. We also found that the effect of problem format depended upon whether or not a reference class was provided. When a reference class was not provided, there were no significant differences in the number of participants who committed conjunction errors in a ranking probability format, a numeric probability estimation format, or a frequency format. When a reference class was provided, however, significantly more participants committed conjunction errors in a ranking probability format than they did in either a numeric probability format or a frequency format. In the first four experiments, a reference class size of n = 100 was used, which may have given an unfair advantage to the numeric probability format over the frequency format, thus offering an alternative explanation for why there was no signifi-

https://doi.org/10.1017/S1930297500001480 Published online by Cambridge University Press

cant difference in conjunction errors between the numeric probability estimation and frequency formats. An additional experiment that manipulated the size of the reference class (n = 50, n = 100, n = 279) also found no significant differences between numeric probability and frequency formats, however. The findings from Experiment 5 rule out the alternative explanation that a reference class size of 100 psychologically equates probability and frequency judgments, and provide further evidence that reference class, not problem format, influences judgments.

Taken together, the findings reported in this study do not provide compelling evidence for the advantage of a frequency format over a numeric probability estimation format. The exception to this conclusion might have been found in the results from Experiment 2, in which more participants committed conjunction errors in the ranking probabilities and numeric probability estimation formats than they did in the frequency format, contrary to the results obtained in Experiment 1. Experiment 3 resolved this discrepancy, however, by showing that the difference in findings between Experiment 1 and Experiment 2 could well have been the result of using two different formulations of the frequency format for the Linda problem. In Experiment 2, the frequency format used may have actually structured participants' responses for them by giving them a concrete reference class ("imagine that there are 100 women who are like Linda") and asking them to use it while providing their frequency estimates (e.g., "How many women like Linda are bank tellers? _____ out of _____"). Interpreting the findings from Experiment 2 as evidence of a frequency effect rather than as evidence for the facilitating effect of problem structure seems logically inconsistent with the findings from the other four experiments reported in this study; however, such an interpretation cannot be completely ruled out at the present time.

Overall, the results from this study suggest that improving the coherence of probability judgments rests largely with whether or not a reference class is presented as part of the problem description, and is not dependent upon whether the problem is formatedd in terms of estimating numeric probabilities or frequencies (Experiments 1, 3, 4 and 5). The findings also provide evidence that frequency effects do not occur in the absence of a reference class, indicating that such effects may be a secondary consequence of how the problem is structured (Experiment 4). Why so many participants commit conjunction errors in a ranking probability format is still an open and interesting question, and one that future research could address. In particular, it is intriguing to note that the ranking probability format was insensitive to the presence or absence of a reference class (Experiment 4).

The findings from this study further suggest that having a reference class might facilitate representing the judgment problem in a manner that highlights the set-subset structure of task, and that it is in forming a better problem representation that the mechanism by which errors are reduced lies. Some preliminary evidence in support of this explanation was provided by participants' reasoning strategies, which were collected in Experiments 2 and 4. Recall that Experiment 2 controlled the reference class confound by holding it constant across the three problem formats. Initially, we examined participants reasoning strategies within each problem format condition to determine whether frequency formats promote more statistically sophisticated reasoning strategies compared to the two probability formats. We found that, although fewer participants in the frequency format reported using a representativeness-based strategy compared to the two probability formats, they were also less likely to report using an extensional reasoning strategy. The findings from this study, consistent with those from Griffin and Buehler (1999), provide no strong evidence that reframing probabilistic judgment tasks in terms of frequencies results in better statistical reasoning.

In Experiment 4, we collected participants' reasoning strategies in order to examine if they differed depending upon whether or not a specific reference class was provided in the problem statement. There, we found that more participants used a strategy based on representativeness or some other strategy when a reference class was absent than they did when one was provided for them. In addition, we found that more participants mentioned the reference class as part of their reasoning strategy when one was provided than they did when one was not provided. We also found that more participants mentioned using some form of extensional reasoning (e.g., the conjunction rule) in the two probability formats than they did in the frequency format, consistent with the findings from Experiment 2.

Taken together, the findings from the analysis of reasoning strategies suggest that having a reference class specified as part of the problem description reduces the likelihood that participants will adopt a normatively inappropriate reasoning strategy, and increases the likelihood that they will use the reference class to somewhat structure their judgments. There is no evidence that frequency formats produce better statistical reasoning, however. Indeed, the evidence suggests that the application of the normative reasoning strategy is more likely to be seen when the problem is formatted in terms of probabilities then when it is formatted in terms of frequencies. The fact that there was no significant relationship between conjunction errors and whether or not participants had previously had a course in statistics rules out an alternative explanation — namely, that prior training in statistics results in better statistical reasoning. Perhaps most interesting is that the results suggest that participants are not likely to

spontaneously generate their own reference class to use in structuring their judgments, even when they are asked to provide frequency estimates but are not given a specific reference class out of which to provide their responses. Only two participants mentioned a reference class in the frequency format with reference class absent condition of Experiment 4. One participant used "the U.S. Census of about 3 million people," and the other participant said that having a specific number of people out of which to provide frequency estimates would have been useful.

Of course, asking participants to analyze their reasoning strategies post-hoc is a limitation of the methodology used in this study. Participants may not have thought to use a particular strategy until after they had already made their judgments. Indeed, the effect of providing a reference class may be a psychologically subtle one that initially effects system 1 processes but is not strong enough to impact system 2 processes (Stanovich, 1999; Stanovich & West, 2000) unless otherwise prompted (e.g., by asking for post-hoc explanations of judgments). One participant actually wrote that "you can't have more bank tellers who are feminists than you have bank tellers ... I didn't think of that earlier ...." Nevertheless, when we analyzed the relationship between stated reasoning strategy and conjunction errors, we found that most participants who stated using an inappropriate reasoning strategy (e.g., representativeness or "other") committed conjunction errors, compared to participants who mentioned using the reference class or who stated using some form of extensional reasoning strategy. Thus, it appears as though participants' post-hoc reasoning strategies were consistent with their objective judgments. It might have been better to use a "write aloud" protocol, or to actually manipulate reasoning strategy (e.g., most people base their judgments on ...). Future research could benefit by routinely collecting information on how participants are reasoning about the task at hand, and such data would be useful to assist in advancing our theoretical understanding of the processes that underlie judgments under uncertainty.

References

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

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

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

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50,* 123–129.

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

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Griffin, D., & Buehler, R. (1999). Frequency, probability and prediction: Easy solutions to cognitive illusions? *Cognitive Psychology, 38*, 48–78.

Hertwig, R., & Chase, V. M. (1998). Many reasons or just one: How response mode affects the conjunction problem. *Thinking and Reasoning, 4,* 319–352.

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12,* 275–305.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heurstics and biases.* New York: Cambridge University Press.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23,* 645–726.

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

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

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

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

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