Available online at www.sciencedirect.com

SCIENCE DIRECT®

Organizational Behavior and Human Decision Processes 99 (2006) 202–211

ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES

www.elsevier.com/locate/obhdp

Differences between probability and frequency judgments: The role of individual differences in working memory capacity ☆

Amber Sprenger *, Michael R. Dougherty

Department of Psychology, University of Maryland, College Park, MD 20782, USA

Received 31 August 2004
Available online 19 September 2005

Abstract

Most theories of probability judgment assume that judgments are made by comparing the strength of a focal hypothesis relative to the strength of alternative hypotheses. In contrast, research suggests that frequency judgments are assessed using a non-comparative process; the strength of the focal hypothesis is assessed without comparing it to the strength of alternative hypotheses. We tested this distinction between probability and frequency judgments using the alternative outcomes paradigm (Windschitl, Young, & Jenson, 2002). Assuming that judgments of probability (but not judgments of frequency) entail comparing the focal hypothesis with alternative hypotheses, we hypothesized that probability judgments would be sensitive to the distribution of the alternative hypotheses and would be negatively correlated with individual differences in working memory (WM) capacity. In contrast, frequency judgments should be unrelated to the distribution of the alternatives and uncorrelated with WM-capacity. Results supported the hypotheses.

Keywords: Probability judgment; Frequency judgment; Support theory; Working memory

Introduction

Consider the following two types of judgments. A doctor stands before file cabinets that contain all of her patients' records for her entire career and estimates how many of the patients died from cancer. Or, the doctor estimates the likelihood that a randomly chosen file would be of a patient who died from cancer. Normatively, these two types of judgments should have a high degree of correspondence, because the probability of pulling a file of a person who died from cancer should reflect the frequencies of patients who died from cancer versus patients who did not die from cancer. Despite this normative isomorphism, research and theory suggests that there are fundamental differences in the processes underlying judgments of probability and frequency. Moreover, these differences in process have often been purported to underlie the common finding that frequency formats lead to improved accuracy on Bayesian inference problems (Cosmides & Tooby, 1996; Fiedler, 1988; Hoffrage, Gigerenzer, & Krauss, 2002; Gigerenzer & Hoffrage, 1995; but see Sloman, Over, Slovak, & Stibel, 2003).

The purpose of our research was twofold. Our primary purpose was to investigate differences in the cognitive processes used to make probability judgments versus frequency judgments. Although theories of probability judgment assume that focal hypotheses[1] are com-

☆ This material is based on work supported by the National Science Foundation under Grant SES-0134678. Order of authorship is arbitrary. The authors thank Steven Pipe for help in data collection and Paul Windschitl for providing us with the experimental stimuli.

* Corresponding author.
E-mail address: asprenger@psyc.umd.edu (A. Sprenger).

[1] Throughout the introduction, we will use the term focal hypothesis to indicate the hypothesis being considered for judgment.

0749-5978/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.obhdp.2005.08.002

pared with alternative hypotheses and theories of frequency judgment do not assume such a comparison process, few direct tests of differences in comparative versus non-comparative judgment processes in probability and frequency judgment have been published. The second purpose of our research was to investigate the accuracy of frequency and probability judgments. As mentioned above, considerable research supports the idea that frequency judgments often are more accurate than probability judgments. Interestingly, however, most studies comparing frequency and probability judgments have relied almost exclusively on what might be termed *absolute accuracy*–measures that essentially amount to an overconfidence measure (i.e., the degree to which judgments exceed an objective standard). Much less research has examined differences in *relative accuracy*, where relative accuracy is defined as one's ability to discriminate between events that have different objective probabilities. It is important to examine both measures of accuracy, because one's definition of accuracy can be pivotal in determining whether or not one kind of judgment is more accurate than the other (Treadwell & Nelson, 1996).

The present research differs from prior investigations of frequency and probability judgments in two important ways. First, most conclusions that frequency judgments are more accurate than probability judgments have been based on experiments using word-based problems (e.g., Tversky & Kahneman's, 1983 Linda problem) or general knowledge questions. In contrast, in the present study we used a learning-based paradigm where participants learned the objective frequencies of the to-be-judged events. A second major difference between our investigation and previous ones is that we examined two aspects of judgment accuracy: absolute accuracy and relative accuracy. In contrast, most studies comparing probability judgment to frequency judgment have relied exclusively on only one measure of accuracy—absolute accuracy (Cosmides & Tooby, 1996; Fiedler, 1988; Gigerenzer & Hoffrage, 1995; Hoffrage et al., 2002; but see Sloman et al., 2003). Thus, claims that frequency judgments are more accurate than probability judgments have been based on only one definition of accuracy.

Comparative versus non-comparative judgment

One of the primary differences between probability and frequency judgment theories is that probability judgments assume that participants compare the strength of the focal event (e.g., cancer) with the strength of a set of alternative events (e.g., no cancer) (Dougherty, Gettys, & Ogden, 1999; Tversky & Koehler, 1994). In contrast, theories of frequency judgment (Hintzman, 1988; Shiffrin, 2003; Murdock, Smith, & Bai, 2001) assume that participants assess a strength or familiarity dimension of only the focal event (e.g., cancer) and map that feeling onto a frequency scale. Thus, the crucial difference between models of probability and frequency judgment is that probability judgments include a comparison of the focal event with the alternative events.

Support theory (Tversky & Koehler, 1994) provides a general theoretical framework for describing the process of comparing alternative events assumed for probability judgment:

$$P(\mathrm{A},\mathrm{B}) = \frac{s(\mathrm{A})}{s(\mathrm{A}) + s(\mathrm{B})}, \tag{1}$$

where $P(\mathrm{A},\mathrm{B})$ represents the probability of hypothesis A rather than hypothesis B, $s(\mathrm{A})$ represents the support for the focal hypothesis A, and $s(\mathrm{B})$ represents the support for the alternative hypotheses entertained by the decision maker. Tversky and Koehler (1994) noted that people's judgments tend to be subadditive; the probability of an implicit disjunction tends to be lower than the sum of the probabilities assigned to its elements. For example, if one were to judge *p*(cancer, no cancer) it would be judged as less likely than the sum of the probabilities assigned to *p*(lung cancer, no lung cancer), *p*(breast cancer, no breast cancer), *p*(skin cancer, no skin cancer), and *p*(all other cancer). Thus, the judged probability of the inclusive hypothesis, *p*(cancer, no cancer) is subadditive with respect to the sum of the judged probabilities of its elements.

One way to conceptualize the assessment of support of the alternative hypothesis, $s(\mathrm{B})$, within support theory is in terms of the generation of the alternative hypotheses contained within B. Consider Eqs. (2) and (3). Let h represent the focal hypothesis, A, and $\neg h_{\mathrm{i}}$ represent the ith element of the set of B alternative hypotheses. Ideally, people should consider the support for all relevant alternatives when making probability judgments:

$$P(h) = \frac{s(h)}{s(h) + \sum_{i=1}^{N} s(\neg h_i)}, \tag{2}$$

where N is the total number of alternative hypotheses contained in B. Without assuming that participants underestimate the alternatives, $\neg h_i$, Eq. (2) predicts that subjective judgments should be additive. However, if people are limited in the number of alternatives that they can consider at any one point in time, the overall support for the alternatives will be underestimated and the probability of the focal will be overestimated. Dougherty and Hunter (2003a, 2003b) found that the judged probability of a focal hypothesis decreased as the number of alternative hypotheses people generated increased. Thus, we propose that probability estimation is constrained by the number of alternative hypotheses one can compare at one time, as represented in Eq. (3):

$$P(h) = \frac{s(h)}{s(h) + \sum_{i=1}^{k} s(\neg h_i)}, \tag{3}$$

where k represents the number of alternative hypotheses *explicitly* considered. To the degree that k (the number of alternative hypotheses explicitly considered) is less than N (the total number of possible alternative hypotheses), participants' subjective judgments will be larger than the objective probability. Such overestimation of focal judgments will result in judgments that are subadditive; the sum of the probability estimates of each alternative contained in a given sample space would be greater than the probability estimate of the whole sample space. Note that Support Theory, as well as our Eqs. (2) and (3) are descriptions of the judgment process, but are not necessarily normative because one could misperceive the support for the focal and/or for the alternative hypotheses. It is also possible that people differentially weight the focal versus the alternative hypotheses.

The assumption of probability theories that focal hypotheses are compared with alternative hypotheses contrasts with the lack of such an assumption for theories of frequency judgment. Rather than comparing the support for the focal event to the support for the alternatives, frequency judgment processes entail a mapping of the strength of the focal directly onto a judgment scale. For instance, Hintzman's (1988; see also Murdock, Smith & Bai, 2001 and Shiffrin, 2003) Minerva 2 memory model assumes that when making frequency judgments, the features of the focal event are matched against the features of each memory trace stored in long-term memory (LTM). Memory traces are activated to the degree that they are similar to the focal event. The magnitude of the sum of the activation of memory traces is then mapped onto a frequency judgment scale, where the overall activation (or memory strength) is compared to a set of activation criteria to determine between which criteria the memory strength falls. For example, if the strength is greater than the first criterion, but less than the second criterion, a judgment of 1 is rendered, if the strength is greater than the second criterion, but less than the third, a judgment of 2 is rendered, and so forth. In terms of our Eqs. (2) and (3), frequency judgment models would assume that people estimate the numerator, but not the denominator.

The contrast between theories of probability and frequency judgments in terms of the role of the alternatives has important theoretical and empirical consequences. Theoretically, it suggests that processes underlying probability judgments are not isomorphic to those underlying frequency judgments. Empirically, it suggests that the two types of judgments should be affected differently by manipulations directed at changing the input to the focal versus the alternative events and should be differentially related to one's working memory (WM) capacity, which we discuss next.

Working memory limitations

We hypothesized that limitations in WM-capacity constrain the number of alternatives people were able to include in the comparison process when making probability judgments, but are unrelated to frequency judgments which do not entail comparing the focal with alternative hypotheses. Dougherty and Hunter (2003a, 2003b) found that the number of alternative hypotheses people generated correlated positively with their WM capacity. Furthermore, Dougherty and Hunter found that WM correlated negatively with the sum of participants' probability judgments, such that the sum of high spans' judgments were less subadditive than was the sum of low spans' judgments. They posited that WM limits the number of alternative hypotheses one can generate and compare concurrently: the fewer alternative hypotheses used in the comparison process, the greater the judgment. This is because a focal hypothesis appears more likely when one fails to consider all of the alternative hypotheses available.

Effect of strength of alternatives

Windschitl and Wells (1998) found that people's probability judgments were affected by the distribution of the alternative hypotheses. In one experiment, participants were told to imagine they had 21 raffle tickets, and that five other people held either 15, 14, 13, 13, and 12 tickets, respectively (evenly distributed), or that five other people held 52, 6, 5, 2, and 2 tickets, respectively (unevenly distributed). Participants judged the even distribution item as more likely than the uneven distribution item, producing the "alternative outcomes effect." That the distribution of alternative hypotheses affected people's probability judgments supports the hypothesis that people compare the focal to alternative hypotheses when making probability judgments. If they did not compare the focal to alternative hypotheses, changing the distribution of alternative hypotheses would not change the focal probability judgment.

Windschitl and Wells argued that the alternative outcomes effect occurred because when making probability judgments participants used a heuristic in which they compared the focal hypothesis with only the strongest alternatives, rather than considering the entire set of possible alternative hypotheses. When participants compared the focal item (21) with only the most likely alternative, the focal item was compared with the person who held 15 tickets in the even distribution and with the person who held 52 tickets in the uneven distribution. Consequently, although the objective probability is the same in the even and uneven distributions (21/88), participants judged the focal item in the even distribution (21/36) as more likely than the focal item in the uneven distribution (21/73), producing the alternative outcomes effect. In contrast, we propose that the alternative outcomes effects occur because WM-capacity limitations prevent participants from comparing the focal hypothesis with the exhaustive set of alternative hypotheses. When the number of hypotheses in the exhaustive set is greater than the participant's working memory capacity, participants will appear to heuristically consider only the few most likely alternatives even though they are unable to consider the exhaustive set of alternative hypotheses due to working memory limitations. For instance, consider the participant whose working memory capacity allows her to compare the focal hypothesis with four out of the five alternative hypotheses when judging the focal hypothesis in Windschitl and Wells (1998) experiment. If the participant accurately represented the support for each alternative, her even distribution judgment would be

$$\frac{21}{21+15+14+13+13}=\frac{21}{76}$$

and her uneven distribution judgment would be

$$\frac{21}{21+52+6+5+2}=\frac{21}{86}.$$

As a result, the alternative outcomes effect would be found, but it would occur due to working memory limitations rather than to using a heuristic in which the participant compares the focal with only the most likely alternative(s). We argue that people compare as many alternatives as possible, but that WM capacity limits their ability to compare all possible alternatives. The net result of failing to compare all alternatives is that the focal alternative in the even distribution is judged greater than the focal alternative in the uneven distribution.

Absolute vs. relative accuracy

The issue of relative accuracy versus absolute accuracy is important, in as much as the answer to the question of whether probability or frequency judgments are more accurate depends upon how accuracy is defined. If accuracy is defined in terms of the degree to which the judgment is too high, one might be inclined to claim that frequency judgments are altogether more accurate than probability judgments. However, this conclusion may be unwarranted in circumstances in which the decision maker is more concerned with relative accuracy, that is, being able to discriminate which of two events is more likely to occur. Treadwell and Nelson (1996) noted that discrimination ability (a form of relative accuracy), although rarely examined, is as important in assessing the goodness of a judgment as absolute accuracy in that the judgments of a well-discriminated person are predictive; an event judged more likely than another event is in fact more likely to occur. Several measures of discrimination have been proposed, including the slope score (Yates, 1990) and DI' (Wallsten, 1996). In this paper, we focus on gamma correlations between the subjective judgments and the objective probabilities (for a discussion of various measures of ordinal association, including why gamma is the preferred measure in many instances, see Gonzalez & Nelson, 1996).

Previous research comparing the relative accuracy of probability judgments and frequency judgments has provided inconsistent results. Price (1998) examined two measures of relative accuracy, slope (Yates, 1990) and discrimination (Murphy, 1973), and found no difference between probability and frequency judgments for either measure. In contrast, using the gamma correlation statistic as their measure of relative accuracy, Treadwell and Nelson (1996) found that the relative accuracy of probability judgments was significantly higher than that of frequency judgments. In our study, we compare the relative accuracy of probability and frequency judgments. As we will show, whereas frequency judgments might have better absolute accuracy, there is no evidence that they are better in terms of relative accuracy.

Hypotheses and research questions

Our first hypothesis is that probability judgments, but not frequency judgments will be subadditive. We predict that the sum of probability judgments will be subadditive due to limitations in including all alternatives in the comparison process. This prediction is a direct consequence of Eq. (3), which assumes that participants cannot include all alternative hypotheses in the comparison process. In contrast, because frequency judgments do not entail a comparison process, the sum of frequency judgments should not be subadditive. Previous research has found that aggregate frequency judgments tend to be lower and less biased than single item probability judgments in confidence paradigms (Griffin & Tversky, 1992; Mazzoni & Nelson, 1995; Schneider, 1995; Sniezek, Paese, & Switzer, 1990). Thus, the hypothesis that the sum of frequency judgments will be lower and less subadditive than probability judgments is not a new hypothesis, but is in line with our hypothesis that probability and frequency judgments differ due to the comparison component of probability judgment.

One research question is whether frequency and probability judgments differ in terms of relative accuracy. Although we expect differences between frequency and probability judgments in terms of absolute accuracy (subadditivity), we do not expect a difference in relative accuracy, as measured by rank order gamma correlations between participants' subjective judgments and the corresponding objective frequencies. This is because the rank ordering of participants' judgments should be independent of whether they use a comparison or non-comparison process. Judgments of events that occur often should be rated higher than less frequent items regardless of whether a frequency or probability judgment is being made.

We hypothesize that differences in WM capacity will be negatively correlated with the sum of probability judgments but unrelated to the sum of frequency judgments. People with larger working memory capacities can consider more alternative hypotheses. As shown in past research (e.g., Dougherty, Gettys, & Thomas, 1997; Dougherty & Hunter, 2003a), considering more alternative hypotheses leads people to make lower probability judgments. This has been hypothesized to be due to the fact that probability judgments entice one to engage in a comparison process. In contrast, we hypothesize that frequency judgments will be unrelated to WM-capacity because one need not concurrently consider multiple alternatives for comparison when judging frequency. One consequence of the comparison process is that participants should be affected by the distribution of the alternatives. If this is the case, then probability judgments, but not frequency judgments, should be affected by the distribution of the alternatives. Moreover, our Eq. (3) suggests that the magnitude of the alternative-outcomes effect for probability judgments is due to the number of alternatives explicitly considered. Thus, we expect WM-span be negatively related to the magnitude of the alternative-outcomes effect.

As a final hypothesis, we expect that the reaction time (RT) for making probability judgments will be longer than the RT for making frequency judgments. This is because the comparison process requires participants to estimate both the focal hypothesis and the alternative hypotheses and make a comparative judgment. In contrast, the frequency judgment task requires only that participants estimate the focal hypothesis. This should lead to lower RTs when making frequency judgments.

Method

Participants

One-hundred fifty-seven University of Maryland introductory psychology students participated in the experiment for course credit.

Materials

Ten computer-presented photographs were used to represent the events, and two computer-presented photographs were used to represent the type of event. A postal package picture represented the beginning of each event. Five of the labeled images represented tool events: screwdriver, utility knife, wrench, pliers, tape measure, and five labeled images represented footwear events: canvas shoes, hiking shoes, dress shoes, running shoes, and brown shoes. A picture of a shoebox represented the footwear set and a picture of a toolbox represented the tool set. The stimuli used were identical to those used in the Windschitl et al. (2002) study.

Design and procedure

A 2 (distribution type) × 2 (judgment type) mixed factorial design with WM as a continuous subject variable was employed. Distribution type (even distribution vs. uneven distribution) was manipulated within participants, and judgment type (frequency vs. probability) was manipulated between participants. Even distribution refers to five items that were distributed 30-15-15-15-15 in the presentation phase of the experiment. The uneven distribution refers to the five items that were distributed 30-45-5-5-5 in the presentation phase of the experiment. Four dependent variables were considered important to the hypotheses of this experiment: the sums of participants' judgments which was computed by summing the five judgments within each distribution; the rank order gamma correlations between participants' subjective judgments and the objective frequency that items occurred; judgments of each of the two frequency-30 items, and judgment reaction times. The frequency-30 items are the items that occurred 30 times in each distribution and are the items upon which the alternative outcomes effect predictions are based. The sum of judgments and frequency-30 item judgments measured the absolute accuracy of participants' judgments (i.e., the magnitude of the judgments), whereas the rank order gamma correlations measured the relative accuracy of participants' judgments (i.e., the degree to which participants were able to assign higher judgments to items that occurred more frequently relative to items that occurred less frequently). The learning task design and procedure were developed by Windschitl et al. (2002).

Participants were randomly assigned to the frequency judgment condition ($n = 80$) and the probability judgment condition ($n = 77$). Each participant completed the experiment individually in sessions that lasted approximately 45 min. The experiment consisted of three tasks, the learning task, the judgment task, and the operation-span (o-span) task (a measure of WM). In the learning phase, participants were instructed to imagine that they had two strange and wealthy friends who liked to send them gifts. One friend always sent gifts of tools, whereas the other friend always sent gifts of footwear. Participants were instructed to imagine that after opening the gifts, they threw them onto separate piles (tools or footwear) in their garage. Participants first saw a package representing the type of present they were receiving (toolbox or shoebox) for 1 s, and then saw a picture representing the specific present they had received. To move to the next present, participants pressed the key corresponding to the first letter of the package they had just received. Each participant saw 180 present pictures consisting of 90 shoes pictures and 90 tool pictures, in a random order. The frequency-30 item for each set was always presented 30 times. Category and distribution type were counterbalanced across participants. Additionally, each tool and shoe item was the frequency-30 item equally often across participants and across the two distribution types. After the learning phase, participants engaged in a 3-min filler task which consisted of answering general knowledge questions. The filler task was used to reduce recency bias.

For the judgment task, participants saw a screen, which reminded them that they now had two piles of presents in their garage, one of tools, and one of shoes. Participants in the frequency group were asked "How many packages of (item) did you receive?" Participants in the probability group were instructed to imagine that they had a particular need for one of the presents, and to imagine that they would go to the pile to randomly select a package. Participants were asked, "If you go to the tool (footwear) pile in your garage, what is the chance that when looking for (item) you would happen to pick a box with (item) in it on the first try?" Participants were told to respond by using the numeric keypad on the computer keyboard, and were given a blank text box in which to respond. Participants were instructed to respond with a number between 0% and 100% where 0% meant there was no chance at all that the outcome in question could occur and 100% meant that the outcome was certain to occur. They were further informed that a judgment of 50% was equivalent to the likelihood that a coin flip would land on heads instead of tails. Participants made judgments of each of the five items in each of the two sets (shoes and tools). Participants always judged the frequency-30 items first, and thereafter judged all other items in a random order.

After completing the experimental task, each participant completed the o-span task as a measure of WM-span (Turner & Engle, 1989). The o-span task required that participants retain a growing list of words while solving mathematical problems. For example, on successive presentations participants would be shown: $(4 \times 3) - 3 = 9$? Door; $(4/2) + 3 = 7$? Shoe. Participants were required to read the equation aloud, verify whether or not the equation was true, and then read the word aloud. After saying the word, the experimenter advanced to the next operation-word pair. This continued until the participant was prompted to recall the words from that set in the order in which they were presented. Participants were presented with 15 sets of equation-word pairs with set sizes ranging from 2 to 6. Each set size occurred three times randomly throughout the task. Performance on the o-span task was measured by the number of words recalled in the correct serial position for which the corresponding math problem was correctly verified. The maximum possible score was 60, with higher scores representing larger WM capacity. A detailed description of the operation-span task is presented in Turner and Engle (1989).

Results

The data from 2 participants were eliminated for responding with the same probability for all ten judgments (both participants gave judgments of 20 for all 10 questions). An additional three participants, whose judgment sums were more than three standard deviations above the group mean were identified as outliers and were excluded from the statistical analyses. All three of these participants were in the frequency condition. The data from these three participants are presented in Appendix A. Frequency judgments were scaled to the range of 0–100, to make them comparable to the probability judgments, by dividing each frequency judgment by 90 (the total number of items presented within each distribution) and multiplying by 100. Data were analyzed using the multiple regression approach to model comparisons, with our measure of working-memory capacity entered as a quantitative predictor and the two independent variables (distribution and judgment type) dummy coded and used as predictor variables. This approach enables a test of the interaction between the quantitative variable and the independent variable without concern for heterogeneity of regression slopes.

Absolute accuracy

The top graph of Fig. 1 plots the sum of each participant's frequency judgments as a function of WM span and the bottom graph of Fig. 1 plots the sum of each participant's probability judgments as a function of WM span. Even distribution judgments are represented by diamonds and uneven distribution judgments are represented by squares. Fig. 2 plots each participant's frequency-30 judgments as a function of WM span. Table 1 shows the mean sum of judgments and the mean frequency-30 judgments for each of the distributions. As is evident in these figures and in Table 1, the sum of participants' judgments was more subadditive and frequency-30 judgments were higher when making probability judgments than when making frequency judgments. There was a main effect of judgment type on the sum of judgments, $F(1,296) = 84.62$, $p < .05$, and on the magnitude of frequency-30 judgments, $F(1,292) = 68.39$, $p < .05$ such that probability judgments were larger than frequency judgments. As seen in Fig. 2, we found an alternative outcomes effect for probability but not for frequency judgments. A marginally significant judgment type by distribution type interaction was found for frequency-30 judgments, $F(1,296) = 2.66$, $p = .104$. Planned comparisons revealed that for the probability condition frequency-30

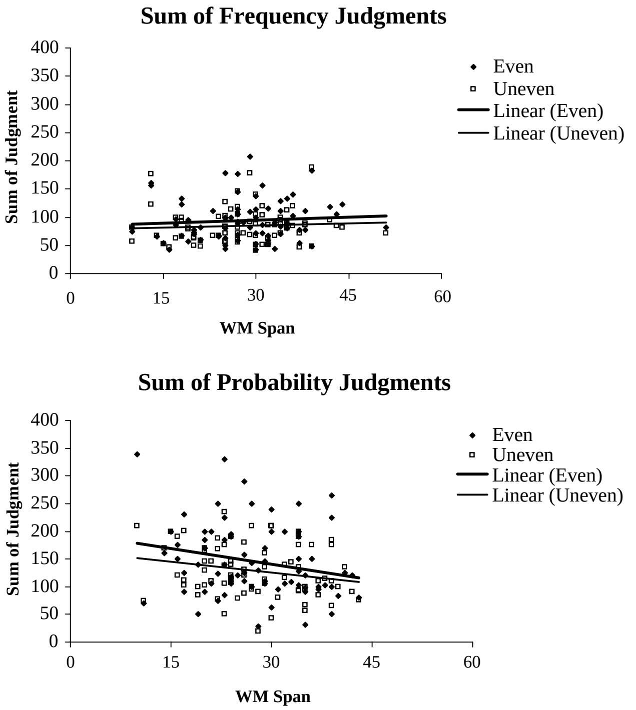

Fig. 1. Sum of judgments as a function of judgment type (probability vs. frequency), distribution type (even vs. uneven), and WM span.

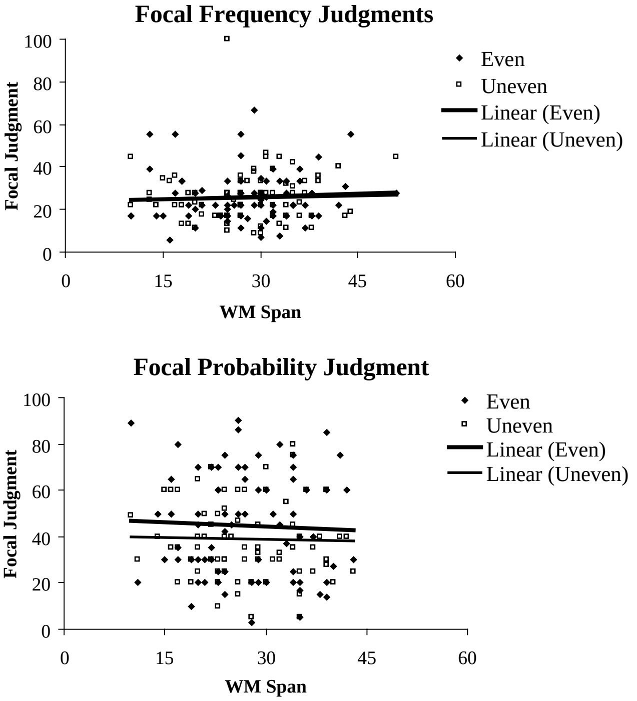

Fig. 2. Frequency-30 judgments as a function of judgment type (probability vs. frequency), distribution type (even vs. uneven), and WM span.

Table 1
Mean judgments as a function of judgment type and distribution type

Distribution type	Mean frequency estimate	Mean probability estimate
Frequency of presentation		
30-15-15-15-15 even distribution		
30 (Focal)	25.63 (1.37)	44.73 (2.64)
15	17.59 (0.95)	26.85 (2.04)
15	17.26 (1.02)	26.21 (1.78)
15	16.51 (0.88)	24.99 (1.76)
15	16.81 (0.88)	23.16 (1.66)
Sum	93.80 (4.11)	145.95 (7.56)
30-45-5-5-5 uneven distribution		
30 (focal)	26.15 (1.47)	38.96 (1.98)
45	36.15 (1.62)	55.96 (2.44)
5	7.32 (0.58)	10.17 (0.90)
5	7.40 (0.70)	11.79 (1.19)
5	7.13 (0.44)	12.23 (1.25)
Sum	84.14 (3.48)	129.11 (5.45)

Judgments are percentages. Standard errors are presented in parentheses after the mean judgments.

judgments were larger in the even distribution than in the uneven distribution, $t(74) = 2.42$, $p < .05$, but for the frequency condition the frequency-30 judgments in the even distribution did not differ from frequency-30 judgments in the uneven distribution, $p > .05$. Thus, consistent with the idea that probability, but not frequency judgments entail a comparison process, only probability judgments were affected by distribution of the alternative events.[2] No interaction between judgment type and distribution type was found for the sum of judgments, $p > .05$.[3]

The regression slopes plotted in Fig. 1 show that sums of probability judgments for high spans are lower than for low spans. Statistical tests supported this hypothesis; WM-span interacted significantly with

[2] We also computed the ratio of the even distribution focal judgment to the uneven distribution focal judgment for each participant. The degree to which the mean ratio of even distribution focal judgment to uneven distribution focal judgment is greater than 1 represents the degree to which the alternative outcomes effect occurred. Consistent with our analyses of the mean focal judgments, for probability judgments the ratio ($M = 1.26$, $SE = .09$) differed significantly from 1, $t(75) = 2.71$, $p < .05$ but for frequency judgments the ratio ($M = 1.07$, $SE = .06$) did not differ significantly from 1, $t(77) = 1.21$, $p > .05$.

[3] Planned comparisons revealed that the sum of judgments was significantly lower in the uneven distribution compared to the even distribution for the probability condition, $t(296) = 2.22$, $p < .05$, with no difference between the distributions for the frequency judgment condition, $p = .19$. However, because there was no judgment type × distribution interaction, $F(1,296) = 0.53$, $p > .05$, we are hesitant to over-interpret the effect of distribution on the sum of judgments in the probability condition. When analyzing the ratio of the even distribution judgment sum to the uneven distribution judgment sum, for probability judgments the ratio ($M = 1.18$, $SE = .08$) differed significantly from 1, $t(75) = 2.25$, $p < .05$ however in contrast with our predictions, for frequency judgments the ratio ($M = 1.13$, $SE = .03$) also differed significantly from 1, $t(77) = 4.81$, $p < .05$. However, the focal judgments actually provide a much cleaner test of the effect of distribution on judgment.

Table 2
Correlation between WM-span and judgments

Distribution type	Frequency	Probability
Frequency of Presentation		
30-15-15-15-15 even distribution		
30 (Focal)	0.05	−0.04
15	0.05	−0.20
15	0.15	−0.25*
15	0.03	−0.25*
15	0.07	−0.20
Sum	0.09	−0.23*
30-45-5-5-5 uneven distribution		
30 (focal)	0.16	−0.02
45	0.10	−0.12
5	−0.16	−0.16
5	−0.17	−0.36*
5	0.13	−0.23*
Sum	0.07	−0.21*

* $p < 0.05$.

judgment type, $F(1,296) = 8.38$, $p < .05$. More specifically, as predicted there was no relationship between WM-span and the sum of participants' frequency judgments, but there was a negative relationship between WM-span and the sum of participants' probability judgments. Interestingly, there was no such interaction for the frequency-30 judgments. Table 2 presents the correlations between WM-span and sums of judgments and individual event judgments for the even and uneven distributions. As can be seen, probability judgments show a negative relationship with WM-span in 12 out of the 12 comparisons (with six correlations being significantly greater than zero at the $p = .05$ level) whereas frequency judgments show negative correlations in only 2 out of the 12 comparisons (with none of the correlations significantly different from zero). Although the correlations between WM-span and judgment are quite modest, that they obtain only in the probability judgment condition is consistent with our theory that probability judgments, but not frequency judgments, entail comparing the focal hypothesis with alternative hypotheses and that this comparison process draws on WM resources.

Relative accuracy

For each participant, we computed two gamma correlations, one for each distribution, between the between the objective frequencies and the participants' judgments. Statistical analyses were conducted on the gammas using the multiple-regression approach to model comparisons.

We predicted and found that gamma correlations between objective frequency and participants' probability judgments ($M = 0.78$, $SD = .53$) did not differ significantly from gamma correlations for frequency judgments ($M = 0.80$, $SD = 0.50$), $F(1,297) = 0.08$, $p > .05$. This finding suggests that while frequency judgments might be better than probability judgments in an absolute sense, they are not better in a relative sense: that is, frequency judgments did not have better relative accuracy than probability judgments.[4]

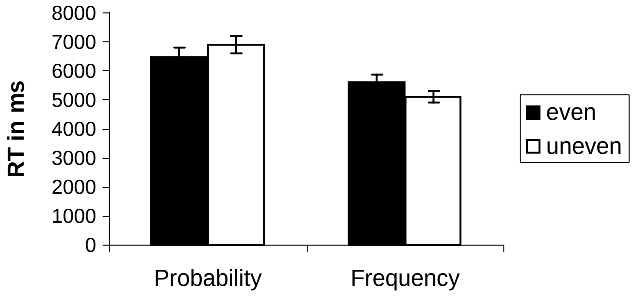

Fig. 3. Average judgment reaction time in milliseconds for Frequency and Probability Judgments.

Reaction times

Fig. 3 presents the average reaction time when making probability or frequency judgments in the even versus in the uneven distributions. The average reaction time when making frequency judgments was significantly shorter than the average reaction times when making probability judgments, $F(1,300) = 24.22$, $p < .05$. No main effect of distribution was found for average judgment reaction time, nor was a significant interaction between distribution type and judgment type found for average judgment reaction time.

Discussion

The primary purpose of our study was to test whether one of the main assumptions of theories of probability judgment, that probability judgments are made using a comparison process, extends to frequency judgments. As implied by theories of probability and frequency judgment, we argued that only probability judgments

[4] Although there was a main effect of distribution on gamma, $F(1,297) = 31.09$, $p < .05$, this is likely an artifact of the gamma correlation statistic. One would expect higher correlations between subjective judgments and the objective frequencies in the uneven distribution merely because of the differences between the objective frequencies in the uneven distribution compared to the even distribution. Moreover, gamma excludes ties, and because the two distributions have a different number of ties within the objective frequencies, the two gammas are differentially affected by ties in the objective frequencies. Thus, whereas it is possible to compare across judgment type within a distribution, comparisons of gamma between distributions are not meaningfully interpreted. We should also note also that there was no interaction between distribution type and judgment type on mean gamma correlations, $F(1,297) = 0.14$, $p > .05$.

entailed the use of a comparison process to derive the judgment. Several aspects of our results were consistent with this hypothesis. First, only probability judgments were sensitive to the distribution of the alternative hypotheses. Second, probability judgments, but not frequency judgments, were related to individual differences in working-memory capacity. Finally, the RT for making probability judgments was significantly longer than the RT for making frequency judgments. Taken together, these results support the hypothesis that participants in the probability judgment condition, but not participants in the frequency condition, compared the focal hypothesis with alternative hypotheses when making their judgments. From a theoretical viewpoint, our study indicates that the processes underlying probability judgments are not isomorphic to those underlying frequency judgments. An empirical implication of our study is that factors that affect the comparison process will affect probability judgments but not frequency judgments. Thus, the two types of judgments are affected differently by manipulations to the focal versus to the alternative events, and are differentially related to WM capacity.

A second purpose of our study was to examine the accuracy of frequency and probability judgments. Although we found differences in absolute accuracy between frequency and probability judgments, there were no differences in relative accuracy between the two judgment conditions. This is an important finding because it suggests that conclusions that frequency judgments are more accurate than probability judgments may well be limited to one definition of accuracy, absolute accuracy. Only a few studies have reported comparisons in relative accuracy between frequency and probability judgments (Price, 1998; Treadwell & Nelson, 1996). Our results are consistent with those of Price (1998); although frequency judgments lead to lower judgments than probability judgments, they do not lead to better relative accuracy.

One issue of importance is the degree to which regression effects affected the results. Erev, Wallsten, and Budescu (1994) argued that people's tendency to be over-confident could be attributed, at least in part, to regression towards the mean. Accordingly, judgments for highly probable events should be underestimated and judgments for highly improbable events should be overestimated. Within the context of our experiment, there are two components on which regression may operate when making judgments: estimation of the to-be-judged item and estimation of the alternatives. Similar to previous studies (Fiedler & Armbruster, 1994; Hintzman, 1988; Windschitl et al., 2002), we found that frequency judgments were regressive: high frequency items (e.g., items presented 30 or 45 times) were underestimated, low frequency items (e.g., items presented five times) were overestimated, and medium frequency items (e.g., items presented 15 times) were estimated almost accurately (see Table 1). If, as we theorize, participants assess the subjective frequency of alternatives to make probability judgments, regression effects for assessments of subjective frequency could influence the frequency-30 probability judgment in the opposite direction than that predicted by the alternative outcome effect, as Windschitl et al. (2002) noted. In the uneven (30-45-5-5-5) distribution, because the amount of regression decreasing the assessment of the highest alternative was less than the amount of regression increasing the assessment of the lowest frequency alternatives, the sum of the subjective frequencies for all non-frequency-30 outcomes (58) underestimated the objective value (60). Underestimating the frequency of the alternatives could in turn inflate the frequency-30 probability judgment. In fact, we found that the sum of the non-frequency-30 alternatives in the uneven distribution (58) was less than the sum for the even (30-15-15-15-15) distribution (68.17). Thus regression effects in assessing the frequency of alternatives could have caused the frequency-30 alternative in the uneven distribution to seem more likely than in the even distribution, as the assessment of the alternatives is lower in the uneven distribution than in the even distribution. Thus, any effect of regression would be to enhance the perceived probability of the frequency-30 for the uneven distribution relative to the frequency-30 for the even distribution—an effect that counters our hypothesis and the results of our study. Consequently, if regression effects played a role in our experiment, the net result would have been to reduce the statistical power of our statistical tests by reducing the magnitude of the alternative outcomes effect.

In sum, our results support the hypothesis that the main difference between frequency and probability judgments is that probability judgments, but not frequency judgments, require a comparison between the focal event and the alternative events. The findings that probability judgments, but not frequency judgments, were related to measures of WM-capacity and that probability judgments but not frequency judgments, were sensitive to the distribution of the alternatives suggest that a comparison process underlies probability but not frequency judgments. Finally, our results suggest that although frequency judgments are more accurate than probability judgments in terms of absolute accuracy, frequency judgments are equally accurate to probability judgments in terms of relative accuracy.

Appendix A

Data from the three participants identified as outliers

Participant No.	WM span score	Frequency-30 judgments		Sum of judgments	
		Even distribution	Uneven distribution	Even distribution	Uneven distribution
3052	31	60	60	210	230
3112	25	50	50	198	207
3125	16	35	45	218	178
Mean		48.33	51.67	208.67	205.00

Note. Participants were excluded from analysis if the sum of judgments was more than three standard deviations above the condition means.

References

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

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). Minerva-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*, 180–209.

Dougherty, M. R. P., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes, 70*, 135–148.

Dougherty, M. R. P., & Hunter, J. E. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica, 31*, 263–282.

Dougherty, M. R. P., & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31*, 968–982.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*, 519–527.

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

Fiedler, K., & Armbruster, T. (1994). Two halfs may be more than one whole: Category-split effects on frequency illusions. *Journal of Personality and Social Psychology, 66*, 633–645.

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

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119*, 159–165.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*, 411–435.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multi-trace memory model. *Psychological Review, 95*, 528–551.

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

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that are not attributable to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1263–1274.

Murdock, B., Smith, D., & Bai, J. (2001). Judgments of frequency and recency in a distributed memory model. *Journal of Mathematical Psychology, 45*, 564–602.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meterology, 12*, 595–600.

Price, P. C. (1998). Effects of a relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. *Organizational Behavior and Human Decision Processes, 76*, 277–297.

Schneider, S. L. (1995). Item difficulty, discrimination and the confidence-frequency effect in a categorical judgment task. *Organizational Behavior and Human Decision Processes, 61*, 148–167.

Shiffrin, R. (2003). Modeling memory and perception. *Cognitive Science, 27*, 341–378.

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

Sniezek, J. A., Paese, P. W., & Switzer, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes, 46*, 264–282.

Treadwell, J. R., & Nelson, T. O. (1996). Availability of information and the aggregation of confidence in prior decisions. *Organizational Behavior and Human Decision Processes, 68*, 13–27.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*, 127–154.

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

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547–567.

Wallsten, T. O. (1996). An analysis of judgment research analyses. *Organizational Behavior and Human Decision Processes, 65*, 220–226.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology, 75*(6), 1411–1423.

Windschitl, P. D., Young, M. E., & Jenson, M. E. (2002). Likelihood judgment based on previously observed outcomes: The alternative-outcomes effect in a learning paradigm. *Memory & Cognition, 30*(3), 469–477.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall.