Journal of Experimental Psychology:
Learning, Memory, and Cognition
1998, Vol. 24, No. 3, 754–770

0278-7393/98/$3.00

Are Judgments of the Positional Frequencies of Letters Systematically Biased Due to Availability?

Peter Sedlmeier
University of Paderborn

Ralph Hertwig and Gerd Gigerenzer
Max Planck Institute for Psychological Research

How do people estimate whether a particular letter is more frequent in the 1st versus in a later position? The authors tested 2 precise versions of the availability hypothesis, a hypothesis that assumes that frequency processing occurs on the level of the phonological classes of vowels and consonants, and the regressed-frequencies hypothesis, which assumes monitoring of individual letters. Across 3 studies, it was found that (a) judgments of whether a letter is more frequent in the 1st or the 2nd position generally followed the actual proportions and (b) the estimated relative frequencies in the 1st versus the 2nd position closely agreed with the actual rank ordering, except for an overestimation of low and underestimation of high values. These results favor the regressed-frequencies hypothesis and challenge the conclusions about frequency judgments in the heuristics and biases literature.

How do humans estimate whether a particular letter is more frequent in the first versus in a later position? Tversky and Kahneman (1973) proposed the availability heuristic as an explanation for how these estimates are generated and for why their participants' estimates were systematically biased. Both their finding and their explanation have since been taken for granted and are cited in almost every textbook on cognitive psychology. Surprisingly, there seems to be no single published replication, except for a one-page article (White, 1991). The lesson from the Tversky and Kahneman study is often taken to be something like the following: "Apparently, people make generalizations about frequency based on the *availability* of the instances in memory rather than by an accurate count of actual past experience" (Mayer, 1992, p. 109).

Tversky and Kahneman's (1973) result is puzzling when compared to the findings reported in many studies on humans' competence in encoding frequency of occurrences. A large amount of literature suggests that memory is often (but not always) excellent in storing frequency information from various environments. In particular, several authors have documented that participants' judgments of the frequency of letters and words generally show a remarkable sensitivity to the actual frequencies (e.g., Hock, Malcus, & Hasher, 1986; Johnson, Peterson, Yap, & Rose, 1989; Naveh-Benjamin & Jonides, 1986; Watkins & LeCompte, 1991). The registering of event occurrences for frequency judgments is assumed to be a fairly automatic process; that is, it requires little to no attentional capacity (e.g., Hasher & Zacks, 1979, 1984). In this view, frequency is one of the few attributes (besides spatial location, temporal information, and word meaning) that seems to be encoded automatically. Although the claim of automatic encoding may be too strong and has been criticized (see Barsalou, 1992, chapter 4), there seems to be broad agreement with what Jonides and Jones (1992) described as follows: "Ask about the relative numbers of many kinds of events, and you are likely to get answers that reflect the actual relative frequencies of the events with great fidelity" (p. 368).

If Jonides and Jones (1992) are correct in concluding that humans are able to derive answers that reflect the actual relative frequencies of the events with great fidelity, then one would not expect the estimates of relative letter frequencies to be systematically biased, contrary to Tversky and Kahneman's (1973) findings. In this article, we examine this contradiction.

Four Hypotheses of Letter-Frequency Judgments

In the original study, participants had to estimate the relative frequencies of letters in different positions in English words (Tversky & Kahneman, 1973, pp. 211–212):

> The frequency of appearance of letters in the English language

Peter Sedlmeier, Fachbereich 2–Psychology, University of Paderborn, Paderborn, Germany; Ralph Hertwig and Gerd Gigerenzer, Center for Adaptive Behavior and Cognition, Max Planck Institute for Psychological Research, Munich, Germany. Ralph Hertwig and Gerd Gigerenzer are now at the Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany.

This research was supported by a Feodor Lynen stipend of the Alexander von Humboldt Foundation as well as by a Habilitationsstipendium of the Deutsche Forschungsgemeinschaft. We thank Sylvia Dickgießer of the Institut für deutsche Sprache in Mannheim, Klaudia Obermayr, and Nicola Korherr for their valuable assistance, and Peter White for providing us with his data. Our thanks also go to Larry Barsalou, Valerie M. Chase, Berna Eden, Dan Goldstein, Wolfgang Hell, Johannes Hoenekopp, Peter Juslin, Howard Nusbaum, Terry Regier, Frank Renkewitz, Anita Todd, Tom Trabasso, and Manfred Wettler for many constructive comments.

Correspondence concerning this article should be addressed to Peter Sedlmeier, University of Paderborn, Fachbereich 2–Psychology, 33095 Paderborn, Germany. Electronic mail may be sent to sedl@psycho.uni-paderborn.de.

was studied. A typical text was selected, and the relative frequency with which various letters of the alphabet appeared in the first and third positions in words was recorded. Words of less than three letters were excluded from the count.

You will be given several letters of the alphabet, and you will be asked to judge whether these letters appear more often in the first or in the third position, and to estimate the ratio of the frequency with which they appear in these positions.

Consider the letter *R*.

Is *R* more likely to appear in _____ the first position?
_____ the third position?
(check one)

My estimate for the ratio of these values is _____: 1.

Tversky and Kahneman (1973) used five consonants (*K*, *L*, *N*, *R*, and *V*), each of which was actually more frequent in the third position. Of 152 participants, 105 judged the first position to be more likely for a majority of the letters and 47 judged the third position to be more likely for a majority of the letters. Each of the five letters was judged by a majority of the participants to be more frequent in the first than in the third position; the median estimated proportion in the first position was 67% (ratio of 2:1) for each of the five letters.

Availability Heuristic: Number and Speed

Several researchers have pointed out that the notion of the availability heuristic has been only vaguely sketched and is consistent with several different mechanisms (Fiedler, 1983; Lopes & Oden, 1991; Schwarz et al., 1991; Wänke, Schwarz, & Bless, 1995). The heuristic applied to judgments of relative letter frequencies is a case in point. Tversky and Kahneman (1973) assumed that the relative frequency of a letter, such as *R*, in the first and third positions is estimated by the relative "ease" with which words that start with *R* and words in which *R* is in the third position come to mind. Although they did not use an independent measure of ease in their study, two different ways of how one could measure it were indicated. On the one hand, Tversky and Kahneman (1973) proposed that ease can be measured by actual recall: "the availability of instances could be measured by the total number of instances retrieved or constructed in any given problem" (p. 210). Measuring ease by the actual frequency of instances recalled has become widespread. This practice has received various labels such as the exemplar-retrieval hypothesis (Greene, 1989) and the recall-estimate theory (Watkins & LeCompte, 1991). We term this version of the availability heuristic where we operationalize ease by the number of words recalled *availability-by-number.* On the other hand, Tversky and Kahneman (1973) suggested that

> it is not necessary to perform the actual operations of retrieval or construction. It suffices to assess the ease with which these operations could be performed, much as the difficulty of a puzzle or mathematical problem can be assessed without considering specific solutions. (p. 208)

We term another version of the availability heuristic where we operationalize ease by speed of retrieval *availability-by-speed.*

In particular, we assume the following processes: Availability-by-number states that if asked for the proportion with which a certain letter occurs in the first versus in a later position in words, one produces words with this letter in the respective positions and uses the produced proportion as an estimate for the actual proportion. Availability-by-speed states that if asked for the proportion with which a certain letter occurs in the first versus in a later position, one produces single words with this letter in each position and uses the time ratio (of the consumed retrieval times) as an estimate of the actual proportion. Here the speed of the retrieval of a single word is taken as an indicator for the ease with which more retrieval processes could be performed, whereas availability-by-number defines the actual number of retrieved words for a constant time period as an indicator. If it were true that words with a particular letter in the first position can be produced more easily, as Tversky and Kahneman (1973, p. 211) suggested, both hypotheses predict that all letters are generally judged to be more frequent in the first position, whether or not they actually are.

Letter-Class Hypothesis

Lopes and Oden (1991) pointed out that in Tversky and Kahneman's (1973) study, each of the five consonants used (*K*, *L*, *N*, *R*, and *V*) was more frequent in the third position in the English language. Therefore, they argued that this sample was atypical, because most consonants (12 of 20) are in fact more frequent in the first position. If participants knew that most, but not which, consonants are more frequent in the first position, they might reasonably estimate that the majority of consonants were more frequent in the first position—in ignorance of having been presented a nonrepresentative sample. Consistent with this conjecture, the actual proportion of consonants that are more frequent in the first position (12 of 20, i.e., 60%) gives a good first approximation of the 67% median estimate given by Tversky and Kahneman's (1973) participants.

We used Lopes and Oden's (1991) critique as a starting point to design an alternative hypothesis of how letter frequencies are judged. What we call the *letter-class hypothesis* assumes that, in the absence of knowledge about a particular letter, relative frequencies are inferred from knowledge about the phonological class they belong to: consonants or vowels. This is a testable hypothesis because the proportion of letters at the first (as compared to the first and third) position in English is quite different for vowels (about 30% in the first position) than for consonants (60%). Thus, the letter-class hypothesis assumes that (a) under ignorance of the proportion for an individual letter, a class containing these instances is used as a basis for inference; (b) the class is the phonological class of vowels or consonants; and (c) the proportion in the respective class is reported as the default value. If presented a vowel, people will report the proportion of vowels by default and if presented a consonant, they will report the proportion of consonants.

Regressed-Frequencies Hypothesis

On the basis of a large body of results, Jonides and Jones (1992) suggested that humans are able to judge the relative numbers of many kinds of events in a way that reflects the

actual relative frequencies of the events with great fidelity. There is, however, a well-known phenomenon usually encountered in frequency judgment tasks that does affect the fidelity of the absolute size of frequency judgments but not their rank order: Low frequencies tend to be overestimated and high frequencies underestimated (e.g., Greene, 1984; Hintzman, 1969; Shanks, 1995; Varey, Mellers, & Birnbaum, 1990). The most comprehensive source available for estimating the amount of this "regression effect" in the case of letters seems to be the study by Attneave (1953), in which participants judged the relative frequencies of all letters in the alphabet. He found a correlation of .79 between actual relative frequencies and the medians of the judged frequencies. An analysis of his data shows that the median judgments were regressed toward the mean by about 70%.[1]

We used Attneave's (1953) result to design a simple *regressed-frequencies hypothesis* that assumes that (a) the frequencies with which individual letters occur at different positions in words are monitored (e.g., during reading), and (b) the letter frequencies represented in the mind are regressed toward the mean. Thus, when asked for the relative frequency of a particular letter, people will give judgments of relative letter frequencies that reflect the actual ones, although they will overestimate relative frequencies below the mean and underestimate those above the mean.

Predictions

From the four hypotheses we now derive predictions for individual letters. We use a larger number of consonants and vowels than were used previously, to take into account the critique by Lopes and Oden (1991). Tversky and Kahneman (1973) let their participants make judgments about letters in the first and third positions of words. If one tests a large number of letters, this could lead to a problem. It turns out that the Mayzner and Tresselt (1965) corpus used by Tversky and Kahneman (1973) can be a biased standard against which to compare participants' estimates. The reason is that this corpus covers only letters in words with at least three letters' length, whereas English also has one- and two-letter words, some of which are quite frequent. Thus, the actual proportion of many letters in first position in English is larger than the proportions reported in the corpus. If participants judge the actual proportion of letters in various positions in all English words, then they would be correct in giving higher estimates for the first position than for the figures in the corpus. This may pose little problem for the five letters used in the original study by Tversky and Kahneman (1973), because there exist few one- or two-letter words that include these five letters (but consider, for instance, words such as *no*, or abbreviations such as *lb*, and *km*). The authors also carefully pointed out to their participants that they should not consider words with fewer than three letters. However, if one uses a large number of both consonants and vowels, the comparison between the proportions reported in the corpus and those estimated by participants becomes problematic unless participants succeed in discarding everything they know about one- and two-letter words. To avoid this problem, we relied on a corpus that also includes two-letter words and asked our German participants about the proportion of letters in the first versus second position. Because no one-letter words (such as the English *a* or *I*) exist in German, the total number of letters that occur in the first and those that occur in the second (but not the third) position is the same. We compared participants' estimates with the actual proportions in German, as reported in the "Mannheimer Korpus" (Institut für deutsche Sprache, 1968, 1969). The thick line in Figure 1 shows the actual proportions (expressed as percentages) with which the letters used in our experiments appear in the first position (as compared to the first and second positions), rank ordered from left to right, from *C* (10.3% in the first position) through *G* (98.9% in the first position). Consider the letter *C* as an example. *C* occurred 5,870 times in the first position and 51,034 times in the second position in the German corpus. Considering only words with *C* in the first or second position, *C* occurred first 10.3% of the time and second 89.7% of the time. What are the predictions of the four hypotheses?

Predictions of the Availability Hypotheses

To specify the predictions of the two versions of the availability heuristic, we conducted two studies to obtain numerical values for availability-by-number and availability-by-speed.

Availability-by-number. At the University of Paderborn, Germany, 131 participants in an introductory psychology course completed a production task. They were given a booklet, were admonished not to read ahead, and received the following oral instructions (translated from German):

> Your task is to recall as many words as you can in a certain time. At the top of the following page you will see a letter. Write down as many words as possible that have this letter as the first (second) letter.

Each participant completed four production tasks, that is, worked on two letters and produced words with these letters in the first and second positions, respectively. The letters used and the order of presentation (productions of words with a letter in the first vs. second position) was randomly varied across participants. The number of participants working on each letter ranged from 17 to 21. The letter was always printed at the top of a page, followed by 20 lines

[1] The analysis was as follows: First, the actual frequencies and the median judgments both were transformed to percentages (sum of actual frequencies for all 26 letters = 100%; sum of median judgments for all 26 letters = 100%). This leads to an identical mean, 3.85% (100% divided by 26 [letters]) for both actual frequencies and median judgments. The distances (in percentage points) of the actual frequencies from the mean (AD) were then calculated. And finally, the distances of the judgments from the mean (JD) were expressed relative to the former distances: Amount of regression = 100 − (JD/AD)100. The mean amount of regression over all 26 letters was 69.8% (median regression: 70.6%). The amount of regression was not dependent on the size of the distance from the mean (r = .00002 between actual values and amount of regression).

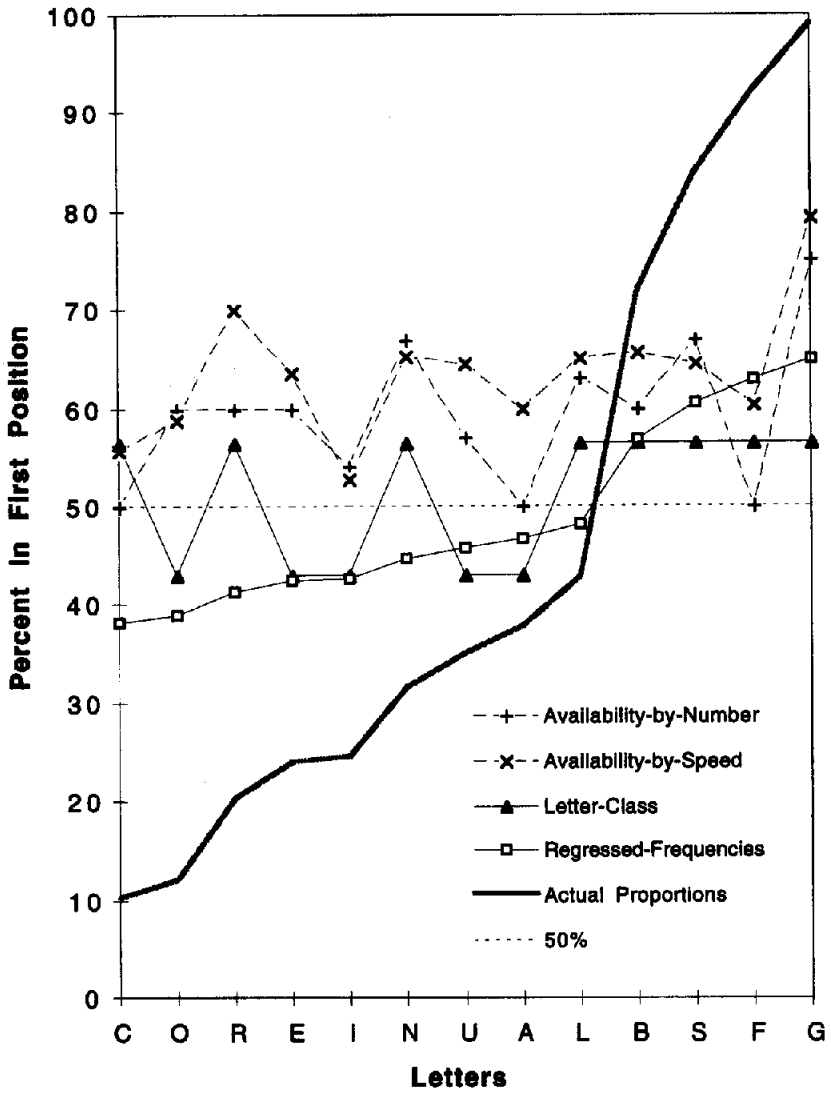

Figure 1. Predictions of the two availability hypotheses (availability-by-number and availability-by-speed), the letter-class hypothesis, and the regressed-frequencies hypothesis. See text for further explanation.

divided into two columns. Participants were given 10 s to produce as many words as possible.

The results of the production task allow us to specify the predictions of availability-by-number for individual letters. It also provides a test for the intuitively appealing assumption that it should be easier to recall words with a certain letter in the first versus a later position (Tversky & Kahneman, 1973, p. 211). To the best of our knowledge, this assumption has never been empirically tested. The production task provides such a test for the first and second positions. Proportions were calculated for each participant by dividing the number of words with a certain letter in the first position recalled by the number of all words recalled (first and second positions). Availability-by-number assumes that the relative frequency estimate is based on the number of words recalled, and its predictions can be stated as follows:

$$\text{Estimate}_{\text{Letter in pos 1}} = \Sigma\text{words}_{\text{Letter in pos 1}}/(\Sigma\text{words}_{\text{Letter in pos 1}} + \Sigma\text{words}_{\text{Letter in pos 2}}),$$

where $\text{Estimate}_{\text{Letter in pos 1}}$ is the relative frequency estimate of how often a certain letter occurs in the first position, and $\Sigma\text{words}_{\text{Letter in pos 1}}$ and $\Sigma\text{words}_{\text{Letter in pos 2}}$ are the numbers of words with that letter in the first and the second position, respectively, that can be recalled during a short period of time.

Figure 1 shows that the actual and the "produced" proportions do not match well. The median produced

proportions (crosses) were consistently at or above 50%, with an average of 59.5%, a minimum of 50% (*C*, *A*, and *F*), and a maximum of 75% (*G*).[2] Thus, the produced proportions are an invalid cue for the actual proportions.

Availability-by-speed. Twenty students from the University of Munich, Germany, received the following oral instructions (translated from German):

> Your task is to recall as quickly as possible one word that has a particular letter as the first (second) letter. You will hear first the position of the letter and then the letter. From the moment you hear the letter, try to recall a respective word and verbalize this word.

The time between stimulus onset (i.e., verbal presentation of the letter by the experimenter) and reaction (i.e., verbalization of a retrieved word) was recorded by the experimenter. The experimenter and the participants were unaware of the purpose of the study. Each participant worked on 13 letters and produced one word with these letters in first and second position, respectively. The letters used and the order of presentation (production of words with a letter in the first vs. the second position) were randomly varied across participants. When the retrieved word did not satisfy the criterion (e.g., "second position, *E*") this response was not used in the analysis. The number of participants working on each letter ranged from 17 to 20.

The proportions used for the predictions of availability-by-speed ($\text{Estimate}_{\text{Letter in pos 1}}$) were calculated as follows: The time needed to recall a word with a particular letter in the second position is divided by the time needed to recall both a word with that particular letter in the first and in the second position. This can be stated as follows:

$$\text{Estimate}_{\text{Letter in pos 1}} = \text{RTword}_{\text{Letter in pos 2}}/(\text{RTword}_{\text{Letter in pos 1}} + \text{RTword}_{\text{Letter in pos 2}}),$$

where $\text{RTword}_{\text{Letter in pos 1}}$ and $\text{RTword}_{\text{Letter in pos 2}}$ are the times needed to recall a word with a certain letter in the first and second positions, respectively.

Figure 1 shows the medians of these proportions (×s). Similar to the number of words produced, the time ratio is an invalid cue for the actual proportions and was consistently higher than 50%: The RT proportions ranged from 53% (*I*) to 79% (*G*), with an average of 63%.[3]

Predictions of the Letter-Class Hypothesis

The letter-class hypothesis assumes that estimated proportions for individual letters in particular positions are inferred from the proportions in the class of consonants or vowels. In German, about 27.1% of the vowels in the first two positions in words occur in the first position (and 72.9% in the second), and 71.6% of all consonants occur in the first position (and 28.4% in the second).[4] Given the general overestimation of low and underestimation of high frequencies, one cannot expect that judgments match the actual proportions exactly. We wanted this hypothesis' predictions to be comparable with the predictions of the regressed-frequencies hypothesis (next section). Therefore, we assumed the same amount of regression (70%, or a factor of .7) here as in those predictions, as derived from Attneave's (1953) study. In Figure 1, we used these "regressed" proportions to derive the predictions for the letter-class hypothesis. For instance, the prediction for the letter *C*, a consonant, is 56.5%. To arrive at this value, we start at 71.6%, the actual value for consonants, and subtract 70% of its deviation from the mean, that is 70% of (71.6% − 50%). For the letter *O*, a vowel, the predicted value is 43.1%. By analogy, we start at 27.1%, the actual value for vowels, and add 70% of the deviation from the mean, that is, 70% of (50% − 27.1%). The predictions of the letter-class hypothesis can therefore be stated as follows:

$$\text{Estimate}_{\text{Letter in pos 1}} = \begin{cases} 56.5\%, & \text{if letter is a consonant} \\ 43.1\%, & \text{if letter is a vowel.} \end{cases}$$

Predictions of the Regressed-Frequencies Hypothesis

The regressed-frequencies hypothesis assumes that the mind keeps track of the frequencies of individual letters in different positions. It further assumes that low frequencies are overestimated and high frequencies are underestimated. The amount of this "regression" (toward the mean of all letter frequencies) is assumed to be 70%, again following Attneave's (1953) results. The prediction of the relative frequency estimate for a particular letter can be stated as follows:

$$\text{Estimate}_{\text{Letter in pos 1}} = \text{Actual} - .7\,(\text{Actual} - 50\%),$$

where Actual is the actual percentage of a particular letter occurring in the first position and the factor of .7 is the amount of "regression." Figure 1 displays the predictions of the regressed-frequencies hypothesis (open squares). For

[2] The mean values were even higher, with an average of 62%. But because of some outliers, medians are reported. The number of words produced ranged from 0 to 7. Means for the words with a specific letter in the first position ranged from 1.48 (*C*) to 3.63 (*E*), and those for words with a specific letter in the second position ranged from 1.41 (*G*) to 2.71 (*E*). Standard deviations did not vary much across letters and positions and were on average 1.06 for words with a specific letter in the first and 0.94 for words with a specific letter in the second position.

[3] The time needed to recall a word with a particular letter in the first or second position ranged from 0.45 s to 35.46 s. Means for the words with a specific letter in the first position ranged from 1.14 s (*R*) to 3.04 s (*C*), and those for words with a specific letter in the second position ranged from 1.99 (*I*) to 8.16 (*G*). Standard deviations varied considerably across letters and positions and were on average 1.05 s for words with a specific letter in the first (range from 0.29 to 2.10) and 3.22 for words with a specific letter in the second position (range from 0.81 to 7.87).

[4] Note that the percentages of vowels and consonants in first position do not have to add up to 100%, as do, for instance, the percentages of vowels (as well as consonants) in first and second positions.

instance, the estimate for the letter *C* is predicted to be 39%, calculated as 13.5% − .7 (13.5% − 50%).

Studies

We conducted three studies to test which of the hypotheses, if any, can predict human judgments. In all studies, participants were asked whether a certain letter, for instance, *R*, is more frequent in the first or the second position in all German words. We also asked them about the proportions of each letter occurring in the first and the second position. These two tasks are called the choice task and the estimation task, respectively.

Choice and Estimation Tasks

The basic task performed in all studies was a German adaptation of the Tversky and Kahneman (1973) task. All participants received the following written introductory information (translated from German):

> In a study, the frequency of occurrence of letters in the German language was analyzed. A representative sample of fiction, trivia literature, scientific and popular literature, memoirs, newspapers, and magazines was assembled. One of the findings of the study was the frequency with which various letters of the alphabet occurred in the first and the second position in words (an umlaut counts as one letter).[5] All words with at least two letters were included in the count.

Then participants first completed the choice task:

> Which result do you think emerged for the following letter? (please check one):
>
> | more frequent in the | first position | second position |
> |---|---|---|
> | *R* | —— | —— |

and later the estimation task:

> Please give an estimate of the ratio of the frequency of occurrence in the first position compared to that in the second position for the following letter:
>
> Ratio of *R* in the first position/second position
>
> ——/——

The predictions specified so far are those for the estimation task. What are the predictions for the choice task? These predictions can be derived from the estimation task. There are two versions corresponding to a maximizing and a matching strategy. First, if participants maximize, they should choose the first position in all cases where they estimated the proportion to be above 50% and they should choose the second position in all cases where their estimate was below 50%. Second, if participants follow a matching strategy, then the proportion of participants choosing the first position should coincide with the predictions specified for the estimates (Figure 1). Because we do not know which of the two strategies participants use, we will look at the interval specified by the two versions. Take the letter *C* as an example: According to the availability hypotheses, the percentage of participants choosing *C* to be more frequent in the first position should be somewhere between 100% and 50% (availability-by-number) or 56% (availability-by-speed); the letter-class hypothesis would predict a percentage between 100% and 56.5%; and the regressed-frequencies hypothesis would predict a percentage between 0% and 39%.

Three Criteria for Goodness of Fit

Apart from visual inspection of the data, we used three criteria to decide on the goodness of fit between data and predictions, the first two for estimates and the third for choice data. The first criterion is the distance between prediction and data. We used root mean squared deviations as the distance measure. This criterion, however, does not take into account the patterns (ups and downs) predicted by the individual hypotheses. For instance, a straight line at 50% (see Figure 1) could result in a small squared error for the letter-class hypothesis, although it could not be taken as support for that hypothesis. The second criterion, therefore, measures whether the pattern of estimates monotonically follows the pattern of the prediction. We chose contrast analysis (Rosenthal & Rosnow, 1985; Winer, 1971) as the measure for the covariation of predictions and estimates.[6] Contrast analysis is sensitive to monotonic changes and it allows comparison of the results within and across studies in terms of effect sizes. Similar to a correlation coefficient, it is not affected by absolute values but only by relative differences between values. The third criterion, for the choice task, is the proportion of choices that falls into the intervals specified by the predictions for choice data.

---

[5] This clarification was added because an umlaut in German (e.g., *ü*) can also be written as two letters (*ue*).

[6] Unlike the conventional use of omnibus *F* tests in analysis of variance, where the question is whether there exists any difference between conditions, contrast analysis addresses the question of whether there exist specific differences between conditions that can be derived from theoretical considerations. Weights for contrasts add up to 0. For the calculation of the weights, first the average of the results of the production task (for availability-by-number), the recall time task (for availability-by-speed), the average of the proportions of the respective letter classes (for the letter-class hypothesis), or the average of the actual proportions in the first position for individual letters (for the regressed-frequencies hypothesis) were calculated (see Figure 1). Then the deviation of the proportion for a single letter from the respective average proportion was used as the weight for that letter. Note that the weights for the different contrasts are not orthogonal. Note also that the absolute size of the weights is not relevant, only the relative difference between weights is. $MS_{contrast}$ (= $SS_{contrast}$, because $df_{contrast}$ is always 1) is calculated as: $L^2/n\Sigma\lambda^2$, where the λ are the theoretically derived weights, n is the number of observations in each condition (the harmonic mean of the overall number of observations in the case of unequal ns), and L is the sum of all weighted (by λ) condition totals (Rosenthal & Rosnow, 1985).

Study 1

Study 1 differed from Tversky and Kahneman's (1973) study in that we used both consonants and vowels, which is necessary to test the letter-class hypothesis. In contrast to their study, we also examined whether estimates differ depending on whether a letter is judged in isolation or in the context of other letters. Why should they? It could be argued that judgment by letter class is a quick, useful strategy when a single letter is judged. When two or more letters of the same class are judged, however, participants might resort to other strategies, such as that proposed by the availability hypotheses. If this were true, participants who judge one letter should give estimates according to the letter-class hypothesis but should switch to a different strategy when they subsequently learn that they have to judge more letters. For instance, when participants do not know that there are more judgments to come, their estimates for the first consonant should correspond to the class of consonants, whereas their estimates for the same consonant given last in a series should not.

Stimulus materials were four lists consisting of eight letters each—two that consisted solely of consonants and two that were "mixed": List 1 (only consonants): *R, F, C, B, L, G, S, N*; List 2 (List 1 in reversed order): *N, S, G, L, B, C, F, R*; List 3 (mixed list): *O, F, E, B, I, G, A, U*; and List 4 (List 3 in reversed order): *U, A, G, I, B, E, F, O*.

For instance, in List 1, participants began with the letter *R*, which is an atypical consonant (as is *N*) in the sense that it occurs 20% of the time in the first position, whereas on average, consonants occur 71.6% of the time in the first position. According to the letter-class hypothesis, participants should estimate that this letter occurs more frequently in the first position about 56.5% of the time. The same result should hold if the letter *R* occurs in the last position of the list (List 2). If, however, the letter-class hypothesis were restricted to the first judgment, then judgments about *R* in Lists 1 and 2 should not be equal. For instance, if the availability heuristic (availability-by-number) were activated in multiple-letter judgments, the estimate should be about 60%; if one takes the regressed actual values as predictions, the estimate should be around 41.3%. Predictions for the other letters in Lists 1–4 can be derived in the same way. In sum, Study 1 examined which of the four hypotheses predicts the data best and whether judging a letter in isolation versus in the context of other letters made a difference.

Method

One hundred fourteen students of the University of Salzburg, Austria, were paid for their participation. One to eight ($M = 5$) participants took part in each session. There were four groups of participants, each receiving one of the four letter lists. At the beginning of the study, participants were handed a small booklet. First, they read the introductory information about the text corpus (see the *Choice and Estimation Tasks* section discussed earlier) and then performed the choice and estimation tasks for one letter. They were not informed that there were more judgments to come. Then they were asked to continue the tasks for the other seven letters on seven successive pages.

Results and Discussion

Study 1 attempted to replicate Tversky and Kahneman's (1973) study with a larger number of letters and aimed to examine whether judging a letter in isolation versus in the context of other letters makes a difference. A first, visual comparison of the results of Study 1 with the four hypotheses' predictions displayed in the four panels of Figure 2 (estimation tasks) and Figure 3 (choice tasks) indicates that the regressed-frequencies hypothesis fits the data best. The presentation of the results in Figures 2 and 3 follows Figure 1. The letters are rank ordered according to their proportion in the first position from left to right. The ratios obtained in the estimation task were transformed into "percent in the first position" (Figure 2), and the percentages in the choice task reflect the proportion of the participants who chose "first position" for a particular letter (Figure 3).

Did it make a difference whether letters were judged in isolation or in the context of other letters? No: The mean difference across all letters between judgments in List 1 and in List 2 was −2.8 percentage points, and no systematic difference between letters occurring in the initial and final positions in lists were found. A similar result was found with the mixed lists (Lists 3 and 4), where the mean difference was even smaller. Therefore, the results shown in Figures 2 and 3 are collapsed across lists.

We now turn to our three criteria for goodness of fit. One might argue that the availability hypotheses are at a disadvantage by not relying on regressed values as do the letter-class and regressed-frequencies hypotheses. We could not think of a compelling reason, nor do we know of any empirical evidence, for why the production data should be regressed. Moreover, it is not clear against which value the predictions based on production data should be regressed. Should it be the mean of a representative sample of production data or should it be 50%? Nonetheless, we decided to treat the availability predictions analogously to the other two predictions. Because compared to no regression and to regression toward the sample mean, regression toward 50% yielded the most favorable results for the availability hypotheses in all studies, all of the following analyses are based on these regressed values. Note that the availability predictions in the upper panels of Figures 2 and 3 are those from Figure 1 but now regressed toward 50%.

Figure 4 (leftmost two groups of columns) shows that the root mean squared deviations are smallest for the regressed-frequencies hypothesis for both the mixed and the consonant lists. Table 1 shows the results of the contrast analysis. It also shows the effect sizes associated with the four hypotheses, expressed as Pearson *r* correlation coefficients (e.g., Rosenthal & Rosnow, 1991). For each study and each of the relevant predictions, the table shows MS_{contrast}, MS_{error}, df_{error}, and *r*. The larger the (positive) *r*, the more the data monotonically follow the predictions of the hypotheses. For both mixed and consonant lists, the regressed-frequencies hypothesis competes best. Note that not absolute but relative deviations

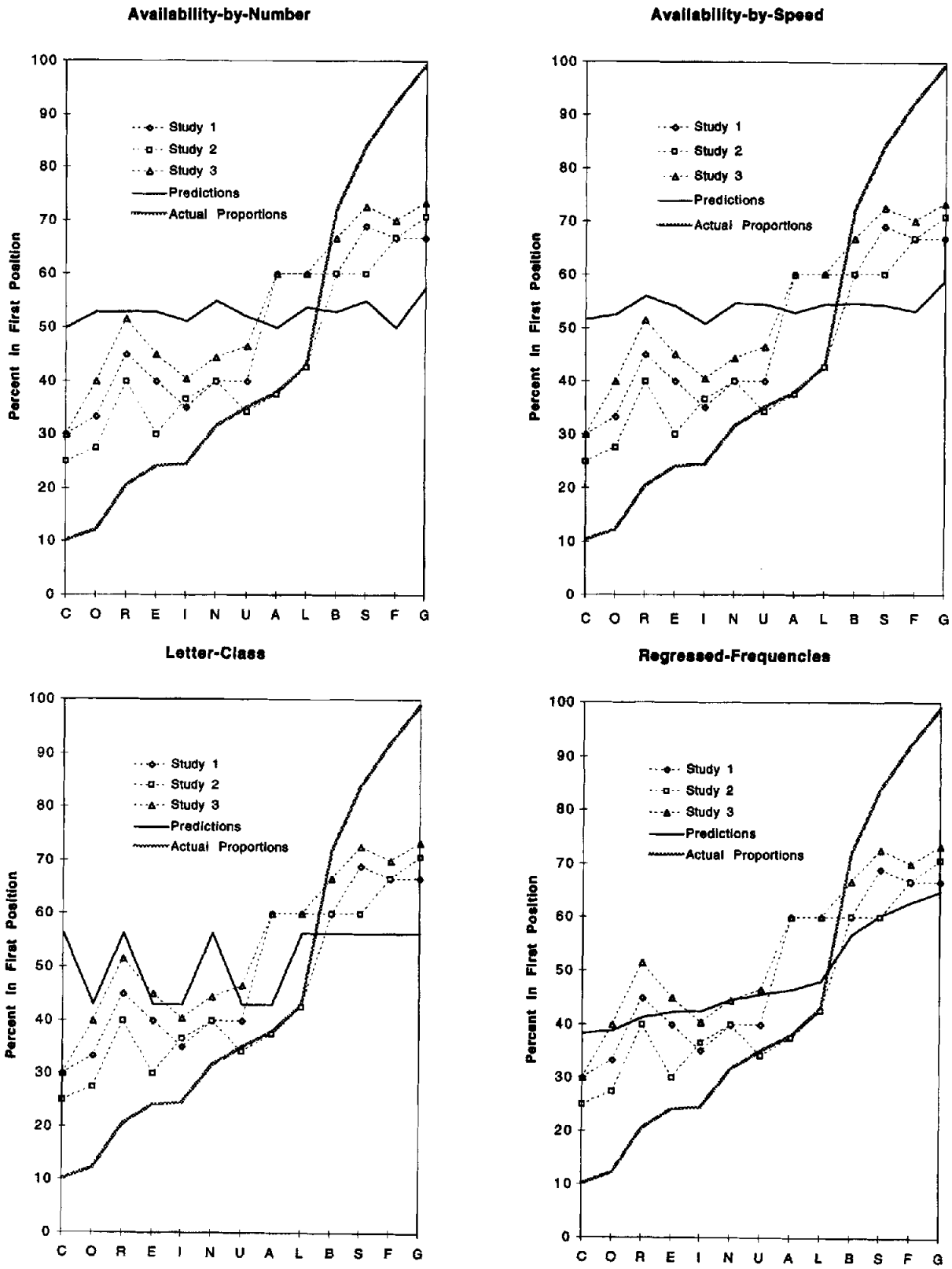

Figure 2. Results of estimation tasks obtained in Study 1, Study 2, and Study 3 plotted against the four hypotheses' predictions. All predictions are regressed toward 50%. Data are medians of ratio estimates transformed into percentage in the first position.

determine the results in a contrast analysis. That means that one obtains identical results for nonregressed and regressed predictions. For instance, in the case of the letter-class hypothesis, it does not make a difference whether a contrast analysis uses predictions of 71.6% versus 27.1%, or 56.5% versus 43.1% for consonants and vowels, respectively.

Also the third criterion for goodness of fit, the percentage of choices consistent with the hypothesis' prediction favors the regressed-frequencies hypothesis, although it is closely followed by the letter-class hypothesis (Figure 5, first group of bars).[7] Thus all three criteria for goodness of fit favor the regressed-frequencies hypothesis followed by the letter-class and the availability hypotheses.

Study 2

Participants' judgments in Study 1 did not differ systematically depending on whether a letter was judged in isolation (i.e., first in a series of letters) or in the context of other letters (i.e., last in a series of letters). This manipulation was introduced to examine whether the letter-class hypothesis held under a condition that appears to favor the application of this strategy, that is, judgment in isolation. However, only the letters *N*, *R*, *O*, and *U* were placed in the first and last position of the letter lists. Such a small sample may limit the generalizability of this test. In addition, participants in Study 1 had received a booklet that might have led some of them to anticipate more letter judgments to come, and thus to discard the letter-class strategy.

To control for these two issues, we conducted Study 2. Study 2 is another test of the letter-class hypothesis under conditions more favorable to it. We examined judgments in isolation for all of the letters used in Study 1. To make sure that participants perceived their task as a judgment in isolation, all of them had to judge only one letter. Study 2 also investigated whether the result found in Study 1 that favored the regressed-frequencies hypothesis over the letter-class hypothesis and the availability hypotheses could be replicated.

Method

Three hundred ninety students from the University of Munich, Germany, participated. They received the instructions used in Study 1 and performed the choice and estimation task for one single letter only. We attempted to avoid participants' anticipations of more letters by printing the task on a single page, and by informing them that the task would take only a few moments. Every letter was judged by 30 participants who were approached individually on campus.

Results and Discussion

Visual inspection shows that none of the four consonants, *C*, *R*, *N*, and *L*, for which letter-class and regressed-frequencies hypotheses make divergent predictions was judged over 50% (Figure 2). This is inconsistent with the letter-class hypothesis. The general pattern of results follows that of Study 1: The three criteria for goodness of fit support this conclusion. First, the root mean squared deviations are markedly smaller for the regressed-frequencies hypothesis than for any other hypothesis (third group of bars in Figure 4). Second, the fit as expressed by the contrast analysis is best for the regressed-frequencies hypothesis, followed at some distance by the other hypotheses (Table 1). Finally, the choice data favor the regressed-frequencies hypothesis over the letter-class and the two availability hypotheses (Figures 3 and 5). We conclude that the results do not favor the letter-class hypothesis even under conditions that are favorable to this hypothesis: If all letters are judged in isolation, the letter-class hypothesis does not fit the data as well as the regressed-frequencies hypothesis.

Study 3

The results of Study 1 and 2 provided little support for the availability explanation. As we did with the letter-class hypothesis, however, we wanted to know whether the availability explanation can be retained under conditions that should be favorable to the application of the availability heuristic. In Study 3, we tested whether the availability explanation can be retained in the special case in which production precedes estimation. The estimation task directly followed the production task. The question is: Do proportion estimates differ depending on whether or not a production task is performed first? Study 3 was identical to Study 1 except for the inclusion of a production task. If the production task had the effect of triggering a "recall-estimate" strategy in the subsequent estimation task, then the goodness of fit of availability-by-number should markedly increase.

Method

Seventy-seven students at the University of Salzburg served as participants and were paid for their participation. The same materials were used as in Study 1. Participants had to produce words with the letters *O*, *U*, *N*, and *R* in the first and the second position. The production task lasted 90 s for each letter. After that, participants proceeded as in Study 1. The study was conducted in small groups of an average of 5 persons.

Results and Discussion

The results suggest that the preceding production task had a quantitative but not a qualitative effect on choices and

[7] The a priori probability that choices fall into the interval predicted for choice data is not the same for the four hypotheses. With increasing average size of the predicted intervals, one would expect an increasing percentage of choice data corresponding to our predictions by chance alone. For all studies, the respective percentages are 47.2%, 46.0%, 43.3%, and 41.8% for availability-by-number, availability-by-speed (both based on regressed values that were used in all analyses), the letter-class hypothesis, and the regressed-frequencies hypothesis. Thus, the a priori probabilities slightly favor the availability hypotheses.

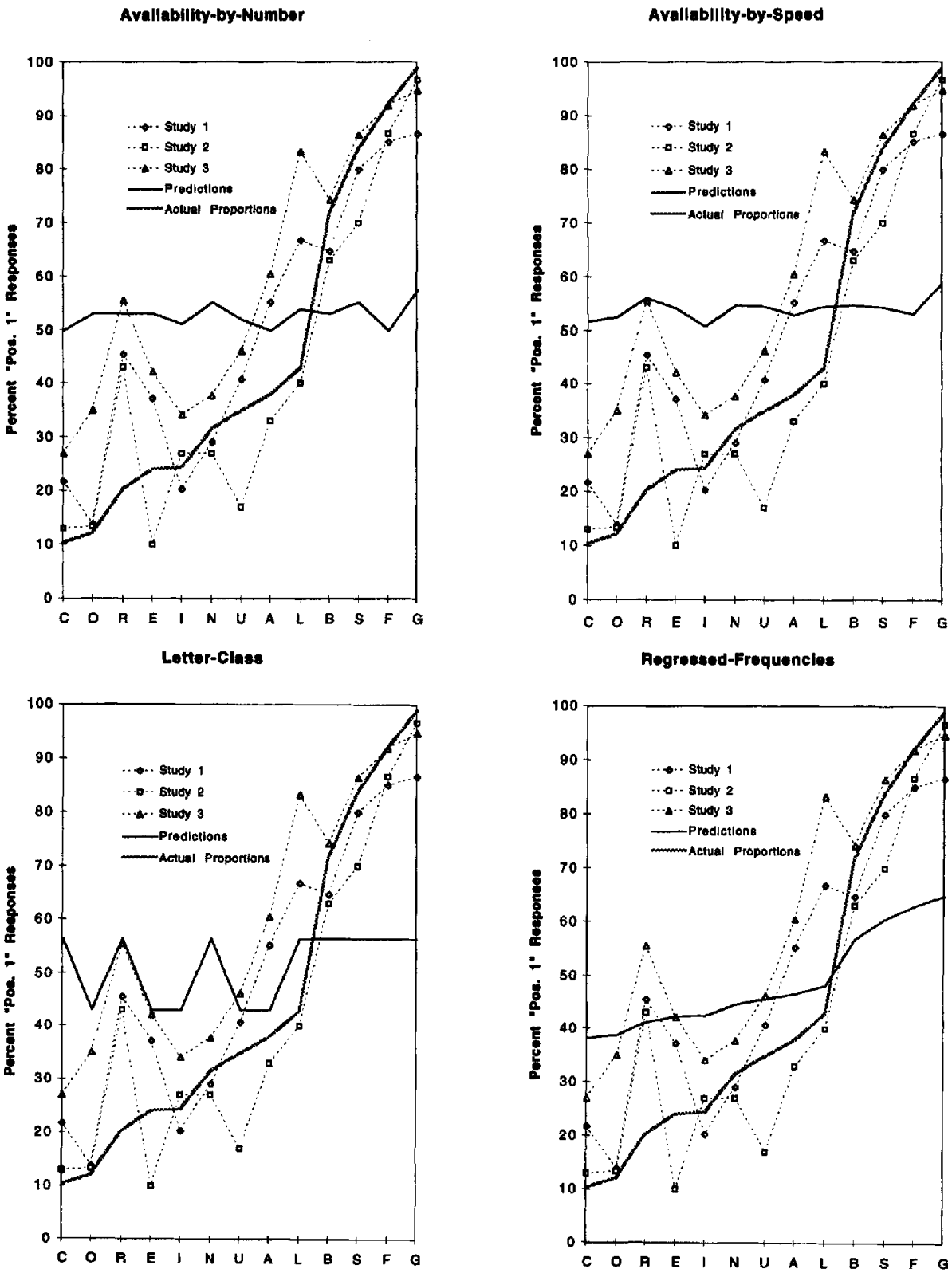

Figure 3. Results of choice tasks obtained in Study 1, Study 2, and Study 3 plotted against the four hypotheses' predictions. All predictions are regressed toward 50%. Data are percentages of participants who decided that a particular letter occurs more often in the first position. Pos. = position.

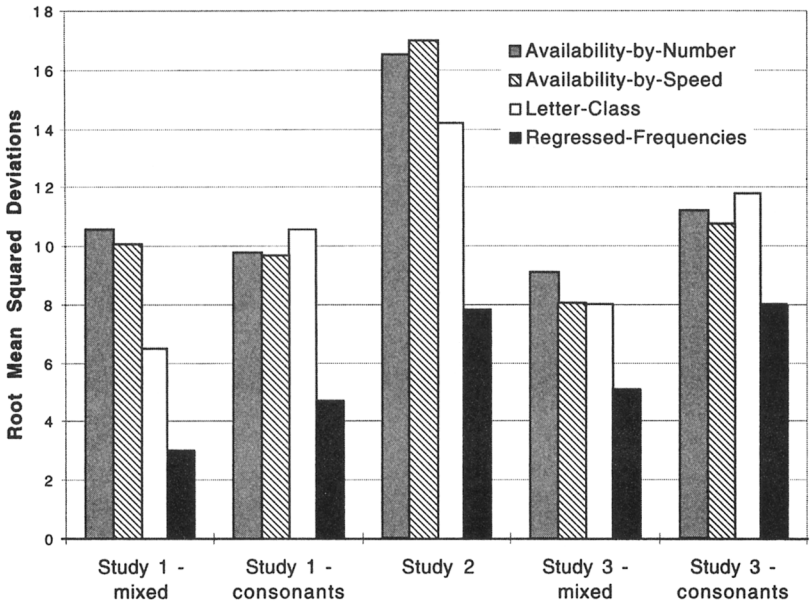

Figure 4. Root mean squared deviations between predictions derived from the four hypotheses and estimates obtained in the three studies.

estimates; that is, estimates and number of choices for the first position were slightly larger than in previous studies but the rank ordering was unaffected. This effect is more pronounced with the choice data (Figure 3).

Figure 2 shows that despite the preceding production task, the estimated proportions in Study 3 follow the same pattern as in Studies 1 and 2. The three criteria for goodness of fit support this conclusion. First, the root mean squared deviations are smallest for the regressed-frequencies hypothesis (Figure 4, fourth and fifth groups of bars). Second, the order relation was best predicted by the regressed-frequencies hypothesis as found in the contrast analysis (Table 1). Third, the analysis of the choice data gives a result similar to the ones obtained in Studies 1 and 2: The regressed-frequencies hypothesis achieves the best fit, followed by the letter-class hypothesis and the availability hypotheses (Figure 5, third group of bars). In summary, the availability explanation cannot be retained even under circumstances that are favorable to the application of the availability explanation.

General Discussion

Many textbooks on cognitive psychology report a classical demonstration of availability: Judgments of relative letter frequencies are systematically biased. Research on judgments of event frequency in a wide range of environments, however, indicates that estimates often reflect the actual relative frequencies of the events with great fidelity. We were concerned with this puzzle. We proposed and tested four hypotheses of judgment of relative frequencies of letters in different positions: two versions of the availability heuristic, the letter-class hypothesis, and the regressed-frequencies hypothesis. Three studies showed consistently that people's judgments conformed best to the predictions of the regressed-frequencies hypothesis. The three criteria for goodness of fit—root mean squared deviations, the results of contrast analyses, and the analysis of the choice data—converged toward this result. This provides strong evidence against the classical availability heuristic explanation.

How Much Better Does the Regressed-Frequencies Hypothesis Do?

In all studies, the root mean squared deviations for the regressed-frequencies hypothesis were smaller than those for the letter-class hypothesis (Figure 4),[8] and the effect sizes consistently favored the former (Table 1). Moreover, the number of choice tasks that fulfill the hypotheses' criteria was always larger for the regressed-frequencies hypothesis (Figure 5). Another way of comparing the two hypotheses is to quantify their difference by comparing the respective contrasts (Rosnow & Rosenthal, 1996, p. 256).[9] This

[8] From the fit between the data and the predictions of the regressed-frequencies hypothesis (lower right portion of Figure 2), it appears that the 70% regression found in Attneave's (1953) study describes the current results well. On average, the exact amount of regression that would minimize the root mean squared deviations would be only slightly smaller. The optimum amount of regression would be 67% and 69% for the mixed and consonant lists in Study 1, respectively, 50% in Study 2, and 71% and 66% for the mixed and consonant lists in Study 3, respectively.

[9] Two contrasts can be compared by creating a new contrast out of the differences between the original contrast weights. The $SS_{contrast}$ for the difference between two contrasts is calculated as $nL^2/\Sigma\lambda_{diff}^2$, where λ_{diff} are the new weights and L is the weighted (by λ_{diff}) sum of the means in all conditions. The k new weights are obtained by first standardizing the weights of the two original

Table 1
Contrast Analyses for the Results in the Estimation Task of Studies 1, 2, and 3

Study and prediction	$MS_{contrast}$	MS_{error}	df_{error}	r
Study 1				
Mixed				
Availability-by-number	4,790.2	326.1	399	.19
Availability-by-speed	25,935.6	326.1	399	.41
Letter class	40,014.5	326.1	399	.48
Regressed frequencies	50,504.7	326.1	399	.53
Consonants[a]				
Availability-by-number	7,487.8	334.4	364	.24
Availability-by-speed	11,085.8	334.4	364	.29
Regressed frequencies	42,614.1	334.4	364	.51
Study 2				
Availability-by-number	15,019.0	281.4	377	.35
Availability-by-speed	26,270.8	281.4	377	.45
Letter class	28,051.6	281.4	377	.46
Regressed frequencies	79,435.5	281.4	377	.65
Study 3				
Mixed				
Availability-by-number	2,873.7	258.4	259	.20
Availability-by-speed	10,582.0	258.4	259	.37
Letter class	20,850.5	258.4	259	.49
Regressed frequencies	25,015.3	258.4	259	.52
Consonants[a]				
Availability-by-number	6,984.7	258.3	245	.32
Availability-by-speed	8,016.2	258.3	245	.34
Regressed frequencies	27,240.0	258.3	245	.55

Note. Letters used for calculating the contrasts were *C, O, R, E, I, N, U, A, L, B, S, F,* and *G* in Study 2. For the mixed lists in Studies 2 and 3, the letters *O, E, I, U, A, B, F,* and *G* were used to calculate the contrasts, and for the consonant lists of these two studies, the letters used were *C, R, N, L, B, S, F,* and *G*. The *F* value can be calculated by dividing $MS_{contrast}$ by MS_{error}. The correlation coefficient as a measure of effect size (last column) is calculated by the formula $r = [F/(F + df_{error})]^{1/2}$ (e.g., see Rosenthal & Rosnow, 1991).
[a] In the case of an all-consonants list, there is no contrast to be predicted by the letter-class hypothesis.

comparison reveals a weighted (by *df*) mean effect size of $r = .24$ (see Table 2), corresponding to a small to medium effect according to Cohen's (1992) classification. Because this value is based on over 1,000 *df*, it reliably expresses the difference between the two hypotheses.[10] Larger effect sizes were obtained when comparing the regressed-frequencies hypothesis to the availability hypotheses. The weighted (by *df*) mean effect sizes corresponding to the differences between the regressed-frequencies hypothesis and the availability-by-number and the availability-by-speed hypotheses are $r = .33$ and $r = .27$, respectively, corresponding to a medium-sized effect according to Cohen's conventions. Table 2 shows that differences of about this size were consistently obtained in all studies.

The Availability Heuristic Reconsidered

Tversky and Kahneman (1973) reported that their participants' estimates of relative letter frequencies were severely biased, which they attributed to application of the availability heuristic. Using a larger set of letters, we consistently found that both estimation and choice judgments followed the predictions of the regressed-frequencies hypothesis. Thus, it seems that humans are able to estimate relative letter frequencies—albeit in a regressed form—after all. However, the validity of this conclusion depends on our operationalization of the alternative hypotheses in our studies, and possible limitations need to be addressed.

Vagueness of availability. We have proposed two ways to measure availability: as the number of instances retrieved within 10 s (availability-by-number) and the ease with which a word with a letter in a certain position comes to mind (availability-by-speed). The specific parameters chosen (e.g., the 10-s interval) were motivated by our observation that when people are asked to perform the choice task, it takes them very little time to respond. The concept of "availability" has not been precisely defined (Gigerenzer, 1996a), and we cannot test and refute its every possible incarnation, but our results speak strongly against two plausible ones.

Language of participants. Tversky and Kahneman's (1973) original study was conducted in English, whereas our studies were conducted in German. We assume that the input on which letter-frequency estimates are based is visual information. Because that information is the same for both languages, there is no reason to expect that our results would turn out differently if the experiments were conducted in English. If the input format were phonological, however, then this generalization would be harder to defend because the correspondence between a written letter and its pronunciation is much looser in English than in German (e.g., Wimmer & Hummer, 1990), and frequency counting based on English phonemes could lead to diverging estimates for written letters.

We mentioned earlier that there exists, to the best of our knowledge, only one attempt to replicate Tversky and Kahneman's (1973) original study on the first versus third position in English with a larger set of letters (White, 1985, 1991). The results of this study can be used to evaluate the generalizability of our results—at least for the choice data—from German to English. White (1985) used Tversky and Kahneman's (1973) five consonants (*K, L, N, R,* and *V*), all more frequent in the third position, and five other

contrasts (by dividing them by the standard deviation of the weights as defined by $\sigma_\lambda = (\Sigma\lambda^2/k)^{1/2}$) and then taking the difference between the standardized weights (Rosnow & Rosenthal, 1996, p. 256).

[10] Even if one only trusts in results of significance tests, already the smallest effect alone (Study 3–mixed, see Table 2) yields a large *F* value: $F(1, 259) = 6.46$, $p = .012$.

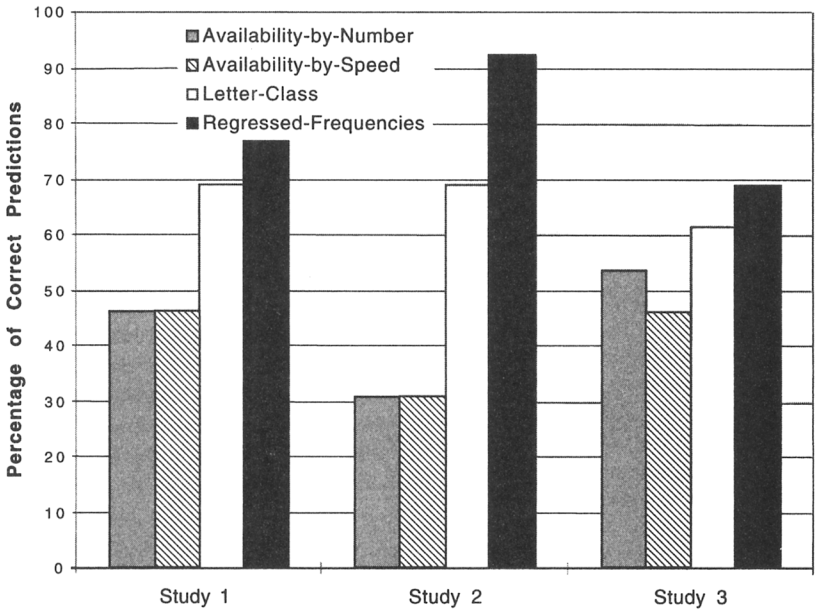

Figure 5. Percentages of choices that fell into an interval specified by predictions derived from the four hypotheses.

consonants (*F, J, G, M* and *Q*) that are more frequent in the first position. White reported the results for individual letters for the choice task only. All five consonants that appear more often in the first position were judged to do so. Moreover, three of the five consonants that appear more often in the third position (*N, R,* and *V*) were judged to appear in that position more frequently. Judgments for the letters *K, L, R,* and *V,* which were also used by Tversky and Kahneman (1973), showed the highest upward deviations from the actual values. White's (1985) finding suggests that a larger set of letters (such as that used in our studies) would lead to a similar result in English and German: Participants' choices roughly follow the actual proportions.

Types versus tokens. As pointed out earlier, because the notion of the availability heuristic has been only vaguely sketched, it is consistent with several different mechanisms (Fiedler, 1983; Lopes & Oden, 1991; Schwarz et al., 1991). Applied to the letter task, one point of vaguenesses concerns the distinction between types and tokens. Word types are represented by words in the dictionary, each of which appears only once, whereas word tokens are instances of a word type, and thus can represent repetitions of the same word. In proposing that "the availability of instances could be measured by the total number of instances retrieved or constructed in any given problem," Tversky and Kahneman (1973, p. 210) did not specify whether "instances" are types

Table 2
Predictive Power of the Contrasts for the Regressed-Frequencies Hypothesis Relative to Those for the Other Hypotheses

Study	MS_{error}	df_{error}	Letter class		Availability-by-number		Availability-by-speed	
			$MS_{contrast}$	r	$MS_{contrast}$	r	$MS_{contrast}$	r
Study 1								
Mixed	326.1	399	5,669.7	.20	18,219.7	.35	5,739.3	.21
Consonants[a]	334.4	364			11,023.1	.28	8,602.4	.26
Study 2	281.4	377	12,755.9	.33	20,773.6	.40	14,821.1	.35
Study 3								
Mixed	258.4	259	1,668.7	.16	7,801.4	.32	4,098.4	.24
Consonants[a]	258.3	245			5,089.5	.27	4,795.0	.27
Weighted mean (by *df*)				.24		.33		.27

Note. New contrasts were created out of the differences between the original contrast weights (see Rosnow & Rosenthal, 1996). Results are based on the estimation task of Studies 1, 2, and 3. Letters used for calculating the contrasts were *C, O, R, E, I, N, U, A, L, B, S, F,* and *G* in Study 2. For the mixed lists in Studies 2 and 3, the letters *O, E, I, U, A, B, F,* and *G* were used to calculate the contrasts, and for the consonant lists of these two studies, the letters used were *C, R, N, L, B, S, F,* and *G.* The *F* value can be calculated by dividing $MS_{contrast}$ by MS_{error}. The correlation coefficient r as a measure of effect size is calculated by the formula $r = [F/(F + df_{error})]^{1/2}$ (e.g., Rosenthal & Rosnow, 1991).

[a]In the case of an all-consonants list, there is no contrast to be predicted by the letter-class hypothesis.

or tokens. If one measures availability in terms of number of words produced (availability-by-number), and if participants only produce types (as most of our participants did by not reiterating the same word), then comparison to a token text corpus such as Tversky and Kahnman's (1973) and ours is inappropriate. This problem does not arise in availability-by-speed because participants only have to produce a single word. However, neither formulation of the availability hypothesis could account for people's judgments, which closely followed the rank order of the actual token frequency in the text corpus.[11]

Possible explanations for discrepant conclusions. Although we cannot be sure why the present results differ from those of Tversky and Kahneman (1973), we can think of two possible explanations. First, Tversky and Kahneman's (1973) participants might not have been able to disregard one- and two-letter words in their judgments as they were instructed to do and therefore overestimated the frequency of occurrence of letters in the first position.

The second explanation has to do with Brunswik's (1955) notion of representative sampling. If a person has experienced a representative sample of objects from a reference class, his or her mental model should be better adapted to that environment than if he or she happened to experience an unrepresentative sample. If one makes the plausible assumption that people experience a representative sample of letters and their positional frequencies (e.g., during reading), then their mental models should be adapted to a representative sample presented by the experimenter. In the letter estimation task, Tversky and Kahneman (1973) presented participants with an unrepresentative sample of letters, namely, the consonants *K*, *L*, *N*, *R*, and *V*. This sample is unrepresentative because each of these five consonants is more frequent in the third than the first position (see Lopes & Oden, 1991), whereas the majority of consonants (12 of 20) are actually more frequent in the first position.

How could testing people on an unrepresentative sample of consonants produce a conclusion discrepant from ours? Tversky and Kahneman (1973) assumed that there is a "bias favoring the first position" (p. 212). Now assume that we had tested only consonants that are more frequent in the second position (in our studies, these were *C*, *R*, *N*, and *L*), that is, an unrepresentative sample like Tversky and Kahneman's (1973). In that case, we would have reported that the frequency of those letters in the first position (see Figure 2) is overestimated, just as Tversky and Kahneman's notion of availability suggests. However, if we consider all of the consonants in Figure 2—*C*, *R*, *N*, *L*, *B*, *S*, *F*, and *G*—then we find that the frequency of those that are more frequent in the first position is actually underestimated. This result is inconsistent with Tversky and Kahneman's (1973) conclusion that there is a bias favoring the first position. We interpreted it as a regression effect.

The issue of representative versus selected sampling is important for more than the estimation of letter frequencies. For instance, Gigerenzer, Hoffrage, and Kleinbölting (1991) proposed that selected sampling of general-knowledge questions produces the overconfidence bias. In a meta-analysis that compares over- and underconfidence in published data with representative versus selected samples of general-knowledge questions, Juslin, Hoffrage, and Gigerenzer (1997) showed that there is a clear effect of item selection, there being less overconfidence for representative samples. Moreover, Gigerenzer, Hell, and Blank (1988) found that random sampling reduces base-rate neglect.

Feature Overlap—A Possible Explanation for Regression

What could be the mechanisms that produce the two main results, the close agreement in the rank ordering of real and estimated proportions and the regression toward 50% found in the estimates? We propose that regression in frequency judgments can be explained by feature overlap. Consider, for instance, the letters *B* and *P*, whose upper halves perfectly overlap. Feature overlap in the input may reduce discriminability of letters (and other encoded events), which in turn may lead to regression toward equal estimated frequencies. This hypothesis can be tested by means of simulation. We used two different classes of models, exemplar and neural network models (see Estes, 1991, for a detailed discussion of these classes).

We chose MINERVA 2 (Hintzman, 1988) as an exemplar model and ELF as a neural network model (Sedlmeier, Hertwig, & Gigerenzer, 1995).[12] The input to both models was identical and consisted of pairs of visual patterns. Each pattern pair was represented as a vector of features. In the case of letter strings, a feature can be taken to be a pixel on a screen or on a sheet of paper that is either part of a letter (feature is present) or not (feature is absent). We let both models encode sets of pattern pairs. Within all sets, each of the five patterns used appeared with a fixed probability in the left and in the right half of the pattern pair. The sets differed only in the amount of overlap between patterns. Both models' "estimates"[13] of relative frequencies in left position reflected the original rank ordering and strongly correlated with amount of overlap among patterns. For patterns that did

[11] People also might use a mixed strategy, that is, use types and weigh them by their frequency of occurrence. This could affect estimates if, for instance, some high frequency words contain a particular letter in first or second position. Unfortunately, the Mannheimer corpus reports only the overall letter frequency at a certain position, and therefore the impact of differential word frequency could not be included in our predictions. However, people do not seem to use such a mixed strategy: From the 390 participants of Study 2 who were individually asked how they had arrived at their answers, not a single one indicated having weighted the estimates by the frequencies of specific words.

[12] MINERVA 2 is described in several publications (e.g., Hintzman, 1984, 1988). ELF is a feed-forward network, consisting of an input and an output layer that are fully connected. It learns with a variant of the competitive learning mechanism (e.g., Grossberg, 1976; Rumelhart & Zipser, 1985). For MINERVA 2, encoding a new pattern of features (e.g., a pair of letters) means adding the corresponding vector to an ever-growing matrix of vectors. ELF, in contrast, stores all encoded vectors in an indirect way by changing the connection weights between input and output nodes.

not overlap, there was a negligible amount of regression, but for those that overlapped to a substantial degree, the degree of regression toward 50% was also substantial.

Thus, despite considerable differences in architecture and encoding processes, both the exemplar and neural network models can in principle account for the results found in the present studies, suggesting that the effect of feature overlap on letter discriminability can explain regression in frequency estimates. MINERVA 2 and ELF provide possible starting points for developing a general model of frequency processing.

Conclusion

Tversky and Kahneman's (1973) findings on letter frequency judgment have become one of the stock-in-trade examples of a "bias" in the heuristics and biases literature. The results of three studies indicate that this chapter in the heuristics and biases literature needs to be rewritten. These findings also cast doubt on the "reality of cognitive illusions" in frequency and probability judgments (see the debate between Kahneman & Tversky, 1996, and Gigerenzer, 1996a). There is increasing evidence that these so-called cognitive illusions largely disappear if participants are given frequency information and asked for frequency judgments instead of probability judgments. For example, Gigerenzer (1994; Gigerenzer et al., 1991) showed that the "overconfidence bias" disappears when participants estimate the number of correct answers instead of the probability that a particular answer is correct (see also May, 1987; Sniezek & Buckley, 1993). Fiedler (1988) and Hertwig and Gigerenzer (1997; Hertwig, 1995) showed that the "conjunction fallacy" in the Linda problem and similar conjunction problems largely disappears (from about 80%–90% to 0%–20% of participants violating the conjunction rule) when the task requires estimation of frequencies instead of single-event probabilities (see also Reeves & Lockhart, 1993; Tversky & Kahneman, 1983). Bayesian reasoning improves in lay people (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995) and experts (Gigerenzer, 1996b; Hoffrage & Gigerenzer, 1996) when Bayesian-type problems are presented in a frequency format rather than in a single-event probability format. Frequency formats have also proven very helpful in training people in conjoint and conditional probability judgment and Bayesian inference (Sedlmeier, 1997a, 1997b; Sedlmeier & Gigerenzer, 1996). These findings are consistent with the idea that minds are tuned to reason about frequencies and stand in sharp contrast to the claim that people are systematically biased in judging frequencies as a result of their reliance on the availability heuristic. Of course, these results should not be interpreted to mean that frequency judgments are always accurate. To the extent that we succeed in modeling the mechanisms underlying frequency judgments, we will be able to predict when frequency judgments are not accurate as well as when they are (e.g., Gigerenzer & Hoffrage, 1995; Gigerenzer et al., 1991).

More generally, there exists a time-honored philosophical position according to which the mind unconsciously tallies event frequencies (Gigerenzer et al., 1989). Hume (1739/1975), for instance, insisted that psychological mechanisms that monitor frequencies are extremely finely tuned: "When the chances of experiments on one side amount to ten thousand, and on the other to ten thousand and one, the judgment gives the preference to the latter, upon account of that superiority" (p. 141). Our results do not substantiate Hume's strong claim, but they do support the claim of recent research on frequency processing that estimates of the relative frequencies of many kinds of events reflect their actual frequencies with great fidelity. Computational models such as MINERVA 2 and ELF might prove useful in building a precise model of how the mind achieves such accuracy in frequency judgments.

[13] When encoding was finished for one specific set of patterns, both models were given a prompt consisting of a pattern in either first or second position of the pair. The response of a model on a pattern in the first position was divided by the sum of the responses elicited by that pattern in both positions and so yielded an "estimate" of relative frequency. In the case of MINERVA 2, the response reflects some kind of match (called "echo intensity"—Hintzman, 1988) between the probe and all the vectors stored in the matrix, and in the case of ELF, the response consists of the sum of activations over all output units.

References

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology, 46,* 81–86.

Barsalou, L. W. (1992). *Cognitive psychology: An overview for cognitive scientists.* Hillsdale, NJ: Erlbaum.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

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

Estes, W. K. (1991). Cognitive architectures from the standpoint of an experimental psychologist. *Annual Review of Psychology, 42,* 1–28.

Fiedler, K. (1983). On the testability of the availability heuristic. In R. W. Scholz (Ed.), *Decision making under uncertainty* (pp. 109–119). Amsterdam: North-Holland.

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

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129–162). New York: Wiley.

Gigerenzer, G. (1996a). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103,* 592–596.

Gigerenzer, G. (1996b). The psychology of good judgment: Frequency formats and simple algorithms. *Journal of Medical Decision Making, 16,* 273–280.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 513–525.

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.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life.* Cambridge, England: Cambridge University Press.

Greene, R. L. (1984). Incidental learning of event frequencies. *Memory & Cognition, 12,* 90–95.

Greene, R. L. (1989). Negative practice effects on frequency discrimination. *American Journal of Psychology, 102,* 225–232.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics, 23,* 121–134.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108,* 356–388.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39,* 1372–1388.

Hertwig, R. (1995). *Why Dr. Gould's homunculus doesn't think like Dr. Gould: The "conjunction fallacy" reconsidered.* Konstanz, Germany: Hartung-Gorre.

Hertwig, R., & Gigerenzer, G. (1997). *The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors.* Manuscript submitted for publication.

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of repetitions. *Journal of Experimental Psychology, 80,* 139–145.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, and Computers, 16,* 96–101.

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

Hock, H. S., Malcus, L., & Hasher, L. (1986). Frequency discrimination: Assessing global-level and element-level units in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 232–240.

Hoffrage, U., & Gigerenzer, G. (1996). The impact of information representation on Bayesian reasoning. In G. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 126–130). Mahwah, NJ: Erlbaum.

Hume, D. (1975). *A treatise of human nature.* Oxford, England: Clarendon Press. (Original work published 1739).

Institut für deutsche Sprache. (1968, December). *Forschungsberichte des Instituts für deutsche Sprache, Band 2.* Mannheim, Germany: Institut für deutsche Sprache.

Institut für deutsche Sprache. (1969, April). *Forschungsberichte des Instituts für deutsche Sprache, Band 3.* Mannheim, Germany: Institut für deutsche Sprache.

Johnson, M. K., Peterson, M. A., Yap, E. C., & Rose, P. M. (1989). Frequency judgments: The problem of defining a perceptual event. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 126–136.

Jonides, J., & Jones, C. M. (1992). Direct coding for frequency of occurrence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 368–378.

Juslin, P., Hoffrage, U., & Gigerenzer, G. (1997). *Connecting minds to environments: A neo-Brunswikian approach to judgment and decision making.* Unpublished manuscript.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103,* 582–591.

Lopes, L. L., & Oden, G. D. (1991). The rationality of intelligence. In E. Eels & T. Maruszewski (Eds.), *Poznan studies in the philosophy of the sciences and the humanities* (Vol. 21, pp. 225–249). Amsterdam: Rodopi.

May, R. S. (1987). *Realismus von subjectiven Wahrscheinlichkeiten: Eine kognitionspsychologische Analyse inferentieller Prozesse beim Over-confidence Phänomen* [Calibration of subjective probabilities: A cognitive analysis of inference processes in overconfidence]. Frankfurt, Germany: Lang.

Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York: Freeman.

Mayzner, M. S., & Tresselt, M. E. (1965). Tables of single-letter and bigram frequency counts for various word-length and letter-position combinations. *Psychonomic Monograph Supplements, 1,* 13–32.

Naveh-Benjamin, M., & Jonides, J. (1986). On the automaticity of frequency coding: Effects of competing task load, encoding strategy, and intention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 378–386.

Reeves, R., & Lockhart, R. (1993). Distributional versus singular approaches to probability and errors in probabilistic reasoning. *Journal of Experimental Psychology: General, 122,* 207–226.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* Cambridge, England: Cambridge University Press.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Rosnow, R. L., & Rosenthal, R. (1996). Contrasts and interactions redux: Five easy pieces. *Psychological Science, 7,* 253–257.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science, 9,* 75–112.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61,* 195–202.

Sedlmeier, P. (1997a). BasicBayes: A tutor system for simple Bayesian inference. *Behavior Research Methods, Instruments, and Computers, 29,* 328–336.

Sedlmeier, P. (1997b). *How to improve statistical thinking: Choose the task representation wisely and learn by doing.* Manuscript submitted for publication.

Sedlmeier, P., & Gigerenzer, G. (1996). *Teaching Bayesian reasoning in less than two hours.* Manuscript submitted for publication.

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1995, August). *Encoding and representing relative letter frequencies: From availability to connections strengths.* Paper presented at the 15th Biannual Conference on Subjective Probability, Utility, and Decision Making, Jerusalem, Israel.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 48*(A), 257–279.

Sniezek, J. A., & Buckley, T. (1993). Decision errors made by individuals and groups. In N. J. Castellan (Ed.), *Individual and group decision making.* Hillsdale, NJ: Erlbaum.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232.

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

Varey, C. A., Mellers, B. A., & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 613–625.

Wänke, M., Schwarz, N., & Bless, H. (1995). The availability heuristic revisited: Experienced ease of retrieval in mundane frequency judgments. *Acta Psychologica, 89,* 83–90.

Watkins, M. J., & LeCompte, D. (1991). Inadequacy of recall as a basis for frequency knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 1161–1176.

White, P. A. (1985). *The availability heuristic and retrieval strategies in judgments of letter frequency.* Unpublished manuscript, University of Auckland.

White, P. A. (1991). Availability heuristic and judgments of letter frequency. *Perceptual and Motor Skills, 72,* 34.

Wimmer, H., & Hummer, P. (1990). How German-speaking first graders read and spell: Doubts on the importance of the logographic stage. *Applied Psycholinguistics, 11,* 349–368.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Received August 6, 1996
Revision received August 11, 1997
Accepted September 3, 1997 ■