Memory & Cognition
2009, 37 (4), 383-393
doi:10.3758/MC.37.4.383

On splitting and merging categories: A regression account of subadditivity

Klaus Fiedler, Christian Unkelbach, and Peter Freytag
University of Heidelberg, Heidelberg, Germany

Frequency judgments tend to be subadditive: A category's frequency is judged to be lower than the summed frequency of its subcategories. Thus, by splitting or merging categories, subjective frequencies increase or decrease, respectively. We offer an account of this phenomenon that is based on the statistical principle of regression. Because empirical information is never noise-free, high frequencies are underestimated, and low frequencies are overestimated. The underlying regression principle explains available evidence on subadditivity and allows novel predictions. The findings from two experiments supported predictions derived from the regression account of frequency estimates for split and merged categories: Subadditivity varied systematically as a function of the two parameters determining regression (extremity and reliability). More extreme frequencies and reduced reliability led to increased regression effects. Theoretical implications for subadditive judgments (of frequency, probability, and/or value) are discussed. Although other factors may contribute to subadditivity, their influence needs to exceed the baseline expected from the regression model alone.

Subjective quantities tend to be subadditive. That is, the subjectively experienced quantity of a compound or aggregate $s(A \vee B)$ is typically less than the sum of the segregated quantities $s(A) + s(B)$.[1] For example, winning \$100 once is worth less than winning \$50 twice. Completing some unpleasant work for 5 h on a single day usually causes less dissatisfaction than doing the unpleasant job 1 h on each of 5 work days. Driving 200 miles all at once is not twice as much, subjectively, as driving 2×100 miles. When aggregating quantities, the merged whole is often less than the sum of its additive parts. Conversely, when segregating quantities, splitting the whole into additive parts serves to increase the subjective quantity; the sum of all parts is more than the whole. The present article offers an explanation of this phenomenon that is based on the basic notion of statistical regression. In two experiments, we show that regression can explain the standard pattern of subadditivity, and we create new predictions that go beyond previous accounts of splitting and merging categories (Fiedler & Armbruster, 1994; Tversky & Koehler, 1994).

Scope and Origins of Subadditivity

The phenomenon of subadditivity in judgments and subjective experience is found for quantitative value (*how much?*) and probability (*how likely?*). The psychophysics of hedonic experience has been concerned with subadditive utilities, showing that positive and negative experiences can be enhanced by segregation and reduced by aggregation (Fernandez & Turk, 1992; Linville & Fischer, 1991; Mellers, 2000; Morewedge, Gilbert, Keysar, Berkovits, & Wilson, 2007). Likewise, research on subjective probabilities has revealed that the subjective likelihood $s(A \vee B)$ of a disjunctive category $(A \vee B)$ is lower than the summed subjective component probabilities $s(A) + s(B)$. This means that both the subjective value and the likelihood of an outcome can be increased by splitting (or *unpacking*) a superordinate category into several subordinate categories (Fiedler, 2002; Fiedler & Armbruster, 1994; Kruger & Evans, 2004; Rottenstreich & Tversky, 1997). Thus, thinking about several different causes of death, rather than death as an overall category, should render the topic more unpleasant and the possibility of dying more likely.

Cognitive underpinnings. Subadditivity is a natural result of universal psychological laws and processes. Most psychophysical functions have a decreasing slope.[2] The threshold for detecting increments of a quantity increases with the absolute level of the quantity (i.e., the Weber–Fechner law). Another well-established law that predicts subadditivity is the superiority of distributed learning over massed learning (Cull, 2000; Hintzman, 1969). This notion implies that segregating or distributing a learning experience in time and space should increase the resulting memory strength, and that this should, in turn, lead to inflated subsequent subjective frequency judgments. Retrieval operations, too, should be facilitated when splitting a category label into several subcategory labels increases the number of retrieval cues (Rottenstreich & Tversky, 1997).

Formal representation. One general account of subadditivity is called *support theory* (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994; White & Koehler,

K. Fiedler, kf@psychologie.uni-heidelberg.de

 © 2009 The Psychonomic Society, Inc.

2006); it provides an algebraic framework for the analysis of subadditive frequency and probability judgments. However, whereas support theory affords an elegant formal representation of the logical and psychophysical structure of aggregation and segregation, it is mute about *why* the support function assumed to underlie subjective judgments is nonadditive. Tackling this question was the purpose of the present study. We investigated subadditive frequency estimates of stimulus categories that were either split into subcategories or merged as a superordinate category. Splitting and merging a category should increase and decrease the subjective frequency estimates, respectively. We propose a model that draws on an incontestable property of all information transmission under uncertainty: the principle of *regression*. Our regression approach can be shown to predict the occurrence and the strength of merging and splitting effects on judgments and decisions. These effects will be shown to vary as a function of exactly those factors that determine the strength of regression—that is, reliability and extremity.

A Regression Account of Subadditive Judgments

Statistical regression is often treated as an artifact of empirical research, but it can serve as an explanatory construct in its own right (Fiedler, 1991, 1996; Moore & Small, 2007; Sedlmeier, 1999). Regression is a universal property of the empirical world, and it affords a well-founded basis for scientific explanations. Moreover, as the implications of regression continue to be neglected (Furby, 1973; Rulon, 1941), novel predictions and insights can be gained from exploring regression phenomena. In particular, a regression analysis of subadditive judgments offers a distinct integrative account of the quantitative illusions resulting from the splitting and merging of stimulus categories.

According to Furby (1973), a necessary and sufficient condition for regression is an imperfect correlation between two variables. In a probabilistic world, therefore, virtually all information processing must produce outputs that are regressive on the input. Large input quantities are underestimated in the judgmental output; small input quantities are overestimated. This regressive loss of systematic variance depends on two parameters: extremity and reliability. First, regression increases with the extremity of quantities. Extremely large (small) quantities are underestimated (overestimated) more than are less extreme quantities. Second, the resulting shrinkage of the differences between high and low stimulus frequencies increases with the noise of the information input and the unreliability of the judgment. The extremity and reliability parameters together determine the regression effect.

The standard parametric regression model. According to the parametric assumptions of statistics books, the precise way in which extremity and reliability determine regression is best described in deviation terms. Let Y represent the deviation of a variable X from its mean, M_X. Then the imperfect reproduction of this quantity regresses to $Y^* = Y \cdot r$, where r denotes the reliability. If $r = .8$, then Y^* shrinks to 80% of its original deviation from the mean. If $r = .5$, then it shrinks to half its original extremity. The shrinkage effect, which amounts to $Y \cdot (1 - r)$, is proportional to the extremity or deviation score, Y. Thus, the absolute degree of regressive shrinkage is stronger for extreme values than for moderate values in a distribution. It is this proportional shrinkage of extremity that justifies the phrase "regression to the mean."

A weaker nonparametric model. In reality, when the strong parametric assumption that X (just as Y) is normally distributed on a metric interval scale cannot be upheld, the $Y^* = Y \cdot r$ rule may not strictly apply, and regression may not move precisely toward the mean. Rather, a weaker nonparametric version of regression is sufficient for many theoretical purposes. It simply assumes that, for any set of stimulus objects {A, B, C, . . . , K}, the actually existing differences will be increasingly underrepresented in subjective judgments as the reliability decreases. With an increasing amount of noise, or *error*, in the system, the original differences will be more and more lost. Moreover, the absolute loss of systematic variance resulting from this regression of differences will increase with the extremity of the stimuli being judged.

Such a weaker formulation of the regression principle is sufficient for generating a set of empirically testable predictions that follow neither from support theory per se nor from the other theoretical principles mentioned at the outset (i.e., psychophysical functions, diminishing marginal units, and spaced learning). Rather, these predictions reflect the distinct theoretical potential of the law of regression.

A regression account of category-split effects. To develop the predictions, we used the same task as in the experiments reported below. The task involves estimating the frequencies with which four stimulus categories (A, B, C, D; viz., exemplars of four types of insects) have occurred in a stimulus list. In one condition, the actual frequencies are $f(\mathrm{A}) = 4$, $f(\mathrm{B}) = 10$, $f(\mathrm{C}) = 16$, and $f(\mathrm{D}) = 22$. By subtracting these objective frequencies from the respective frequency estimates $j(\mathrm{A})$, $j(\mathrm{B})$, $j(\mathrm{C})$, and $j(\mathrm{D})$, we can measure the sign and size of judgment inaccuracies.[3]

The basic regression effect (see under "Nonsplit" in Table 1) shows that subjective judgments underestimate actual frequency differences. A and B should be overestimated, whereas C and D should be underestimated. The expected variance in the four quantities decreases relative to the original variance. If these inaccuracies reflect regression, the overall amount of shrinkage should increase with experimental manipulations of reliability. Higher reliability (i.e., less uncertainty) should result in smaller deviations of $j - f$, whereas lower reliability should increase the deviations. At a given level of reliability, regression should be more pronounced for extreme (A, D) than for moderate (B, C) categories.

Now consider what happens after splitting or unpacking one or more categories into subcategories (see under "Split" in Table 1). For example, when the least frequent category, A, is split into two subcategories, A_1 and A_2, this results in even more extremely small subfrequencies. Thus, by presenting two different versions of A (e.g., the same insect in different colors), $f(\mathrm{A}) = 4$ is decomposed into $f(\mathrm{A}_1) = 2$ plus $f(\mathrm{A}_2) = 2$. When judges are then asked

Table 1
Theoretically Expected Frequency Judgments (j_{exp}) of Four Categories (A, B, C, D), As a Function of Different Category Frequencies (f) and Two Levels of Assumed Reliability (r)

	Nonsplit				Split				
	f	$e = f - 13$	$e^* = e \cdot r$	$j_{exp} = 13 + e^*$	f	$e = f - 13$	$e^* = e \cdot r$	$j_{exp} = 13 + e^*$	$jj_{exp} = 2j_{exp}$
					$r = .5$				
A	4	−9	−4.5	8.5	2 + 2	(−11) + (−11)	(−5.5) + (−5.5)	7.5 + 7.5	15
B	10	−3	−1.5	11.5	5 + 5	(−8) + (−8)	(−4) + (−4)	9 + 9	18
C	16	+3	+1.5	14.5	8 + 8	(−5) + (−5)	(−2.5) + (−2.5)	10.5 + 10.5	21
D	22	+9	+4.5	17.5	11 + 11	(−2) + (−2)	(−1) + (−1)	12 + 12	24
					$r = .67$				
A	4	−9	−6	7	2 + 2	(−11) + (−11)	(−22/3) + (−22/3)	5.67 + 5.67	11.33
B	10	−3	−2	11	5 + 5	(−8) + (−8)	(−16/3) + (−16/3)	7.67 + 7.67	15.33
C	16	+3	+2	15	8 + 8	(−5) + (−5)	(−10/3) + (−10/3)	9.67 + 9.67	19.33
D	22	+9	+6	19	11 + 11	(−2) + (−2)	(−4/3) + (−4/3)	11.67 + 11.67	23.33

Note—Subtracting the mean frequency from the original frequency f yields a deviation or extremity measure $e = f - 13$; that is, mean(A, B, C, D) = 13. Multiplying e by r gives the extremity e^* after regression. The expected judgment is obtained by adding the mean again: $j_{exp} = e^* + 13$. In the case of split categories, the sum of both subcategory judgments is $jj_{exp} = 2 \cdot j_{exp}$.

to provide separate estimates—$j(A_1)$ and $j(A_2)$—the sum of these two judgments of a split category should result in a particularly strong overestimation, because a regressive overestimation of extremely small values of 2 should be doubled. As a general rule, because splitting always results in lower subfrequencies, this will increase the relative overestimation error, thus producing subadditivity [$j(A_1) + j(A_2) > j(A)$].

Again, if subadditivity reflects regression proper, the category-split effect should decrease as the reliability of the frequency estimation increases (e.g., from $r = .5$ to $r = .67$, as is shown in Table 1). The influence of extremity on the degree of category-split effects is especially pronounced for small categories (e.g., A), because both the basic regression effect and the split effect increase the judges' subjective estimates, thus leading to a particularly strong overestimation. For a large category (e.g., D), though, a regressive underestimation will counteract the overestimation resulting from the split, thus resulting in a weaker deviation of judgments from the actually presented stimulus frequency.

An intriguing implication of the (parametric) regression model is that the theoretically predicted difference between split and nonsplit judgments—the category-split effect—remains constant across categories A, B, C, and D, as illustrated in Table 1 for two levels of reliability (i.e., $r = .5$ and $r = .67$, producing constant split effects of +6.5 and +4.33, respectively, under parametric assumptions). For example, assuming $r = .5$, the deviation or extremity score for $f(A) = 4$ (in the first row of Table 1) is −9 (i.e., $e = 4 - 13$); the expected judgment of Category A before the split should thus be $13 + r(-9) = 8.5 = j_{exp}(A)$. By comparison, splitting A into $f(A_1) = 2$ and $f(A_2) = 2$ yields two deviation scores of −11, which leads to two regressive judgments of $(-11/2) + 13 = 7.5$, totaling an expected judgment sum of $jj_{exp}(A_1 + A_2) = 15$. Thus, the category-split effect amounts to a positive increase of $15 - 8.5 = +6.5$. As is evident from the columns labeled j_{exp} and jj_{exp} in Table 1, this increment of +6.5 is constant across all categories. When the reliability increases from $r = .50$ to $r = .67$, the constant increment shrinks proportionally to 4.33. Deviations from this constant increment can be expected within the nonparametric model if internally represented frequencies do not form an interval scale or are not perfectly translated onto numerical judgment scales.

Thus, whereas the reliability parameter should influence the basic regression effect (nonsplit categories) and the category-split effect (split categories) equally, the extremity parameter should affect the basic regression effect only. This is because the split effect and the basic regression effect should both contribute to overestimations of small categories (e.g., A). For large categories (e.g., D), though, the absolute size of the split effect is larger, but the basic regression effect works in the opposite direction, resulting in the same net effect.

Experimental predictions. The regression model leads to the following predictions about the moderating and accompanying conditions of subadditivity, which will be tested in Experiments 1 and 2. First, subjective frequency estimates of the nonsplit categories should exhibit the basic regression effect, which consists in the overestimation of small frequencies and the underestimation of large frequencies. Second, this basic regression effect should increase with the extremity of the categories being judged. Third, cognitive load should amplify the basic regression pattern, due to decreased reliability. Fourth, the summed judgments of split categories should exceed the aggregate judgment of the original category. This category-split effect should also increase with cognitive load, but it is not predicted to depend on extremity. Two parameters of the regression model suffice to explain the impact of the experimental variables on both the basic regression effect and the category split effect, as explained in Table 1, and the strength of both effects should be correlated across judges. Neither the algebraic beauty of support theory nor the concave form of psychophysical functions can in and of itself generate this refined pattern of predictions. No other approach can particularly explain the double influence of extremity and reliability on basic frequency judgments (i.e., reduction of systematic variance) and the amplifying role of unreliability in category-split effects (i.e., subadditive judgments).

In Experiment 1, we tested the foregoing predictions. Participants first observed a stimulus series and then judged the frequency of occurrence of different stimulus categories. They observed a list of 52 insects and then judged the frequency of four categories of insects: A, B, C, and D (cf. Table 1). To manipulate reliability, we used a secondary task to increase cognitive load during observation. To observe the independent effects of category split and regression, either the extreme (very frequent [A] vs. very infrequent [D]) insect categories or the moderate (B vs. C) insect categories were split at judgment into two subcategories.

EXPERIMENT 1

Method

Participants and Design. Forty students of the University of Heidelberg (32 female; mean age, 24.45 years) participated for payment of €4 or partial course credit. They were randomly assigned to one of the four experimental groups created by the orthogonal combination of the two between-participants manipulations (load vs. no-load, extreme-split vs. moderate-split). For all participants, the actual presentation frequencies of A, B, C, and D were 4, 10, 16, and 22, respectively. The between-participants manipulation of main interest was cognitive load (load vs. no-load), as operationalized by a secondary task during stimulus presentation. The second manipulation referred to the subset of categories that were split (extreme vs. moderate). For half of the participants, the extreme categories (A, D) were split for judgments; for the other half, the moderate categories (B, C) were split.

Materials and Procedure. We used butterflies as objects in the categories; the presentation frequencies were 4, 10, 16, and 22 for Categories A, B, C, and D, respectively, as is indicated in Table 1. The butterflies were made of simple geometric forms. Category membership was determined by the form of the wings: A butterfly's wings consisted of quadrangles, triangles, circles, or half ellipsoids. The assignment of wing form to category (A, B, C, D) was randomly determined for each participant, in order to avoid confounding form and category frequency. Butterflies in the nonsplit condition always appeared in crème white. When a category was split, one subtype of butterflies was pale blue, and the other was pale yellow. The contrast between the colors used to represent stimuli from split and nonsplit categories was too weak to affect frequency estimates. A separate test using 16 participants confirmed that pale blue or pale yellow butterflies presented at the same rate did not receive different frequency estimates. The reason for presenting split categories in distinct colors was that, logically, one additional feature is required to introduce any subcategory distinction.[4] We used a Microsoft Visual Basic program to present these stimuli and record the dependent variables.

The experimental sessions included up to 4 participants. Upon arrival, each participant was greeted and seated in a cubicle with a PC used for the computerized experiment. The participants were told that they should take the role of a researcher who wants to know how often four different types of butterflies (the four categories) occur in a given area. They were then presented with examples of butterflies and were informed that some categories included subspecies that still belonged to the same type but differed in wing color.

In the load condition, an additional instruction was given. There were supposedly snakes in the observation area, and if one of these snakes appeared, the participant would have to double-click on it to fend it off and avoid a snake bite. Finally, the task was repeated: The participant was to observe this area (the screen) and try to remember how often each butterfly type occurs. When the participants had no further questions, the observation phase started immediately.

Butterflies appeared on-screen for 2 sec at a random location in front of a photograph of a lake; the full butterfly was always visible. The width of each butterfly was set to 25% of the screen width, and the height was set to 75% of the screen width. There was an interstimulus interval of 750 msec between the appearances of each butterfly. In the no-load condition, participants observed the 52 occurrences and continued with the rating phase. The load condition was realized as follows: A second, smaller picture appeared on-screen at a random location 750 msec after a butterfly appeared. In half of the trials, it was a drawing of a bird or bee. No reaction was required for these stimuli. In the other half, it was a drawing of a snake, and the participants had to double-click on this picture. If they failed to do so within 1 sec, the program enlarged the snake picture to full-screen size until it was double-clicked. All other timers were on halt until then. Thus, participants in the load condition had to monitor a second set of pictures and react accordingly if a drawing of a snake appeared.

In the judgment phase, the participants were presented with an exemplar of each type of butterfly they had seen (six types: two exemplars from the nonsplit categories and two exemplars from the split categories). The participants answered three questions for each type: "How many times did this kind of butterfly occur?" "What is the probability of occurrence in percent (0–100)?" "How aesthetically pleasing is this butterfly?" The rating order was newly randomized for each participant, and the questions were presented consecutively for each type. After finishing these ratings, the participants were fully debriefed, thanked, and paid by the experimenter. The experimental sessions ranged from 10 to 15 min.

Results

The validity of the reported findings is contingent on the experimental manipulations' actually inducing the intended independent variables, as we assume they do. Whereas the level and extremity of the category frequencies could be controlled objectively, we use a manipulation check in order to demonstrate that the cognitive load treatment actually affects the unreliability of the judgment process. For a simple and straightforward measure of unreliability, we analyzed the unsigned inaccuracies (i.e., each judge's mean differences between all six judgments and their associated correct frequencies, disregarding the sign of difference) as a function of cognitive load. As we expected, this measure of unreliability was significantly higher in the load group ($M = 56.42$) than in the no-load group ($M = 28.81$) [$t(38) = 2.64, p < .05$], supporting the assumption that load reduced the reliability as intended. Convergent results were obtained when the correlations between subjective and objective frequencies, rather than the absolute discrepancy, were used as an alternative measure of unreliability. The average correlation amounted to .443 in the load condition, as compared with .693 in the no-load condition [$t(38) = 2.08, p < .05$]. In any analysis, unsystematic error increased with load.

Prior to the data analysis, judgments were normalized, so that the sum of each judge's estimates equaled 52, the actual total stimulus frequency.[5] Double estimates for split categories were totaled. From each judge's four category-level estimates (for A, B, C, D, whether split or not), we calculated additive deviation scores [ADS; j(category) − f(category)]. Overestimation was evident in positive ADS > 0; ADS < 0 implied underestimation.[6]

Basic regression effect. The upper part of Table 2 gives the average (normalized) estimates for the four categories per experimental conditions, along with the corresponding ADS measures, averaging across only the

Table 2
Actual Frequencies, *f*, Normalized Mean Judged Frequencies, *j*, and Mean Additive Deviation Scores (ADS) Obtained in Experiment 1, As a Function of Experimental Condition

		No Load			Load		
		Normalized *j*			Normalized *j*		
Category	Presented *f*	*M*	*SD*	ADS	*M*	*SD*	ADS
Estimates of Nonsplit Categories							
A	4	4.35	1.70	0.35	6.62	3.47	2.62
B	10	8.89	3.66	−1.11	6.77	2.83	−3.23
C	16	12.22	4.83	−3.78	9.94	3.83	−6.06
D	22	18.59	5.67	−3.41	12.30	4.94	−9.70
Estimates of Split Categories							
A	4	8.74	4.49	4.74	10.33	4.24	6.33
B	10	14.73	3.85	4.73	17.58	3.68	7.58
C	16	14.33	3.67	−1.67	15.51	4.17	−0.49
D	22	20.15	8.21	0.15	24.96	7.60	2.96

judges who saw the respective categories in the nonsplit condition. Note that the data for the extreme (A, D) and moderate (B, C) categories, as well as for the load and no-load conditions, stem from different participant groups. The basic regression effect is clearly borne out in these judgments of nonsplit categories. As the upper row of charts in Figure 1 illustrates, estimates of small frequencies (open squares) were above the actual frequencies (the straight line), whereas large frequencies were underestimated (estimates below the straight line). As predicted, regression was stronger for extreme (A, D) than for moderate (B, C) categories, and it increased under cognitive load. This is clearly visible when one compares the middle and right charts in the upper row in Figure 1: Regressive over- and underestimation of small and large frequencies, respectively, are pronounced under cognitive load. This amplifying influence of cognitive load is mainly due to the extreme condition. This is exactly the theoretically predicted pattern.

For a statistical test, a basic regression score was defined as the difference between each judge's two nonsplit ADS scores—that is, the ADS for a low-frequency category (A or B, depending on whether extreme or moderate categories were in the nonsplit condition) minus the ADS for a high-frequency category (C or D). The higher this difference score, the stronger the individual strength of basic regression on nonsplit categories.

Across all participants, this basic regression score was clearly positive [$M = 5.15$, $SD = 6.69$; $t(39) = 5.10$, $p < .001$]. Moreover, a two-factorial ANOVA with extremity of the category being judged (extreme vs. moderate) and cognitive load (load vs. no load) as between-participants factors yielded the predicted main effects of extremity [$F(1,36) = 10.15$, $p < .01$], a main effect of cognitive load [$F(1,36) = 6.90$, $p < .05$], and a significant interaction [$F(1,36) = 6.39$, $p < .05$]. As predicted, stronger regression was obtained for extreme ($M = 7.81$) than for moderate ($M = 2.74$) categories and for the load ($M = 7.32$) than for the no-load ($M = 3.18$) condition, and the load effect was stronger for estimates of extreme ($M = 12.32$ vs. 3.76) than for those of moderate ($M = 2.83$ vs. 2.66) categories.

Category-split effects. A comparison of the upper and lower parts of Table 2 reveals that the summed estimates of segregated categories were generally higher than the corresponding aggregate-category estimates. For a graphical illustration of this category-split effect, the bottom row of charts in Figure 1 repeats the curves for nonsplit categories (open squares; the basic regression effect), along with the corresponding curves for split categories (filled squares).

Recall that the judgments in the split and nonsplit conditions always stem from different participants. We can thus conduct only four between-participants comparisons for judgments of split versus nonsplit categories, of which two pairs (A, D and B, C) rely on the same groups. In general, the summed ADS scores for the split condition were significantly higher than those for the corresponding nonsplit conditions (cf. upper and lower parts of Table 2). Pooling across load conditions, the *t* statistics for categories A, B, C, and D are $t(38) = 3.30$, $p < .005$; $t(38) = 6.75$, $p < .001$; $t(38) = 2.71$, $p < .01$; and $t(38) = 3.36$, $p < .001$, respectively.

For an overall test of the split effect, we calculated an individual split score as the difference between each judge's average ADS on the two split categories minus his or her average ADS for the nonsplit categories, regardless of whether extreme or moderate categories were split. Note that the actual total frequency in both extremity conditions always summed to $26 = (10 + 16) = (4 + 22)$. If category-split effects reflect the same uncertainty or noise as the basic regression effect, then the individual split scores should correlate with the same individuals' basic ADS regression scores. Indeed, the correlation amounts to $r = .38$, $p < .02$.

When extremity (0 = *moderate categories split*, 1 = *extreme categories split*) and load (0 = *no load*, 1 = *load*) were included together with the basic regression score in a regression analysis using the ADS split score as a criterion, the multiple correlation increased [$R = .52$; $F(3,37) = 6.84$, $p < .001$]. The basic regression score remains a significant predictor when the other predictors are controlled for [$\beta = .46$; $t(37) = 2.77$, $p < .01$]. Extremity

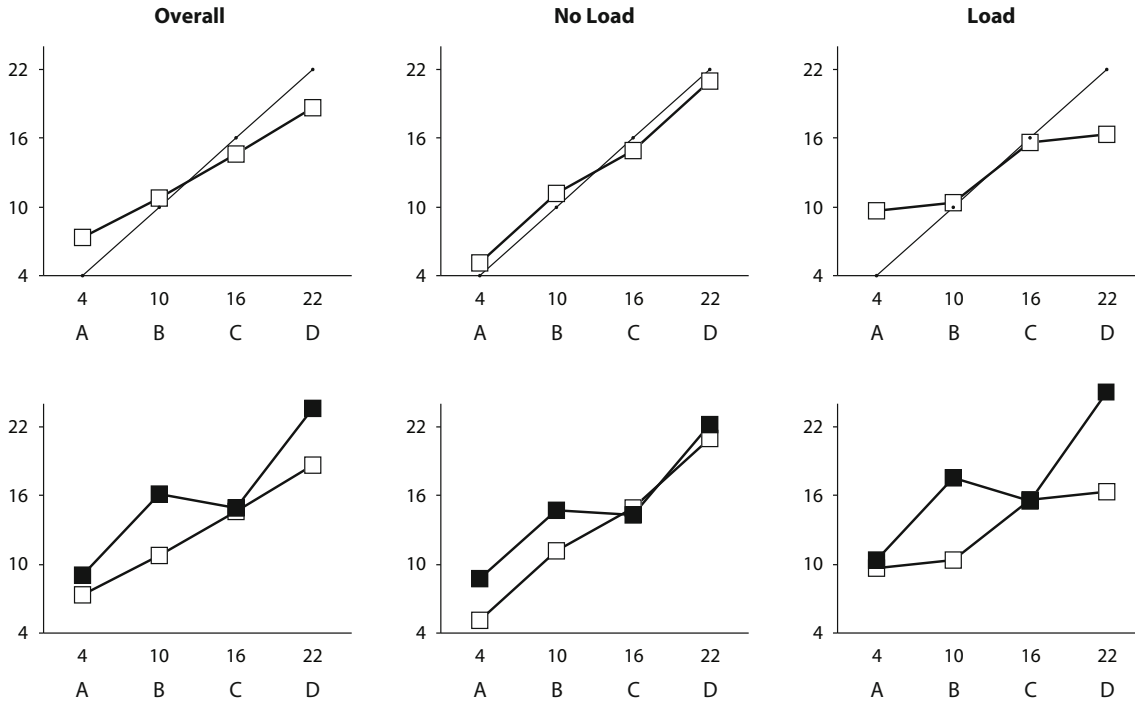

Figure 1. Mean estimates of nonsplit categories (open squares) and split categories (filled squares), as a function of actually presented stimulus frequencies (diagonal) and of the cognitive-load manipulation in Experiment 1.

(i.e., whether split categories are extreme or not) is also a significant predictor in this analysis [$\beta = .33$; $t(37) = 2.13$, $p < .05$], but cognitive load is not [$\beta = .16$; $t(37) = 1.07$, n.s.]. Cognitive load did not exert an independent influence. Load is theoretically supposed to be mediated by regression, so it makes sense that load receives no independent weight, above and beyond regression. Consistent with this interpretation, load does become a significant predictor if the basic regression score is omitted [$\beta = .31$; $t(38) = 2.04$, $p < .05$].

It should be noted that extremity (i.e., whether split categories are extreme or moderate) did contribute to the prediction, although this cannot be derived from the regression model, which implies a constant split effect of extreme and moderate categories. Therefore, the enhanced split effect of extreme categories must reflect something other than regression proper. In the absence of a cogent explanation, we tend to attribute this unpredicted finding to the differential use of moderate and extreme regions of the numerical response scale.

Discussion

The obtained differences in subadditive judgment biases can all be explained by our nonparametric regression model. First, the estimates of aggregate categories reflect the basic regression effect that is typical of frequency judgment studies. Second, the degree of this regression effect depends on the two crucial determinants of regression: reliability and extremity. Third, the inflation of subjective category size that results from splitting categories into smaller subcategories also increases with the decreased reliability that is due to cognitive load. Finally, the interpersonal variation in the strength of the category-split effect correlates in a distinct fashion with the same individuals' basic regression effects. No other model or theoretical conception that is commonly used to explain subadditive judgment biases can account for this precise pattern, derived in all detail from the regression account.

EXPERIMENT 2

To further substantiate and corroborate this account, we conducted a second experiment, using a different distribution of stimulus frequencies. The frequencies used for Categories A, B, C, and D were reduced from 4, 10, 16, and 22, respectively, in Experiment 1 to 10, 10, 16, and 16, respectively, in Experiment 2. This left the total stimulus frequency at 52 and the subtotals for A + D and B + C at 26. Given only one frequency difference of 6, the range of the distribution is thus greatly reduced to one third.

The purpose of this second experiment was not only to replicate and extend the Experiment 1 findings to a task setting characterized by smaller frequency differences. The newly introduced constraints in the stimulus setup—$f(A) = f(B) = 10$, and $f(C) = f(D) = 16$—also afford a direct test of predictions that could not be tested in the previous experiment. First, as was already mentioned, we can test for regression and subadditivity with

less extreme frequency differences. Second, the inclusion of two equally infrequent and two equally frequent categories allows us to directly compare split and nonsplit categories within participants. Third, the frequency-range reduction offers a suitable between-participants test of the extremity assumption (when we compare the results from both experiments). And, finally, we can rule out the numerical-response-scaling account of the regressive pattern of findings.

Theoretically, our regression account is conceived as a genuine cognitive-process model, rather than as a superficial or artificial influence of the judgment task. Thus, it is assumed that noise in a transmitting system causes information loss, consistent with the operation of noise parameters in connectionist memory models (Dougherty, Gettys, & Ogden, 1999; Fiedler, 1996). Alternatively, however, one might assume that the regressive judgment output simply reflects the scaling of internally represented judgments, which may not themselves accord to regression law, onto a numerical response scale. Judges may for some reason not use the full range of the response scale, due either to a central tendency that reflects a lack of confidence or to strategic concerns (e.g., to leave the extreme parts of the response scale open for even more extreme items; see Haubensak, 1992; Unkelbach & Memmert, 2008).

If the obtained pattern of regressive judgments merely reflects the mapping of internally represented frequencies onto a numerical response scale, rather than a genuine influence on the internal representation itself, then the drastic reduction of the frequency range should alter the translation rule. In Experiment 2, the same two frequencies, 10 and 16, should be mapped onto more extreme locations of the numerical scale than in Experiment 1, in which 10 and 16 were moderate levels, rather than endpoints of the frequency range. If, however, regressive judgments reflect information loss in frequency learning and the mapping on a cardinal frequency scale is relatively accurate (cf. Gigerenzer & Hoffrage, 1995), then the results for the same two frequencies from both experiments should converge. Moreover, since 10 and 16 represent only moderate, nonextreme positions, the basic regression effect should be weaker. The category-split effect should be similarly strong as in Experiment 1, because the size of this effect should not (or should only weakly) depend on extremity (cf. Table 1). Finally, the generally weaker regression effect resulting from the marked reduction of extremity cannot be expected to correlate strongly with the category-split effect.

Method

Participants and Design. Forty-one students (32 female; mean age, 26.17 years) of the University of Heidelberg, who were recruited from the same pool as in the first study, participated for payment of €4. They were randomly assigned to either the load or the no-load condition. For all participants, we varied the frequencies—10, 10, 16, and 16—for the categories A, B, C, and D, respectively. Furthermore, one of the high-frequency (16) categories and one of the low-frequency (10) categories were split. Technically, the design involved the same second between-participants factors as in Experiment 1—namely, whether Categories A and D or B and C were split, equivalent to the moderate versus extreme manipulation of Experiment 1. However, the factor is inconsequential here, because the frequencies for A and B are identical, as are those for C and D.

Materials and Procedure. Upon arrival, the participants were greeted by an experimenter, who followed the same procedure as in Experiment 1. Exactly the same computer program controlled the entire experiment, using the same instruction text, stimulus display, and dependent measures as in Experiment 1. The only difference resulted from the use of different stimulus frequencies. Categories A, B, C, and D were presented 10, 10, 16, and 16 times, respectively. Splitting infrequent and frequent categories resulted in subcategory frequencies of 5, 5 and 8, 8, respectively. For approximately one half of the participants, Categories A and D were split; for the remaining half, Categories B and C were split. Thus, two parallel pairs of judgments for infrequent and frequent categories were obtained within each participant: one for the nonsplit category and one for the split category. Upon completing the study, participants were fully debriefed, thanked, and paid. The experimental sessions included up to 4 participants and lasted between 10 and 15 min.

Results and Discussion

Basic regression effect. Table 3 shows that the basic regression effect is still visible, but clearly reduced, after the reduction of the frequencies in Experiment 2.[7] Still, infrequent categories (A, B) were overestimated, whereas frequent categories (C, D) were underestimated. The upper

Table 3
Actual Frequencies, *f*, Normalized Mean Judged Frequencies, *j*, and Mean Additive Deviation Scores (ADS) Obtained in Experiment 2, As a Function of Experimental Condition

		No Load			Load		
		Normalized *j*			Normalized *j*		
Category	Presented *f*	*M*	*SD*	ADS	*M*	*SD*	ADS
Estimates of Nonsplit Categories							
A	10	10.59	5.22	0.59	9.34	4.06	−0.66
B	10	8.91	3.14	−1.09	8.62	1.69	−1.38
C	16	12.33	1.90	−3.67	11.21	3.94	−4.79
D	16	15.65	6.06	−0.35	11.59	2.87	−4.41
Estimates of Split Categories							
A	10	12.11	2.09	2.11	13.03	4.43	3.03
B	10	11.16	4.67	1.16	15.09	5.12	5.09
C	16	14.60	5.64	−1.40	15.99	4.88	−0.01
D	16	18.65	3.68	2.65	19.14	5.54	3.14

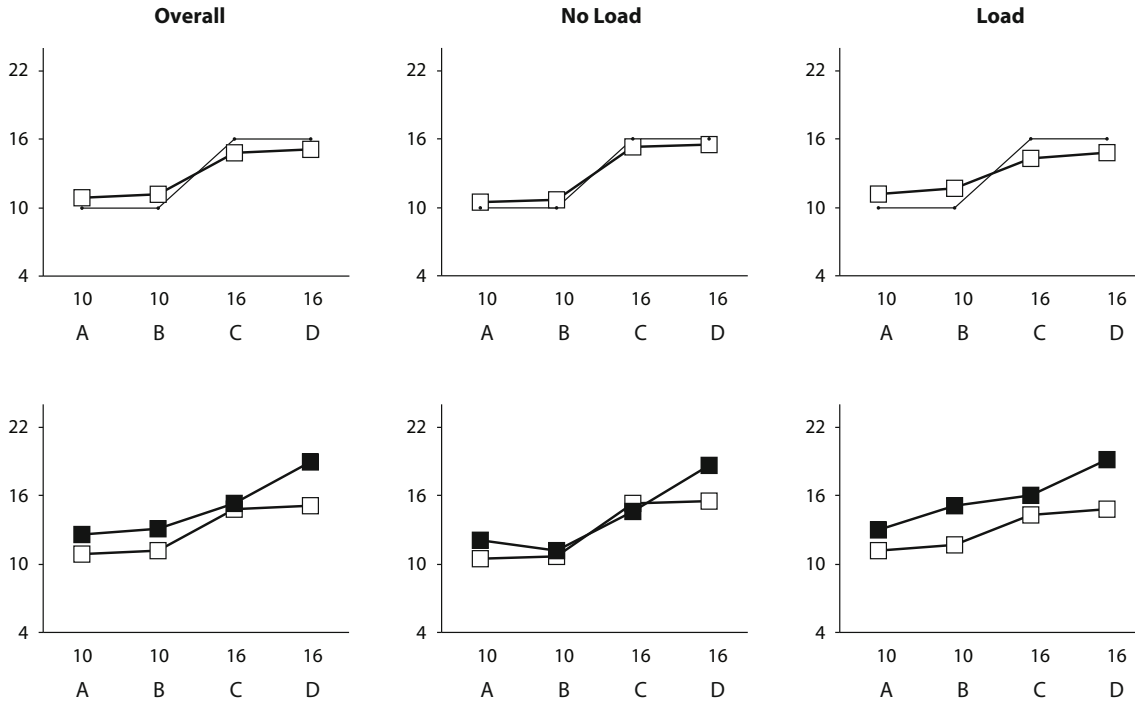

Figure 2. Mean estimates of nonsplit categories (open squares) and split categories (filled squares), as a function of actually presented stimulus frequencies (thin lines) and of the cognitive-load manipulation in Experiment 2.

charts in Figure 2 plot the nonsplit categories against the actual frequencies. Apparently, equally frequent categories received similar ratings, as is evident from the step curves in the upper charts in Figure 2. Across all judges, though, the mean basic regression score (as was derived in Experiment 1, the ADS difference between small and large nonsplit category estimates) is significantly above zero [$M = 2.69$, $SD = 5.18$; $t(40) = 3.32$, $p < .001$].

When the same ANOVA design used in Experiment 1 was applied to the basic regression effect in Experiment 2—including the cognitive-load factor and the A, D versus B, C contrast—only the main effect of cognitive load was significant [$F(1,37) = 3.02$, $p < .05$, one-tailed]. The basic regression effect increased from the no-load ($M = 2.26$) to the load ($M = 5.63$) condition. Since the other manipulation (i.e., whether Categories A and D or B and C were split) was only a technical control factor in Experiment 2, it is not surprising that this factor did not exert any effect.

However, although the extremity effect cannot be tested within Experiment 2, a direct comparison between Experiments 1 and 2—which used exactly the same methods and procedures, as well as the same participant pool—allows for a between-participants test of the extremity effect. Indeed, the exclusion of extreme stimulus frequencies in Experiment 2 resulted in lower basic regression scores ($M = 2.69$, $SD = 5.18$) than those found in the moderate-split condition of Experiment 1 ($M = 7.81$, $SD = 7.68$), which included more extreme categories [$t(58) = 3.04$, $p < .01$]. However, no difference was obtained between Experiment 2's findings and the extreme-split condition results of Experiment 1 ($M = 2.74$, $SD = 3.70$), in which the basic regression effect referred to the same moderate category frequencies (10, 16) as in Experiment 2 [$t(60) = 0.042$, n.s.].

These highly comparable results corroborate the contention that participants in Experiment 2, in which the extreme frequencies (4, 22) were excluded, did not map the smaller range of stimulus frequencies (10, 16) onto a broader range of the judgment scale than did Experiment 1's participants, who had to reserve extreme scale positions for more extreme frequencies. The between-experiments test of the extremity effect cannot, therefore, be attributed to different uses of the numerical judgment scale.

Category-split effect. The design of Experiment 2 allowed a more refined test of the category-split effect because of its orthogonal within-participants manipulation of category frequency (10 vs. 16) and category split. The basic split effect is visible in the lower charts of Figure 2, in which the split (filled squares) and nonsplit (open squares) frequency estimates are compared. As was predicted from the regression model, a within-participants ANOVA yielded two main effects [category frequency, $F(1,40) = 9.47$, $p < .01$; category split, $F(1,40) = 15.69$, $p < .001$], but no interaction [$F(1,40) = 0.05$]. Table 3 and Figure 2 show that large frequencies were underestimated, whereas small frequencies were overestimated.

ADS scores for split categories exceeded those for nonsplit categories. When cognitive load was included as a between-participants factor, load interacted with the split effect [$F(1,39) = 2.96$, $p < .05$, one-sided], but not with category frequencies [$F(1,39) = 1.24$, n.s.]. Thus, the reduction of the frequency range, which already caused a significant reduction of the basic regression effect, also reduced the strength of the category-split effect. The overall pattern is again completely consistent with the predictions of the regression model.

A comparison of the split scores (i.e., each judge's average judgment difference for split minus nonsplit categories) obtained in both experiments again showed that the split effect in Experiment 2 ($M = 4.00$, $SD = 6.46$) tended to be lower than in the extreme-split condition of Experiment 1 ($M = 6.99$, $SD = 7.38$) [$t(60) = 1.64$, $p < .06$], but of similar strength to that in the moderate-split condition of Experiment 1 ($M = 4.97$, $SD = 6.32$) [$t(60) = 0.54$, n.s.].

GENERAL DISCUSSION

Both theoretical analyses and empirical results corroborate the notion that statistical regression alone affords a sufficient account of subadditive frequency judgments in the context of a category-split task. Because the correlation between subjective estimates and the objective presentation frequencies of different stimulus categories is imperfect, judgments were regressive. Large frequencies were underestimated, whereas small frequencies were overestimated. Moreover, as was predicted by the regression model, the degree of regression increased with cognitive load (supposedly because load reduces judgment reliability) and with the extremity of frequency to be judged. This basic regression phenomenon, which reflects an incontestable property of the empirical world, could then be expanded to provide a sufficient condition for subadditive category-split effects, too. Because splitting a category produces smaller subcategory frequencies, regression implies that each subcategory will be overestimated, relative to the nonsplit category. This category-split effect again tended to be enhanced under cognitive load, and it was correlated across participants with the basic regression effect. Although not predicted by a parametric regression model, subadditivity also tended to increase slightly with extremity, which is however tolerated by a nonparametric formulation of regression. Moreover, when the frequency range was reduced to one third, the regression effect was reduced accordingly, thus ruling out an alternative explanation in terms of the mapping of frequency estimates onto a constant range of the response scale.

Although our approach is not bound by the quantitative predictions of a parametric regression model, such a model happens to provide a good approximation of the reported judgment data (as summarized in Figures 1 and 2). However, we continue to refrain from too strong a formulation of the regression account. We also refrain from assuming that regression can explain the entire variance of the judgment illusion, and equally so for different kinds of subadditivity (cf. Rottenstreich & Tversky's [1997] distinction of implicit from explicit subadditivity). Just as every theory—in psychology, as in other disciplines—that postulates an impact of X on Y cannot prevent other, non-X causes from influencing Y as well, we do not exclude the possibility that factors independent of extremity and reliability (i.e., the determinants of regression) may also influence the occurrence and degree of subadditivity.

For example, the similarity structure of subcategories (Rottenstreich & Tversky, 1997) or their typicality (Sloman, Rottenstreich, Wisniewski, Hadjichristidis, & Fox, 2004) may influence the memory retrieval process in ways that cannot be sensibly reduced to reliability and extremity. Also, the kinds of central tendencies or polarization tendencies on the judgment scale that we have controlled in the present research may influence final judgments under different conditions. However, although regression is not a necessary cause of all manifestations of subadditivity, we have shown that regression alone provides a sufficient condition for the joint occurrence of a basic regression effect and a subadditive category-split effect.

Even though the relationship between unpacking or category-split effects and statistical regression has been noted earlier (Fiedler, 2002; Fiedler & Armbruster, 1994; Rottenstreich & Tversky, 1997), no previous research has spelled out the implications of a mere regression account systematically. The present study is the first to analyze basic regression tendencies and subadditive judgment biases as a function of the two constituents of regression: extremity and reliability. Nevertheless, the more general notion that ordinary regression is at the heart of judgment biases receives support from several prior studies, including those that have dealt with overconfidence (Erev, Wallsten, & Budescu, 1994), consensus judgments (Weaver, Garcia, Schwarz, & Miller, 2007), probability judgments (See, Fox, & Rottenstreich, 2006), and personality judgments (Fiedler & Walther, 2004).

Treating regression as an explanatory construct, rather than as a statistical artifact, has important theoretical and practical implications. Theoretically, regression applies to all judgments of imperfect reliability, not just frequency and probability judgments, the home domain of support theory (Tversky & Koehler, 1994). Evidence for subadditive judgments of hedonic value corroborates this contention (e.g., Mellers, 2000; Morewedge et al., 2007). Indeed, the aesthetic-quality ratings that we collected after the frequency estimates exhibit the very same pattern of biases as the subjective frequency judgments, highlighting regression effects and category-split effects on both frequentistic and hedonic or aesthetic judgments. These additional findings, which will be published in a separate article, provide strong evidence for exposure effects (Bornstein & D'Agostino, 1994) originating in illusory, rather than in actually presented, frequencies.

With regard to the cognitive processes and boundary conditions leading to subadditive judgment biases, a number of interesting and testable hypotheses can be derived from the present approach. The cognitive origins of uncertainty or unreliability that amplify regression effects are manifold, including memory load, stimulus sample

size, encoding conditions, familiarity with the stimulus materials, ease of retrieval, or calibration to the response scale. In a similar vein, the extremity of the quantity to be judged depends on the nature of the judgment problem, the partitioning of the category system, and the framing and scaling of the quantity to be judged. The regression account is not restrictive regarding the origins of regression; it suggests that many heterogeneous phenomena should converge in the same mediation process, which is based on variation in reliability and extremity.

At the practical level, all kinds of treatments that increase reliability and reduce the extremity of stimulus quantities suggest themselves as potential means of debiasing. Consistent with this prediction, regressive tendencies have been shown to decrease with increasing subjective confidence (See et al., 2006), increasing sample size (Fiedler, 1996), and improved encodability of the stimulus materials (Fiedler & Armbruster, 1994).

We believe that this tension between a broad domain and a distinct mediating process highlights the theoretical fertility and the explanatory power of this account. Even when regression alone does not fully explain a judgment phenomenon, the expected degree of regression should be controlled as a baseline to be considered in the interpretation of other sources of judgment bias.

AUTHOR NOTE

The research underlying this article was supported by a grant from the Deutsche Forschungsgemeinschaft to the first author and a research scholarship by the Humboldt Foundation to the second author. Correspondence concerning this article should be addressed to K. Fiedler, Psychologisches Institut, Universität Heidelberg, Hauptstrasse 47-51, 69117 Heidelberg, Germany (e-mail: kf@psychologie.uni-heidelberg.de).

REFERENCES

Bornstein, R. F., & D'Agostino, P. R. (1994). The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attributional model of the mere exposure effect. *Social Cognition*, **12**, 103-128.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, **14**, 215-235. doi:10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

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.

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.

Fernandez, E., & Turk, D. C. (1992). Sensory and affective components of pain: Separation and synthesis. *Psychological Bulletin*, **112**, 205-217.

Fiedler, K. (1991). Heuristics and biases in theory formation: On the cognitive processes of those concerned with cognitive processes. *Theory & Psychology*, **1**, 407-430.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. *Psychological Review*, **103**, 193-214.

Fiedler, K. (2002). Frequency judgements and retrieval structures: Splitting, zooming, and merging the units of the empirical world. In P. Sedlmeier & T. Betsch (Eds.), *Etc. Frequency processing and cognition* (pp. 67-87). Oxford: Oxford University Press.

Fiedler, K., & Armbruster, T. (1994). Two halfs may be more than one whole: Category-split effects on frequency illusions. *Journal of Personality & Social Psychology*, **66**, 633-645. doi:10.1037/0022-3514.66.4.633

Fiedler, K., & Walther, E. (2004). *Stereotyping as inductive hypothesis testing*. New York: Psychology Press.

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology*, **8**, 172-179. doi:10.1037/h0034145

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

Haubensak, G. (1992). The consistency model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 303-309.

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

Kruger, J., & Evans, M. (2004). If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology*, **40**, 586-598.

Linville, P. W., & Fischer, G. W. (1991). Preferences for separating or combining events. *Journal of Personality & Social Psychology*, **60**, 5-23.

Mellers, B. A. (2000). Choice and the relative pleasure of consequences. *Psychological Bulletin*, **126**, 910-924.

Moore, D. A., & Small, D. A. (2007). Error and bias in comparative judgment: On being both better and worse than we think we are. *Journal of Personality & Social Psychology*, **92**, 972-989.

Morewedge, C. K., Gilbert, D. T., Keysar, B., Berkovits, M. J., & Wilson, T. D. (2007). Mispredicting the hedonic benefits of segregated gains. *Journal of Experimental Psychology: General*, **136**, 700-709.

Rottenstreich, Y. [S.], & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review*, **104**, 406-415.

Rulon, P. J. (1941). Problems of regression. *Harvard Educational Review*, **11**, 213-223.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, NJ: Erlbaum.

See, K. E., Fox, C. R., & Rottenstreich, Y. S. (2006). Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **32**, 1385-1402.

Sloman, S., Rottenstreich, Y. [S.], Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 573-582.

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

Unkelbach, C., & Memmert, D. (2008). Game management, context effects, and calibration: The case of yellow cards in soccer. *Journal of Sport & Exercise Psychology*, **30**, 95-109.

Weaver, K., Garcia, S. M., Schwarz, N., & Miller, D. T. (2007). Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. *Journal of Personality & Social Psychology*, **92**, 821-833.

White, C. M., & Koehler, D. (2006). Assessing evidential support in uncertain environments. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 261-298). New York: Cambridge University Press.

NOTES

1. The function *s* pertains to the subjective experience of quantities.

2. The argument holds for concavity of logarithmic, power, or any other function.

3. Alternatively, one could calculate the proportional deviation scores, defined as the ratio between judged and actual frequencies *j*(category)/*f*(category).

4. We deliberately decided to conserve this feature in the way we operationalized a category split, giving split categories a distinct color. For several reasons, this cannot, however, account for the supposed split effect: We empirically checked that the pale colors we used did not affect frequency estimates, as already mentioned. Analogous split effects were found in several previous experiments (cf. Fiedler, 2002) without colors,

and an alternative account, in terms of color salience, cannot explain the distinct interactions with cognitive load and extremity that were derived from the regression account.

5. Because each judge estimates only two nonsplit categories, the frequencies of which always totaled 26, it was appropriate to normalize basic regression scores to 26. Using the total presentation frequency (i.e., 52) would have confounded the basic regression effect with the general underestimation of nonsplit, relative to split, categories. Analyses of split-category estimates or comparisons of nonsplit and split categories, though, were based on scores normalized across all four categories (totaling 52).

6. We also computed proportional deviation scores [PDS = *j*(category)/*f*(category)], which were highly redundant with ADS scores and yielded virtually the same results. However, the normalization of raw estimates (to total 56) renders PDS inappropriate, because this linear transformation presupposes an additive measure for large and small frequencies. The reported analyses were therefore based on ADS.

7. Because of this restricted range, with only two slightly different frequencies, the no-load manipulation check was based on inaccuracy scores in Experiment 2; the slightest regression toward the mean would have left hardly any latitude for a load effect. Inaccuracy still increased with load ($M = 49.0$ vs. 41.9).

(Manuscript received August 12, 2008;
revision accepted for publication December 18, 2008.)