The Psychophysics of Number Integration: Evidence from the Lab and from the Field

Benjamin Scheibehenne

University of Geneva, Switzerland

Corresponding author

Benjamin Scheibehenne

University of Geneva

Geneva School of Economics and Management

40 bd du Pont-d'Avre

1211 Geneva, Switzerland

Phone: +41 22 379 9498.

E-mail: benjamin.scheibehenne@unige.ch

Acknowledgements

The author thanks Antonia Krefeld-Schwalb, Sarah Kuhn, and Sebastian Olschewski for their helpful comments and Anita Todd for editing the manuscript. This research was supported by research grants from the Swiss National Science Foundation (No. 100014_130149 and 100014_149846).

Abstract

The subjective integration of numbers that are encountered sequentially is an elementary judgment process that is highly relevant in research (e.g. in a decisions from experience paradigm) and in everyday life alike (e.g. when keeping track of spending during a shopping trip). Towards a better understanding of how people perceive and integrate numerical information, participants in a laboratory experiment (n = 40) repeatedly estimated the sum of a number sequence briefly presented on a computer screen. Results indicate a systematic bias towards underestimation that could be captured with a compressive power function. The observed underestimation depended on the sequential order in which the numbers were presented but not on the shape of the underlying frequency distribution. Similar results were obtained in a field study where customers in a grocery store (n = 966) systematically underestimated the total value of their shopping basket prior to checkout. A model comparison approach revealed that the observed underestimation in the lab study was best captured by a compressed mental number line when evaluating single items while in the field study, the bias rather stem from a systematic error during information integration. The field study further indicated that underestimation increased with age but was not due to a simple rounding strategy or the systematic forgetting of unhealthy items such as sweet or fatty snacks. The results yield novel insights into how people perceive and integrate numbers.

Keywords: Psychophysics, Number estimation, mental arithmetic, consumer behavior, cognitive judgment processes, mathematical intuition

The subjective perception and integration of discrete numerical information is an elementary cognitive process that is highly relevant for research in psychology, cognitive science, and economics alike (Anderson, 1981; Ashcraft, 1992; Dehaene et al., 1999). The mental arithmetics that govern information integration also provide the basis for measures such as expected utility and the evaluation of monetary gambles which are fundamental for many theories on judgment and decision making (Anderson, 1996; Stewart, Chater, & Brown, 2006). Finding approximate solutions to arithmetic problems without computing the exact answer is an important component of mathematical cognition (Ashcraft, 1992; Dehaene et al. 1999). Number integration also occurs in many everyday situations where information is accumulated over time. For example, in a consumer context, purchase decisions often depend on price estimations (Alba, Broniarczyk, Shimp, & Urbany, 1994), and keeping track of spending during a shopping trip or at a restaurant can help prevent overspending (Heath & Soll, 1996). Given the importance of this ability, it is important to better understand how people perceive and aggregate sequentially presented numerical information and what factors influence their estimation accuracy.

Factors That Influence Estimation Accuracy

Research on psychophysics indicates that subjective perceptions are often well-described by compressive power functions of objective values (Stevens, 1957). Compressive functions imply a tendency for underestimation that gets stronger as the objective values grow larger. This scaling has also been applied to the mental representation of numbers (Longo & Lourenco, 2007) and it is commonly used to capture the subjective utility of money in an economic context (Bernoulli, 1738/ 1954; Kahneman & Tversky, 1979, 1984). In line with this, customers at a grocery store were found to systematically underestimate the total value of their shopping baskets (Van Ittersum, Pennings, & Wansink, 2010).

According to information integration theory (Anderson, 1981), observed underestimation on a behavioral level could be caused by two qualitatively different cognitive processes. It could either occur during valuation, suggesting a compressive mental scaling of single numerals, or it could occur during integration, suggesting a linear scaling in combination with a systematic integration error. The former case ("scaling first") can be formally described as a sum of compressed numbers (i.e. $\sum f[x]$, where f is a scaling function and x is a sequence of numbers) whereas the latter ("sum first") assumes a bias during integration (i.e. $f[\sum x]$). To account for underestimation, researchers typically implement a power function with an exponent $\theta < 1$, sometimes in combination with a linear scaling factor or "proportionality constant" w (e.g. Stevens 1957):

$$f(x) = w \cdot x^{\theta} \quad (1)$$

Implemented this way, the qualitative difference between the scaling first and the sum first model can be illustrated based on a simplified case of two sequences that have equal sums but different addends (e.g. {50;50} and {1;99}). Here, the sum first model necessarily makes identical predictions irrespective of the specific parameters values while the scaling first model makes different predictions for both sequences as long as the exponent does not equal exactly one or zero.[1] Likewise, it becomes easier to distinguish both models if the scaling is more compressed (i.e. the exponent becomes smaller than 1).

Empirical evidence supporting compressed mental scaling comes from experiments using nonsymbolic numerosities such as clouds of dots presented on a screen (Dehaene, 2007, 2009). On the other hand, research using symbolic numbers (i.e. written Arabic numerals) often finds linear scaling, at least for adult Western populations (Dehaene, Izard, Spelke, & Pica, 2008;

[1] Technically, for the simplified example at hand, one can find a parameter combination where both models make identical predictions. As the number and the length of sequences increase, the models can be rigorously distinguished.

Siegler & Opfer, 2003). However, as symbolic numbers lend themselves to exact counting and adding operations, linear scaling might not necessarily apply to mathematical intuition and approximate numerosity (Dehaene, 2007).

Besides mental scaling and integration functions, estimation accuracy may also depend on the serial order in which information is presented. Past research has found evidence for both primacy and recency effects such that stimuli presented at the beginning and/or the end of a sequence received higher weight (Hogarth & Einhorn, 1992). When estimating the product of a sequence of numbers, primacy or "anchoring" effects seem to prevail (Tversky & Kahneman, 1974), suggesting that sequences starting with relatively low numbers are more likely to be underestimated. Research in which participants estimated the mean of a number sequence often found recency effects, especially for short sequences (e.g. Brezis, Bronfman, & Usher, 2015; Tsetsos, Charter, & Usher, 2012). To formally model such order effects, so-called serial position curves can be estimated that assign a weight to the information depending on its position in the sequence (Anderson, 1996).

Estimation accuracy for a sequence of numbers may also depend on the shape of the underlying frequency distribution. Experimental evidence from research on risky choices indicates that preferences critically depend on the distribution of values that people experienced in the past (Stewart, 2009). Likewise, grocery shoppers can be influenced by the skewness of product prices over time (Niedrich, Weathers, Hill, & Bell, 2009). Such patterns can be explained by several theoretical accounts including the decision by sampling theory (Stewart, 2009) and the range–frequency model (Parducci, 1965) and they align with early research on perception showing that negatively skewed distributions (many large and few small values) lead to lower mean estimates compared to positively skewed distributions (Parducci, Thaler, & Anderson, 1968). Finally, possible over- and underestimation may also be due to simplifying strategies such

as rounding numbers up or down prior to adding them (Lemaire, Arnaud, & Lecacheur; 2004; Van Ittersum et al., 2010).

Thus far, past research commonly tested these factors in isolation, which makes it difficult to evaluate their relative importance. A better understanding of how people integrate sequential information requires a design that tests these influences conjointly. Towards this goal, I conducted two studies: a laboratory experiment in which participants estimated the sum of a sequence of monetary values presented on a computer screen, and a field study in which customers at a grocery store estimated the total value of their shopping baskets prior to checkout.

Laboratory Experiment

Method

Local university students ($N = 40$; *Mdn* age = 21 years; SD = 5.7; 33 females) participated in the experiment in exchange for course credit. The sample size was determined prior to conducting the study. No variables or experimental conditions were dropped. Each participant repeatedly estimated the total sum of 24 numbers described as the prices of fictitious items in a shopping basket. The numbers were sequentially displayed on a computer screen for 0.5 s each, similar to in the Japanese game "flash anzan" (Bellos, 2010). The rapid presentation has been used in previous research. In difference to the summation task at hand, participants in these previous experiments either estimate the mean (e.g. Brezis, Bronfman, & Usher, 2015; Bronfman et al. 2015; Malmi & Samson, 1983; Tsetsos, Chater, & Usher, 2012), chose between sequences (e.g. Zeigenfuse, Pleskac, & Liu, 2014; Tsetsos, Chater, & Usher, 2012), or tried to re-construct the underlying frequency distribution (e.g. Goldstein & Rothschild, 2014).

After observing each sequence, participants typed in their best estimate for the total sum before proceeding to the next trial. The short presentation time inhibited the application of exact

arithmetic strategies but rather required an intuitive approximation. For each sequence, the shape of the underlying frequency distribution, the sequential order in which the numbers were displayed, and the range of these numbers were subject to experimental manipulation. The shape of the distribution was either uniform, positively skewed, negatively skewed, unimodal, or bimodal. The sequential order was either increasing, decreasing, U-shaped, or inversely U-shaped. The numbers ranged from 0.1 to either 5 (average sum = 66), 15 (average sum = 199), or 25 (average sum = 331).[2] For each sequence, a small amount of random noise (±10% of the respective upper range, uniformly distributed) was added. Combining these factors led to 5×4×3 = 60 different sequences. Each participant also saw 15 "baseline" sequences with uniformly distributed random numbers scattered around 2.8, 8.3, or 13.8, the mean within each number range. Due to the way the sequences were constructed, the true sum was comparable across all conditions except for the negatively skewed distributions where the sum was slightly higher, and the positively skewed distributions where the sum was slightly lower than average. All numbers were rounded to two decimal places. Likewise, participants could also enter their estimates with up to two decimal places. Sequences were presented in random order in a within-subject experimental design.

At the end of every 10th round, participants received feedback about their mean estimation accuracy in the preceding rounds, expressed as the percentage by which their estimate deviated from the true sum. As facilitation toward subjectively estimating the sum rather than actually adding the exact numbers, participants also completed a secondary working memory task that required them to memorize four randomly drawn capital letters that were displayed before each trial and had to be recalled in correct order at the end of each trial.

---

[2] Appendix A displays the trajectories and frequency distributions of all presented sequences.

To incentivize accuracy, participants received a bonus at the end of the experiment that was determined by multiplying their mean accuracy across all rounds by the number of trials in which they accurately remembered the letter sequence. This number was divided by 10 and then paid out in Swiss francs (CHF; 1 CHF = approximately 1.1 U.S. dollars).

The accuracy of participants' estimations ($\hat{y}$) was quantified as a bias measure indicating over- or underestimation proportional to the true sum (y):

$$\text{bias} = \frac{\hat{y}-y}{y} \quad (1)$$

This measure is similar to the (exponential) signed order of magnitude error that is sometimes reported in the literature (e.g., Brown & Siegler, 1992).

For 16 of the 3,000 observations, the bias was larger than 1. Closer inspection of these cases suggests that they mainly occurred because participants mistakenly entered one digit too many when typing in the true sum. Thus, they were coded as missing prior to data analysis.

Results

Overall, 65% of all sequences were underestimated[3]. The mean bias across participants was -.055 (SD = .066; Cohen's d = 0.8), and for 35 of 40 participants, the bias was negative, indicating underestimation. This proportion of participants was highly unlikely under the null hypothesis of no bias as indicated by a binomial test ($p < .001$). The corresponding Bayes factor (BF) was greater than 10,000, indicating extreme evidence for the alternative hypothesis (Rouder et al., 2009).

[3] The raw data can be downloaded at: https://osf.io/2xaum/ along with an R-script that reproduces the statistical analyses and Figures presented in this manuscript.

Estimating Stevens's power law. To test if underestimation became stronger for higher sums, as predicted by Stevens's power law, a mixed effects regression with random intercept and slope was estimated on the log-transformed data by means of the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2014). Using log-transformed data solved potential biases due to the skewed distribution of both true and estimated values and the resulting heteroscedasticity in the data. Results yielded an intercept (i.e., a proportionality constant) of 1.1 and an exponent of 0.97. These results indicate that underestimation became slightly stronger for higher sums. For the lowest quintile, participants on average underestimated the true value by about 3.4% while for the highest quintile it was about 7.7%. Figure 1 illustrates this relationship. Figure 1 further shows that the estimation error (i.e. $|\hat{y}-y|$) increased with the true sum, hence revealing a magnitude effect (Ashcraft, 1992).

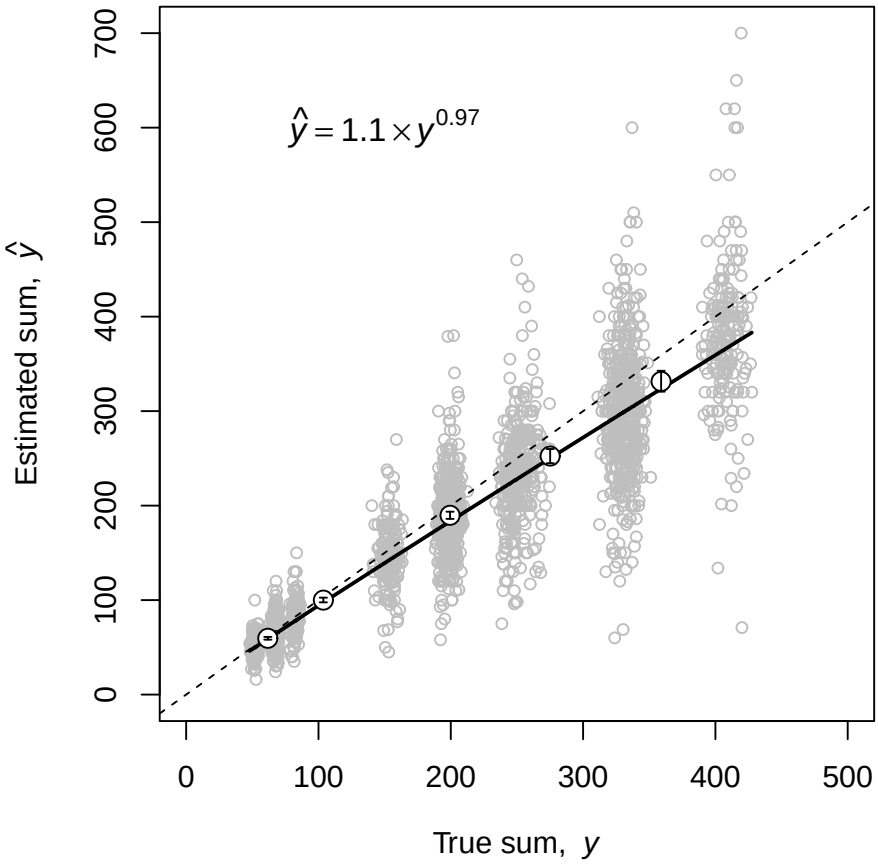


Figure 1. Plot of the true sum against the estimated sum across all participants. Open gray circles represent single sequences. The fitted regression is plotted as a black line. The open black circles indicate the mean quintiles, the error bars inside the black circles are 95% confidence intervals across the individual quintiles (bootstrapped).

Factors that influence the estimation bias. Figure 2 plots the mean estimation bias separately for different sequential orders and for different frequency distributions. As can be seen from the figure, the bias varied depending on the sequential order. Underestimation was most pronounced for increasing and U-shaped sequences that both end with high numbers, while

decreasing and peaked sequences that end with low numbers were estimated more accurately. Together, these patterns suggests relatively stronger underweighting for numbers appearing at the end of the presented sequences and hence a primacy effect. Figure 2 further illustrates that the bias was independent of the underlying frequency distribution.

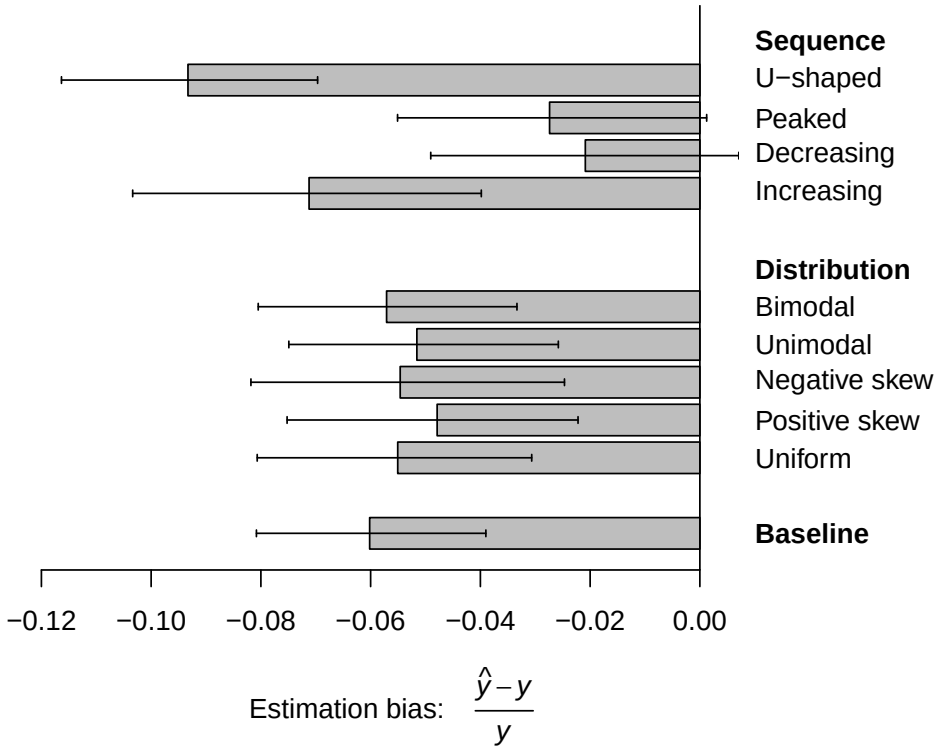


Figure 2. Mean estimation bias depending on the sequential order of the number presentation and the underlying frequency distribution. Error bars indicate 95% confidence intervals across individual means (bootstrapped).

To further corroborate these findings across experimental conditions on statistical grounds, I fit a mixed-effects regression model to the data with random slopes and intercepts

across participants for the true sum and random intercepts for each sequence and each distribution level. All data except the baseline condition were included to make sure that each sequence level was combined with each distribution level in the data.

In a first step, I estimated a baseline model with the intercept as an additional fixed effect, thus assuming a constant bias across all sequences. In comparison to this, an extended model that includes the true sum as an additional predictor (fixed effect) yields a Bayesian information criterion (BIC) difference of 5, which translates into a BF of 12. Thus, in light of the data, the extended model became 12 times more likely than the baseline model, $\chi^2(1) = 12.8$; $p < .001$. As shown in Table 1, the estimated beta coefficient for the true sum was negative, indicating a stronger underestimation bias for larger sums. To improve numerical estimation within the lme4 package, the true sum was scaled to cents (i.e. divided by 100) before entering the model. This transformation did not affect any of the results.

Table 1

Fixed Effect Estimates and Model Comparison for the Laboratory Experiment [a]

	Model				
	Baseline	1	2	3	4
			Fixed effects		
Intercept	-3.86	-1.08	-4.93	-5.13	-4.65
True sum/100		-3.88	-4.02	-4.02	-4.15
Trial position			8.85	8.76	8.77
Sequential order					
Decreasing				3.18	3.18
Peaked				2.96	2.96
U-shaped				-1.34	-1.34
Distribution					
Positive skew					-0.29
Negative skew					0.86
Unimodal					0.02
Bimodal					-0.36
			Model fit		
BIC	-1274	-1279	-1348	-1351	-1322
log(Likelihood)	664.3	670.6	709	722.3	723.2
			Bayes factor		
Model/Baseline	-	12	>10,000	>10,000	>10,000
Model/Model 1		-	>10,000	>10,000	>10,000
Model/Model 2			-	5	<1
Model/Model 3				-	<1

Note. BIC = Bayesian information criterion.

[a] Fixed effect coefficients are reported as standardized scores (i.e., estimate/standard error).

The model fit further improved when including the randomized position of each trial in the experiment as an additional predictor (BIC difference = 69; BF > 10,000; $\chi^2(1) = 76.8$; $p < .001$). Here, the coefficient was positive, indicating a learning effect: Despite the sparse feedback, participants' were less biased for trials presented at the end of the experimental session. Adding the sequential order of the presented numbers within each trial (i.e. "increasing", "decreasing", "u-shaped", and "inversely u-shaped") as a predictor in the model further improved fit. The difference in BIC between this model and the previous one that included just the true sum was 3, BF= 5; $\chi^2(3) = 26.6$; $p < .001$. This result corroborates the differences shown in Figure 2.

Adding the shape of the frequency distribution or any higher moments (variance, skewness, and kurtosis) as additional predictors did not yield a further improvement in model fit as measured by BIC differences, indicating that the estimation bias did not credibly depend on the underlying frequency distribution, as also shown in Figure 2. Furthermore, underestimation also did not depend on the proportion of numbers with "low" cent endings (i.e., 0–49 cents), thus rendering the explanation of a rounding-down strategy less likely. The bias was also independent of the proportion 1-digit numbers within each sequence. Thus, the underestimation was not due to missing large numbers. The proportion of 1-digit numbers were highly correlated with the total sum though, which reduces the power to rigorously test this explanation with the data at hand.

The estimation bias and the estimation error, (i.e. |bias|) were independent of the accuracy in the dual letter task, indicating that participants did not systematically trade-off accuracies in both trials. The low correlation may be partly due to a sealing effect in the letter task with a median of 67 out of 75 correct answers (interquartile range 62 to 69 correct answers).

Number Integration Model. To test if the underestimation occurs at the level of individual items ("scaling first" model) or when integrating ("sum first"), both models were implemented in a hierarchical Bayesian framework. As the previous analysis indicated an

influence of the sequential order, both models also incorporated a serial weighting vector *w* (Anderson, 1981). For the scaling first model, the estimation response ŷ of a single participant *i* for a number sequence *j* was modelled as

$$\hat{y}_{ij} = \sum\left(w_i \cdot {x_{ij}}^{\theta_i}\right) + \varepsilon_i, \quad (2)$$

where ε_i represents a normally distributed error term. Likewise, the sum first model was implemented as

$$\hat{y}_{ij} = \sum\left(w_i \cdot x_{ij}\right)^{\theta_i} + \varepsilon_i. \quad (3)$$

For both models, the weighting vector was a linear function of the serial position vector *s*. In its simplest form, the weighting consists of just a constant or intercept β0 that is independent of the serial position:

$$w_i = \beta 0_i \quad (4)$$

Assuming that numbers are under- or overweighed depending on their position within the sequence (hence allowing for either primacy or recency effects) requires a linear weighting vector where the relative weight is a function of the serial position:

$$w_i = \beta 0_i + \beta 1_i \cdot s \quad (5)$$

An even more flexible model that allows for both, primacy and recency effects can be implemented by means of a quadratic function:

$$w_i = \beta 0_i + \beta 1_i * s + \beta 2_i * s^2 \quad (6)$$

Polynomials of higher degree were not considered here because their shape becomes increasingly difficult to interpret on theoretical grounds.

Bayesian Model Implementation. Preparatory model recovery analyses indicated that the experimental design allowed distinguishing the scaling first and the sum first models across a wide range of possible parameter values on the level of individual participants. To improve estimation efficiency, the serial position *s* was normalized ($M = 0$, $SD = 1$) prior to entering the

model and the models were estimated on the log-transformed data, including the baseline condition. The priors on the individual-level parameters (i.e. the exponent θ and the respective β parameters) were normally distributed with means and standard deviations that were partially pooled through weakly informative normally distributed priors on the group-level. The prior group-level means for β0 and θ were set to zero, for β1 and β2 the group-level means were set to one. The prior standard deviations for these group-level distributions were all set to five. The group-level priors for the standard deviations on an individual level were half-cauchy distributed with location parameter zero and scale parameter two. Finally, the likelihood function was normal with a half-cauchy (0,2) prior on the standard deviation. Posterior estimates were obtained with the Stan software (Stan Development Team, 2015) that was called from R[4]. The sampling procedure within Stan was efficient after some thinning was applied (i.e. R-hat < 1.01).

Model Estimation. Table 2 provides an overview of the fit (log-likelihood) and the predictive accuracy for the different models. The latter was measured with the approximate leave-one-out cross-validation information criterion (LOOIC, Vehtari, Gelman, & Gabry, 2015), and the widely applicable information criterion (WAIC, Watanabe, 2010) that both take model complexity into account. As can be seen from the table, all three measures indicate that "scaling first" models explain the data better than "sum first" models, irrespective of the specific type of weighting function that was used. Across all model implementations, the mean estimate of the exponent (θ) was credibly smaller than zero as indicated by a Bayesian 95% confidence interval. The model estimates were not driven by the choice of priors as a model estimate with no priors using STAN yield similar results. Together, this provides evidence for compressive mental scaling of single numerals prior to integration.

[4] The Stan code and the respective R function to run the models are also available in the online supplement at https://osf.io/2xaum/

Table 2

Quantitative Measures of Model Fit for the Lab Study

	Model					
	Sum First			Scaling First		
Serial Weighting	None	Linear	Quadratic	None	Linear	Quadratic
log likelihood (max)	797	888	927	819	914	954
WAIC	-1,382	-1,509	-1,563	-1,402	-1,536	-1,586
LOOIC	-1,370	-1,495	-1,550	-1,389	-1,522	-1,573

Note: For log likelihood, higher values indicate better fit. For WAIC and LOOIC, lower values indicate better fit.

The model comparison further shows that models with a quadratic weighting vector (i.e. Equation 6) provide a better explanation of the data than models with linear or constant weighting, confirming that sequential order influenced estimation accuracy. The estimated quadratic weighting vector is plotted in Figure 3. As can be seen from the figure, the shape of the weighting vector on the group-level is inversely u-shaped, indicating that—on average—numbers in the middle of the sequence received relatively higher weights as compared to numbers at the beginning or the end. The Figure also reveals noticeable individual differences. For many participants, the weighting curve points downward, indicating relatively higher weight for numbers at the beginning of the sequence and hence a primacy effect. A few participants also exhibit an upward slope and hence a recency effect. The quadratic weighting vector dovetails with the observed patterns in Figure 2 showing that u-shaped sequences were most likely to be underestimated while peaked and decreasing sequences were estimated relatively accurately.

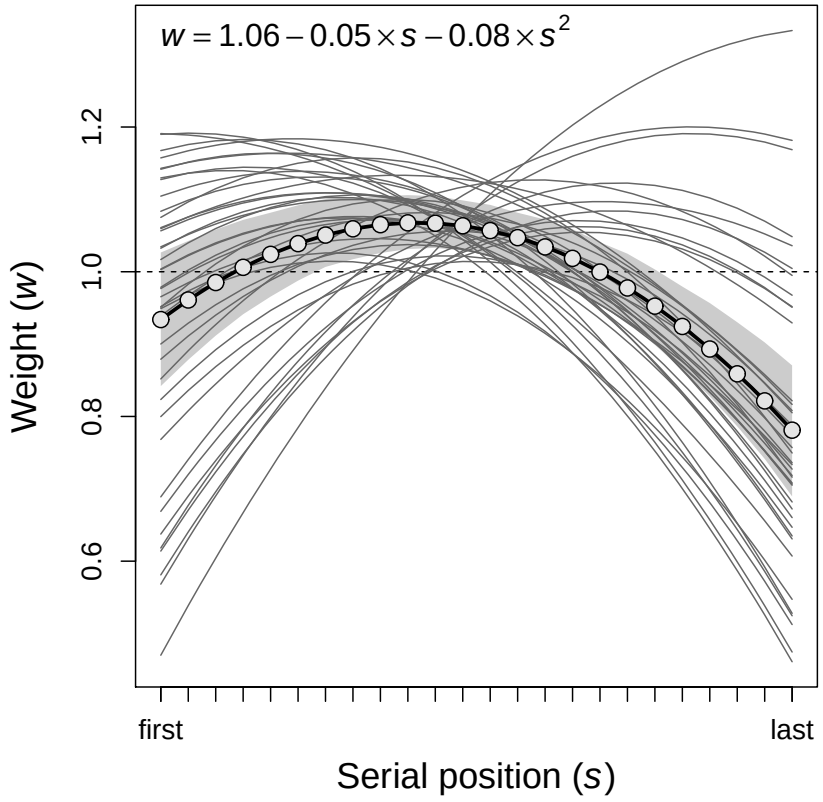


Figure 3. Quadratic weighting vector in the "scaling first" model. The thick line indicates the group-level estimate, the underlying gray area marks the 95% highest posterior density interval and the thin lines represent the estimates for each individual participant.

Field Study

To test if the compressive scaling observed in the lab experiment generalizes to real-world scenarios, I conducted a field study at a local grocery store where customers who had lined up at the checkout gave a spontaneous estimate of the total value of their shopping basket. After the checkout, this estimate was compared to the actual value indicated on their sales receipt. Besides testing for a possible estimation bias, this design also aimed to compare the "scaling first" against

the “sum first” model and it allowed testing a possible influence of the underlying price distribution, the number and types of items purchased, and the mode of payment. The field study also provided an opportunity to test possible influences of participants’ age. Research on cognitive aging indicates that the ability to conduct simple arithmetic operations such as addition declines with age due to a decrease in cognitive resources such as working memory or processing speed (Baltes & Baltes, 1990; Gandini, Lemaire, & Dufau, 2008). While this suggests an increase in estimation error with age, elderly participants may also have had more experience and thus opportunity for feedback in real-world environments such as grocery stores, which might lead to more accurate estimates.

When integrating prices, spontaneous, unplanned purchases may be less likely to be encoded and retrieved from memory (Baddeley, 1992). If so, shopping baskets that contain items that are often purchased on impulse, such as sweet and fatty snacks, could be more prone to underestimation (Erdelyi & Goldberg, 2014). Another indication of planned purchases is the use of external memory aids such as shopping lists (Block & Morwitz, 1999). Thus, customers who use a shopping list might be more accurate in estimating the total value of their baskets and the number of items that they purchased.

Finally, underestimation might be less likely for customers who pay cash and thus want to avoid overspending. In line with this prediction, customers who pay via credit or debit card seem more prone to overspending, perhaps because they underestimate or forget the total value of their basket (Prelec & Simester, 2001).

Method

Data was collected on the sales floor of a local grocery store during two sessions that were one year apart. At each session, customers with a shopping basket were approached by one of two female research assistants. The first session took place on three consecutive days (Thursday

to Saturday) from about 3:00 to 7:00 p.m. on Thursday and Friday and from 10:00 a.m. to 2:00 p.m. on Saturday. The second session took place on a Friday (3:00 p.m. to 7:00 p.m.) and a Saturday (10:00 a.m. to 2:00 p.m.). When standing in line before the checkout, participants gave a spontaneous estimate of the total value of their shopping basket. Participants at the second session also estimated the total number of items in their basket. The order of the two questions was counterbalanced. Estimation accuracy was not monetarily incentivized but participants seemed generally motivated and engaged in the task. Participants at the second session also indicated if they had used a shopping list or not. At both sessions, all customers who were approached agreed to participate in the study. After checkout, customers exchanged their sales receipt for a piece of chocolate or a flower. During this transaction, a research assistant noted participants' gender and estimated their age on a decade-scale (i.e. 20-30, 30-40, etc.). A total of 966 customers participated in the study, 545 in the first session and 421 in the second session. Fifty-eight percent of the participants were female and the mean estimated age was 42 years (*SD* = 12).

Results

Across all customers, the average basket value (calculated as the geometric mean to account for skewness in the data) was 50.6 CHF (first session: 45 CHF, second session: 58.9 CHF). The respective median and arithmetic mean values were 49.9 CHF and 63.79 CHF. The interquartile range (i.e. the middle 50%) was between 32.4 CHF and 79.1 CHF. The (geometric) mean number of items within each basket was 14.5 (first session: 14, second session: 15.1). The respective median and arithmetic mean values were 14 and 16.8. The middle 50% of all baskets contained between 10 and 20 items. Accordingly, the (geometric) mean value of a single item across all baskets was 2.7 CHF (median = 2.9 CHF; arithmetic mean = 3.8 CHF).

The median bias was -0.05 (interquartile range = -0.19 to 0.13; M = -.02; Cohen's d = 0.2). In total, 60% of all baskets were underestimated (first session: 62%, second session: 57%), which was unlikely under the null hypothesis of no bias as indicated by a binomial test ($p < .001$; BF > 10,000).

Estimating a regression on the (log) estimated basket value with the (log) true value as predictor yields an intercept of 1.44 and an exponent of 0.89, indicating that underestimation got stronger for larger sums, while small values were slightly overestimated. In particular, a 10% increase in the true basket value yields only a 9% increase in the estimated value. In other words, a basket worth 51 CHF (the average basket value) was underestimated by 3 CHF or 8% while a basket worth 79 CHF (the upper 75% quantile) was underestimated by 8 CHF or 10%. Figure 4 illustrates this relationship. The estimated curvilinear relationship is quite similar for both sampling sessions (first session: intercept = 1.36; exponent = 0.9, second session: intercept = 1.7; exponent = 0.86) and the compression is slightly more pronounced compared to the laboratory experiment.

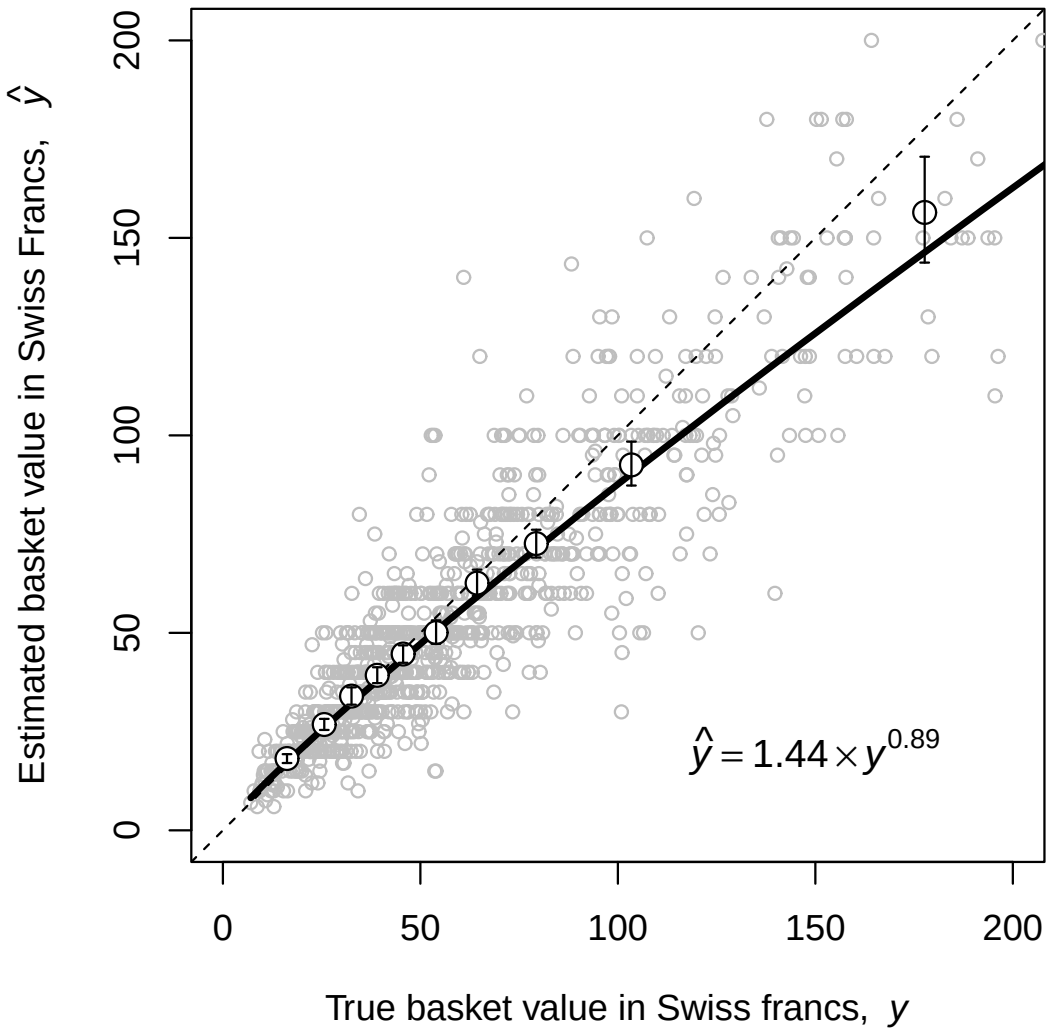


Figure 4. True basket value plotted against estimated basket value. The gray circles are individual estimates. Circles below the diagonal (dashed line) indicate underestimation and circles above overestimation. The circles drawn in black indicate the arithmetic mean of the estimates across 10% percentiles. Error bars indicate the 95% confidence interval of these means (bootstrapped). The estimated regression line is plotted in dark black. The plot does not show 21 data points with values > 200.

To analyze possible factors that influence the amount of underestimation, I estimated a stepwise linear regression with bias as dependent variable, similar to the lab experiment. The analysis excluded 15 participants for whom age was not recorded and another eight for whom the mode of payment (cash or credit) was not recorded. In a first step, a baseline model was estimated with only the logarithm of the true basket value as predictor. The fit of this baseline model improved when participants' age was entered (BF = 297). As shown in Table 3, the estimated beta coefficient was negative, indicating that older participants underestimated the basket value more strongly. In particular, when keeping basket value constant, underestimation further decreased by 3% with every 10 years of age.

Table 3

Regression Coefficients and Model Comparison for the Field Study Data

	Model			
	Baseline	1	2	3
Fixed effects				
Intercept	-0.02	0.41	0.5	0.64
log(True sum)		-0.11	-0.1	-0.16
Age			-3.0E-03	-2.9E-03
Number of items				4.7E-03
Model fit				
BIC	246	179	168	161
log(Likelihood)	-116	-79	-70	-64
Bayes factor				
Model/Baseline	-	> 10,000	> 10,000	> 10,000
Model/Model 1		-	297	6,553
Model/Model 2			-	22

Note. To ensure that the model comparison was not affected by differences in sample size, BIC, log(Likelihood) and Bayes factors for all models were calculated after excluding 21 participants for whom data on age (n = 15) and/or mode of payment (n = 8) was missing.

Adding the actual number of items in the basket as a predictor further improved model fit (BF = 22), indicating that, when keeping basket value and age constant, underestimation increased for baskets that contained a few (relatively expensive) as compared to many (relatively inexpensive) items. Adding the mode of payment (cash vs. credit) or properties of the distribution of items, including its variance and skewness, did not improve model fit further, indicating that these variables did not influence the bias.

In a next step, I estimated and compared the scaling first and the sum first model using Stan. The model implementation was similar to the lab study except for two differences: First, instead of a weighting vector, a simple proportionality constant as in Equation 4 was estimated because no reliable information was available about the sequence in which the items were purchased[5]. Second, no hierarchical structure was implemented as each participant in the field study only contributed one estimate. Like in the lab study, preceding model recovery analyses indicated that the data allowed distinguishing both models across a wide range of possible parameter values.

Results of the model estimates confirm that the exponent was credibly smaller than one. In contrast to the lab experiment however, Table 4 shows that model comparisons based on LOOIC, WAIC and (log) likelihood indicate that the "sum first" model describes the field data slightly better than the "scaling first" model. Thus, the observed underestimation in the grocery

---

[5] Attempts to approximate the item order from its position on the sales receipt confirmed the effects found in the lab study, but analyses based on the position of items on the sales floor led to inconclusive results. Therefore, possible sequential order effects are not included in the model.

store seemed to stem from a systematic aggregation error rather than a compressive scaling of single numbers.

Table 4

Quantitative Measures of Model Fit for the Field Study

	Model	
	Sum First	Scaling First
Log likelihood (max)	-113.8	-117.4
WAIC	234.3	241.8
LOOIC	234.3	241.9

Note: For log likelihood, higher values indicate better fit. For WAIC and LOOIC, lower values indicate better fit.

One possible explanation for the observed underestimation which does not assume compressive scaling of single numbers would be that people did not pay attention to the cent-endings of the items they purchased, for example, because they rounded prices down to the nearest integer. In this case, underestimation would be stronger for baskets that had a higher proportion of items ending between .51 and .99 cents. However, entering this proportion as a predictor into the regression analysis above did not improve model fit, suggesting that underestimation was not due to a rounding-down strategy.

At the second sampling session, 39% of all participants stated that they used a shopping list. The use of a shopping list had no influence on the bias though, suggesting that underestimation was not driven by unplanned or impulse purchases. In line with this,

underestimation did not increase if customers bought potential impulse products such as sweets (40% of all baskets contained at least one sweet item) or alcohol (30% of all baskets contained at least one alcoholic beverage).

Another possible reason why people underestimated the value of their baskets is that they did not remember all the items they had chosen. However, on average, participants at the second session overestimated the number of items in their baskets by about 1.9 (SD = 7.57; Mdn = -1), $t(420) = 5.1$, $p < .001$; BF > 10,000, and only 35% of all participants underestimated the number of items, rendering this explanation unlikely.

General Discussion

Most participants in the laboratory experiment and in the field study underestimated the true sum of a sequence of numbers. This bias increased for larger sums. These results align with previous findings of similar patterns of underestimation with respect to the perception of numerals in general (Dehaene, 2011) and in a consumer context in particular (Van Ittersum et al., 2010). Which factors influenced people's estimation bias?

Influence of Sequential Order

In the lab experiment, underestimation systematically depended on the sequential order in which the numbers were presented. On the group-level, this dependency could be captured with an inversely u-shaped serial weighting curve. Underlying this average were many individual participants who showed a primacy effect. This primacy effect dovetails with similar findings in the judgment and decision making literature (Tversky & Kahneman, 1974) and it corroborates previous research showing that people's accuracy in remembering items declines over the length of a sequence (Hurlstone, Hitch, & Baddeley, 2014). Results from the field study also show

stronger underestimation for elderly participants, suggesting that number integration relies on cognitive processes that decay with age (Baltes & Baltes, 1990).

However, the inversely u-shaped pattern on the group-level differs from previous findings based on similar experimental designs that show recency effects (e.g. Brezis, Bronfman, & Usher, 2015; Tsetsos, Charter, & Usher, 2012). Research on memory suggests that better recall for middle positions is not unheard of but it seems to be the exception rather than the rule (e.g. Jones & Oberauer, 2013). One possible explanation for this difference could be that the observed group-level pattern stems from averaging across individual participants who showed either a primacy or a recency effect. Alternatively, the different patterns could reflect qualitative differences in the underlying cognitive processes for approximating means and sums respectively. While from an algebraic perspective, summing and averaging are closely related calculations , this correspondence may not necessarily hold for mental arithmetic (Gigerenzer, 1991).

Sum First vs. Scaling First

Despite descriptively similar data patterns across both studies, the model comparison approach reveals a qualitative difference in the underlying cognitive processes. In the lab study, the best explanation for the observed underestimation was a compressed mental number line while in the field study, the bias seems to stem from a systematic aggregation error. The latter result accords with the stronger underestimation for elderly participants, but it was not due to a simple rounding strategy, an underestimation of the total number of items, or systematically forgetting of presumably unhealthy items such as sweet or fatty snacks that are prone to unplanned impulse purchases.

The difference between the lab and the field study may also be due to qualitative differences in the task itself. Compared to the field study, participants in the lab were explicitly instructed to estimate the sum prior to the task, got monetarily incentivized, and also received feedback over time, all of which may have triggered different, presumably more conscious or strategic number integration strategies. A further elucidation of the underlying mechanisms awaits additional research.

No Effect of Frequency Distributions

Across both studies, underestimation did not depend on the underlying frequency distribution. While this is surprising given the empirical and theoretical support for such an influence outlined in the introduction, the results are in line with previous research that also did not find such a relationship when information was presented sequentially (e.g. Hutchinson, Wilke, & Todd, 2008). Perhaps the differences in the higher moments of the observed distributions in the field and in the laboratory experiment were not strong enough to be noticed. As mentioned by Stewart (2009), a possible effect also may have been overshadowed by participants' past experiences with price distributions on a daily basis.

Why Did People Underestimate?

The results at hand shed light on the question of how people perceive and process numbers. Past evidence for a compressive mental number line mostly stems from experiments using non-symbolic numerals such as clouds of dots (Dehaene, 2007, 2009). In extension to this, the results from the lab study indicate that compressive scaling also holds for adult Western populations processing Arabic numbers. These results align with results from Brezis, Bronfman, and Usher (2015) showing that participants in an experiment slightly underestimated the mean of

a rapidly presented number sequence. Presumably, in these the compressed scaling because the experimental setting inhibited exact counting and adding and induced people to rely on approximation and mathematical intuition. The compressive scaling found for numbers shares many properties with psychophysical functions and thus may build on similar neural processes (Nieder & Dehaene, 2009; Verguts & Fias, 2004).

In the field study, customers underestimated the value their baskets even though grocery shopping is a common task that provides ample opportunity for feedback and learning, in particular if national inflation rates are low, as was the case at the time of the study in Switzerland. While this finding matches with insights from psychophysics indicating that subjective scales are often resistant to training or experience (Stevens, 1957), the model comparison revealed that the underestimation was rather due to a systematic aggregation error than a compressive mental number line.

What could be the reason for this error in the field study? Past research indicates that people's estimates systematically depend on past experiences in a given context (Stewart et al., 2006). The study was conducted at the end of the week, so if the grocery baskets that participants usually purchase are less expensive, they might have corrected their estimate towards the long-run average (see also Dehaene & Mehler, 1992). While this explanation would also account for the observed overestimation for small baskets, it conflicts with the finding that most participants overestimated the number of items they bought. Thus, rigorously testing this hypothesis requires knowledge of past shopping experiences in a longitudinal design.

Implications for Related Theories

The present research further contributes to recent theoretical work in economics and decision making, where information about risky options is sometimes presented sequentially in a

decision from experience paradigm (Hertwig, Barron, Weber, & Erev, 2004; Zeigenfuse et al, 2014). When choosing among gambles, people often appear risk averse, which can be captured through a marginally decreasing utility function. In a choice paradigm, it is difficult to disentangle perceptual biases and idiosyncratic preferences, though. Here, asking people to estimate the sum of a set of numbers provides a more direct measure of the subjective value of money that does not require fitting highly parameterized models that can be difficult to interpret (Scheibehenne & Pachur, 2015). In the present estimation task, where subjective (risk) preferences did not come into play, the results suggest that the curvature of the utility function partly reflects perceptual, psychophysical biases or systematic errors when aggregating numerals.

From an applied perspective, the results suggest that estimation accuracy can be improved by separately integrating different parts of the number sequence (such as different accounts or product categories) and then adding these separate estimates in a second step (see also Chandon & Wansink, 2007). Likewise, the observed sequence effect indicates that accuracy may increase when expensive items are encountered early on.

References

Alba, J. W., Broniarczyk, S. M., Shimp, T. A., & Urbany, J. E. (1994). The influence of prior beliefs, frequency cues, and magnitude cues on consumers' perceptions of comparative price data. *Journal of Consumer Research, 21,* 219-235.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York, NY: Academic Press.

Anderson, N. H. (1996). *A Functional Theory of Cognition*. Mahwah, New Jersey: Erlbaum Associates.

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition, 44*, 75-106.

Baddeley, A. (1992). Working memory. *Science*, *255*, 556-559.

Baltes, P. B., & Baltes, M. M. (1990). *Successful aging: Perspectives from the behavioral sciences.* Cambridge, England: European Science Foundation.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (R package version 1.1-7). Retrieved from http://CRAN.R-project.org/package=lme4

Bellos, A. (2010). *Alex's adventures in numberland*. London, England: Bloomsbury.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica, 22*, 23-36.

Block, L. G., & Morwitz, V. G. (1999). Shopping lists as an external memory aid for grocery shopping: Influences on list writing and list fulfillment. *Journal of Consumer Psychology, 8*, 343-375.

Brezis, N., Bronfman, Z. Z., & Usher, M. (2015). Adaptive spontaneous transitions between two mechanisms of numerical averaging. *Scientific reports, 5*, 10415.

Bronfman, Z. Z., Brezis, N., Moran, R., Tsetsos, K., Donner, T., & Usher, M. (2015). Decisions reduce sensitivity to subsequent information. *Proceedings of the Royal Society B, 282*. 20150228.

Brown, N. R., & Siegler, R. S. (1992). The role of availability in the estimation of national populations. *Memory & Cognition, 20*, 406-412.

Chandon, P., & Wansink, B. (2007). The biasing health halos of fast-food restaurant health claims: lower calorie estimates and higher side-dish consumption intentions. *Journal of Consumer Research, 34*, 301-314.

Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard & Y. Rossetti (Eds.), *Attention and Performance XXII. Sensori-motor foundations of higher cognition*. Cambridge, Mass: Harvard University Press.

Dehaene, S. (2009). Origins of mathematical intuitions. *Annals of the New York Academy of Sciences, 1156*, 232-259.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics* (2nd ed). New York, NY: Oxford University Press.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition, 43*, 1-29.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284*, 970-974.

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science, 320*, 1217-1220.

Erdelyi, M. H., & Goldberg, B. (2014). Let's not sweep repression under the rug: Toward a cognitive psychology of repression. In J. F. Kihlstrom & F. J. Evans (Eds.), *Functional disorders of memory* (pp. 355-402). New York, NY: Psychology Press.

Gandini, D., Lemaire, P., & Dufau, S. (2008). Older and younger adults' strategies in approximate quantification. *Acta Psychologica, 129*, 175-189.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological review, 98(2)*, 254-267.

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making, 9*, 1-14.

Heath, C., & Soll, J. B. (1996). Mental budgeting and consumer decisions. *Journal of Consumer Research, 23*, 40-52.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*, 534-539.

Hogarth, R. M., & Einhorn, H. J.(1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1-55.

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin, 140,* 339-373.

Hutchinson, J., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: Can a generalist adapt its rules to dispersal of items across patches? *Animal Behaviour, 75*, 1331-1349.

Jones, T., & Oberauer, K. (2013). Serial-position effects for items and relations in short-term memory. *Memory, 21(3)*, 347-365.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-292.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American psychologist, 39*, 341-350.

Lemaire, P., Arnaud, P., & Lecacheur, M. (2004). Adults' age-related differences in adaptivity of strategy choices: Evidence from computational estimation. *Psychology and Aging, 19*, 467-481.

Longo, M. R., & Lourenco, S. F. (2007). Spatial attention and the mental number line: Evidence for characteristic biases and compression. *Neuropsychologia, 45*, 1400-1407.

Malmi, R. A., & Samson, D. J. (1983). Intuitive averaging of categorized numerical stimuli. *Journal of Verbal Learning and Verbal Behavior, 22(5)*, 547-559.

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual review of neuroscience, 32*, 185-208.

Niedrich, R. W., Weathers, D., Hill, R. C., & Bell, D. R. (2009). Specifying price judgments with range–frequency theory in models of brand choice. *Journal of Marketing Research, 46*, 693-702.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, *72*, 407-418.

Parducci, A., Thaler, H., & Anderson, N. H. (1968). Stimulus averaging and the context for judgment. *Perception & Psychophysics, 3*, 145-155.

Prelec, D., & Simester, D. (2001). Always leave home without it: A further investigation of the credit-card effect on willingness to pay. *Marketing Letters, 12*, 5-12.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225-237.

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review, 22,* 391-407.

Stan Development Team (2015). Stan: A C++ Library for Probability and Sampling, Version 2.8.0. http://mc-stan.org/

Stevens, S. (1957). On the psychophysical law. *Psychological Review, 64*, 153-181.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology, 62*, 1041-1062.

Stewart, N., Chater, N., & Brown, D. A. G. (2006). Decision by sampling. *Cognitive Psychology, 53*, 1-26.

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences, 109(24)*, 9659-9664.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Van Ittersum, K., Pennings, J. M. E., & Wansink, B. (2010). Trying harder and doing worse: How grocery shoppers track in-store spending. *Journal of Marketing, 74*, 90-104.

Vehtari, A., Gelman, A., & Gabry, J. (2015). Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models. *arXiv preprint arXiv:1507.04544*.

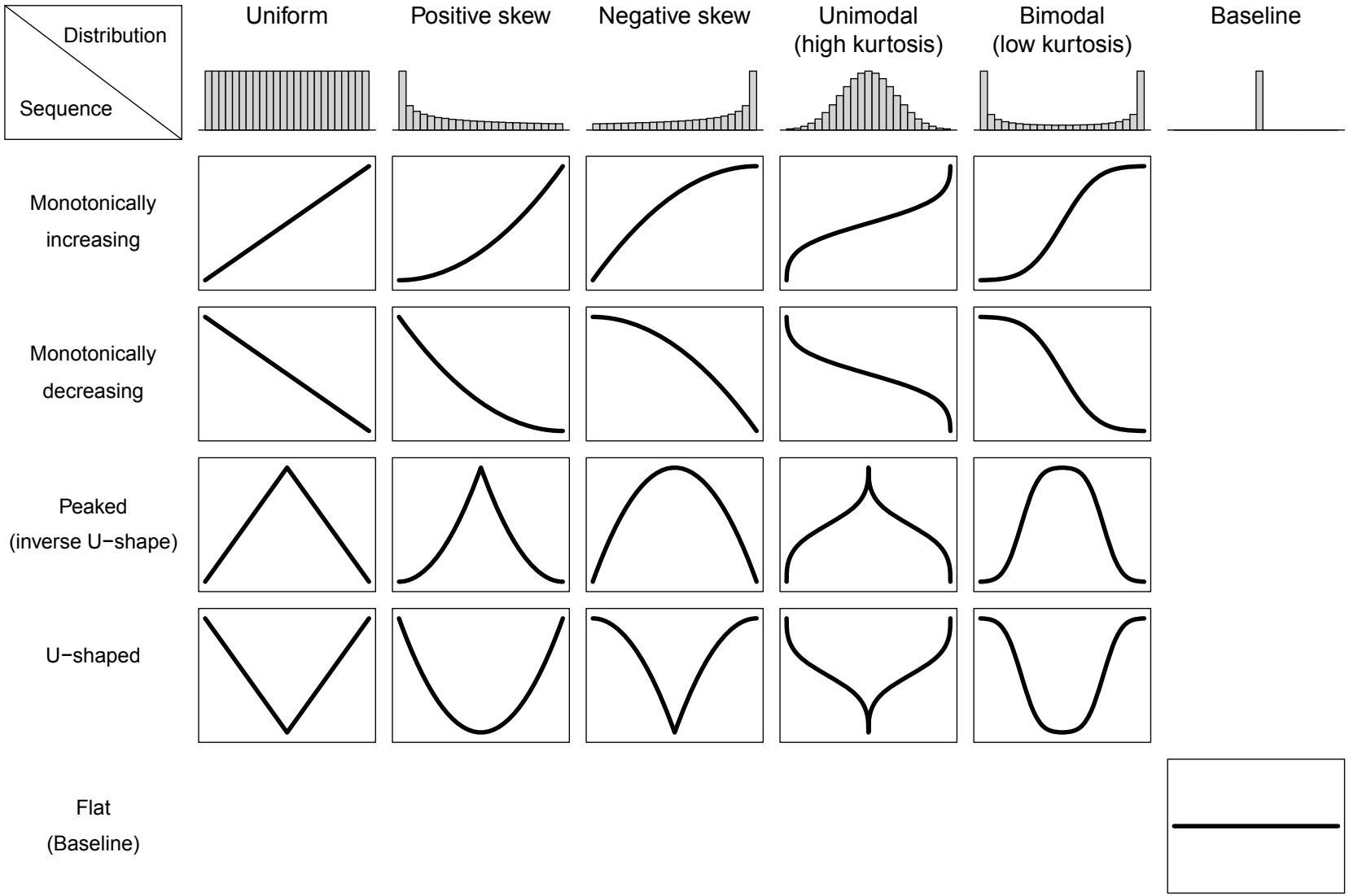
Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: a neural model. *Journal of Cognitive Neuroscience, 16*, 1493-1504.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely application

information criterion in singular learning theory. *Journal of Machine Learning Research 11*, 3571-3594.

Zeigenfuse, M. D., Pleskac, T. J., & Liu, J. (2014). Rapid decisions from experience. *Cognition, 131*, 181-194.

Appendix A: Overview of the presented sequences



Note. The baseline condition was characterized by a flat sequence and hence its frequency distribution consisted of only a single peak.