Are judgments a form of data clustering? Reexamining contrast effects with the k-means algorithm

Eric Boillaud, Guylaine Molina

Are Judgments a Form of Data Clustering? Reexamining Contrast Effects With the K-Means Algorithm

Eric Boillaud

Guylaine Molina

Aix Marseille Université, ENS Lyon, ADEF EA 4671, 13248, Marseille, France

Author Note

Eric Boillaud and Guylaine Molina, Aix Marseille Université, ENS Lyon, ADEF EA 4671, 13248, Marseille, France.

Many of the ideas expressed here have their origins in discussions and arguments with Jean-Marc Fabre, to whom we are deeply indebted. We would also like to thank Allen Parducci and Gert Haubensak, whose comments on earlier versions of our model helped us refine our assumptions.

Correspondence concerning this article should be addressed to Guylaine Molina, Aix Marseille Université, Laboratoire Apprentissages, Didactiques, Evaluation, Formation (ADEF EA 4671), 32, rue Eugène Cas, 13248 Marseille Cedex 4, France. Email: guylaine.molina@univ-amu.fr

Abstract

A number of theories have been proposed to explain in precise mathematical terms how statistical parameters and sequential properties of stimulus distributions affect category ratings. Various contextual factors such as the mean, the midrange, and the median of the stimuli, the stimulus range, the percentile rank of each stimulus, and the order of appearance have been assumed to influence judgmental contrast. A data clustering reinterpretation of judgmental relativity is offered, wherein the influence of the initial choice of centroids on judgmental contrast involves two combined frequency and consistency tendencies. Accounts of the k-means algorithm are provided, showing good agreement with effects observed on multiple distribution shapes, and with a variety of interaction effects relating to the number of stimuli, the number of response categories, and the method of skewing. Experiment 1 demonstrates that centroids initialization accounts for contrast effects obtained with stretched distributions. Experiment 2 demonstrates that the iterative convergence inherent to the k-means algorithm accounts for the contrast reduction observed across repeated blocks of trials. The concept of within-cluster variance minimization is discussed, as well as the applicability of a backward k-means calculation method for inferring, from empirical data, the values of the centroids that would serve as a representation of the judgmental context.

Keywords: judgment, contrast, context, k-means, clustering

Are Judgments a Form of Data Clustering? Reexamining Contrast Effects With the K-Means Algorithm

Category ratings are widely used in cognitive and social research, in opinion surveys, and in consumer reviews, probably because they reflect the way people make value judgments in everyday life. Due to their methodological flexibility and ease of implementation, rating scales are amongst the most widespread tools to collect indications of how stimuli varying along one or several dimensions are perceived. As opposed to absolute identification tasks, in which participants are required to correctly identify stimuli drawn from a set of items by using clearly predefined labels, category scaling typically requires assigning a few category levels to numerous stimuli without any instructions as to what would be a right or wrong response (Stewart, Brown, & Chater, 2005). Correlatively, by enabling subjective assessment of stimuli in systematically manipulated contexts, rating scales largely contributed to isolating mathematical relations between the mean response assigned to any particular stimulus on one hand, and different contextual properties on the other. In the field of judgmental relativity, category ratings are known to be highly sensitive to the frequency distribution. The same stimuli are rated higher when the distribution is positively skewed than when the distribution is negatively skewed, even though both distributions have the same range (e.g., Parducci, 1956). For example, squares receive higher ratings when the smaller sizes are presented more frequently than the larger sizes. Here we examine a k-means clustering reinterpretation of judgmental relativity as a way to account for those contrast effects in category ratings.

Helson's (1948) adaptation-level theory represents one of the earliest attempts at a quantitative framework capable of accounting for such context effects. Helson postulated that the perceptual value of any stimulus is determined by its relation to a prevailing, internal reference

value, the so-called level of adaptation, which acts as a dynamic, background function encompassing the influence of all current and past stimuli of a similar nature. The adaptation level, by combining the effects of present and past experience, serves as a constantly changing baseline to which comparisons are made for assessing the current stimulus. Stimuli of a magnitude close to the adaptation level are judged as neutral. Stimuli occurring above or below the adaptation level are judged as being of positive or negative magnitude, respectively. The adaptation level was initially conceived as the result of an implicit averaging process operating on relevant stimuli, and was practically defined as a weighted logarithmic mean of the current and previously presented stimuli. Multiple experiments were conducted in the 1950s to characterize, in terms of nearness, recency, and salience, the weighting function intended to reflect this averaging process.

Experiments based on systematic control of central tendency indicators in the early 1960s (Parducci, Calfee, Marshall, & Davidson, 1960) demonstrated that manipulating either the midpoint or median of a set of stimuli affected the judgment scale, while manipulating the mean did not. This discovery proved decisive in the emergence of a new framework, the range-frequency theory (RFT), which explained contrast effects as the result of an integration process involving the full contextual series of stimuli, rather than as the manifestation of a comparison process involving one single reference value.

The essential idea of RFT (Parducci, 1965) is that the judgment of any particular stimulus represents a compromise between two principles of judgment: (a) The range of contextual stimuli is divided into as many equal subranges as response categories (the range principle), and (b) the same number of contextual stimuli are assigned to each response category (the frequency principle).

The judgment J_{ic} of Stimulus i in Context c is a weighted average of two values:

, (1)

where R_{ic} is the range value of Stimulus i in Context c (what its judgment would have been if categories actually divided the contextual range into equal subranges); F_{ic} is the frequency value of the same stimulus (what its judgment would have been if an equal number of contextual stimuli were assigned to each category); and w is the weighting parameter describing the compromise between the two principles of judgment.

The range value of Stimulus i in Context c is determined by the proportion of the contextual range lying below that stimulus:

$$R_{ic} = (S_i - S_{\min}) / (S_{\max} - S_{\min}), \quad (2)$$

where S_i is the subjective value of Stimulus i, and S_{min} and S_{max} are the subjective endpoints of the range.

Because the range values are highly dependent on the subjective endpoints, and because the subjective endpoints may differ from the two most extreme stimuli, R_{ic} is generally inferred from the observed responses instead of being calculated beforehand.

The frequency value of Stimulus i in Context c, directly calculable from the stimulus distribution, is determined by the proportion of stimuli lying below that stimulus:

$$F_{ic} = (r_{ic} - 1) / (n - 1), \quad (3)$$

where r_{ic} is the rank of Stimulus i in Context c.

All values (J_{ic}, R_{ic}, F_{ic}, and w) are expressed on an abstract scale ranging from 0 to 1. The mean response to Stimulus i in Context c is obtained by linear transformation of J_{ic}:

, (4)

where k is the number of categories in the response scale.

In the 1970s and 1980s, RFT met undisputed success in establishing clear indications of contextual effects pertaining to perceptual and social judgments. RFT has proven to be a highly robust framework for describing effects obtained with rating scales for a variety of stimuli: lifted weights, numerousness of dots, sizes of squares, lengths of lines, pleasantness of facial expressions (e.g., Fabre, 1993), perception of fair grading (Wedell, Parducci, & Roman, 1989), estimation of comfortable temperatures (Molina & Fabre, 1999), and ratings of test scores (Molina & Fabre, 2000, 2001). RFT also aided in demonstrating the more general principle of context dependency in other types of judgmental tasks: body perception (Wedell, Santoyo, & Pettibone, 2005), judgment of athletes' performance (Damisch, Mussweiler, & Plessner, 2006; Fasold, Memmert, & Unkelbach, 2013), hedonic preference and contrast (Cogan, Parker, & Zellner, 2013; Zellner, Mattingly, & Parker, 2009), hedonic ratings of paintings (Zellner et al., 2010), judgment of price (Matthews & Stewart, 2009), loudness estimations (Parker, Moore, Bahraini, Gunthert, & Zellner, 2012), tempo and pleasantness judgments (Rashotte & Wedell, 2012), comparative optimism (Milhabet, Le Barbenchon, Molina, Cambon, & Steiner, 2012), pain perception (Watkinson, Wood, Lloyd, & Brown, 2013), duration perception (Matthews, Stewart, & Wearden, 2011; Penney, Brown, & Wong, 2013), and randomness judgments (Matthews, 2013). The value of *w* has been proven to be affected by significant variations, depending on how instructions and presentation factors emphasize either the relationship between the stimuli and the endpoints of the contextual set, or the relative frequencies or spacings of stimuli.

In the early 1990s, Haubensak reinterpreted the frequency effect as the concomitant manifestation of an initial central tendency on one hand, consisting in assigning the middle categories to the first presentations, and of a consistency tendency on the other, consisting in

persistently reassigning, throughout the entire experiment, the same categories to the same stimuli (Haubensak, 1992a, 1992b; Tommasi, 2001). The so-called consistency model postulates the prime importance of stimuli presented early in the sequence, with the correlated assumption that more frequent stimuli have a higher probability of occurring earlier. Practically, the consistency model relies on four assumptions partially inspired from the LS-2 model of Atkinson and Crothers (1963). According to the LS-2 model, each newly presented stimulus has a probability *b* of being stored in long-term memory (LTM) along with the response for later use as a standard. If not stored in LTM, the stimulus-response pair still enters short-term memory (STM). At each trial, every stimulus in STM has a probability *f* of being forgotten and a probability $1 - f$ of remaining in memory storage. Assumption 1 of the model, which is a variation of Parducci's range principle, states that each stimulus is judged relative to the subjective value of the standard immediately higher and lower, if there is any. Assumption 2 of the model, which invokes the central tendency, states that, at the start of the task, categories closer to the center of the scale are likely to be selected so as to leave room on both sides for future judgments. Assumption 3 of the model states that stimuli do not enter STM a second time unless they are forgotten. Assumption 4 of the model, which accounts for scale development along the course of the task, states that if a new stimulus matches or even exceeds the highest or the lowest of the current standards, the response is switched to a higher or lower category, if available. With these few assumptions, Haubensak managed in the early 1990s to reproduce most of the known context effects observed so far using the successive presentation method. Subsequent efforts by Haubensak in the field of judgmental relativity were directed towards experimentally decoupling the respective effects of the order of appearance and of the stimulus frequencies (e.g., Haubensak & Petzold, 2003).

These three theoretical approaches emphasize that rating scales can be regarded as context-dependent interval scales (i.e., that the rank difference between the response categories assigned to two stimuli reflects the perceived degree of difference between those two stimuli in a particular context). In essence, these approaches have been developed as measurement theories, in the sense that they share the same underlying premise that any response has a quantitative relation to other responses along a phenomenal continuum. In this paper, we advance the radically different idea that judgment is a clustering partitional process consisting in grouping what is perceived as similar in the same response categories, and in distinguishing what is perceived as distinct by use of different response categories. Judgment can be interpreted as a natural case of unsupervised classification, driven by the objective of expressing hidden structures in unlabeled stimuli by means of response categories. This theoretical premise is developed in the following section by reference to the mechanics of the k-means algorithm, an automated classification approach that has laid the foundation for unsupervised learning.

Data Clustering: A K-Means-Oriented Framework

The k-means algorithm (MacQueen, 1966) has been largely used in the field of data clustering, and, while involving costly iterative calculation, has spawned numerous machine learning applications based on continuously increasing storage capacities and computer processing power. In its standard form, it can be applied to unidimensional or multidimensional numeric data, and with few modifications, to multi-attribute categorical data. The k-means algorithm enables the classification of a given set of n data points into k clusters, based on a procedure composed of two main features: centroids initialization on one hand, and assignment-update iterations on the other. Centroids initialization involves defining a set of k starting centroids, one for each cluster. One of the commonly used initialization methods is the Forgy

method, wherein *k* data points are randomly drawn from the data set and used as starting centroids. The assignment-update iterations consist in assigning each data point to the cluster with the closest centroid, and in calculating the new means to be the centroids of the data points in the new clusters. Final convergence is reached when the assignments no longer change or when the objective function, *J*, measuring the within-cluster sum of squares, is lower than a predetermined threshold, or after a predetermined number of iterations:

, (5)

where *J* is the sum of the Euclidean distances between each data point x_i and the closest centroid m_j.

The vanilla Forgy k-means algorithm relies on the following computation rules:

- Rule 1 (initialization): Randomly draw *k* distinct data points as starting centroids. The initial set of centroids is denoted $c_1, \ldots, c_k$.
- Rule 2 (assignment): Assign each data point x_i to the cluster that has the closest centroid.

. (6)

Here, C_j is the cluster of centroid m_j. Ties must be broken consistently (i.e., always to the lowest centroid, or always to the highest centroid), in order to prevent the algorithm from cycling through non-convergent loops.

- Rule 3 (update): Recalculate the new value of m_j as the mean of all data points assigned to it.

. (7)

- Rule 4: Repeat Rule 2 and 3 until the centroids no longer change, or until *J* is lower than a predetermined threshold, or for a predetermined number of iterations.

There is no guarantee that the algorithm will converge to the global optimum in terms of within-cluster variance minimization, as the result depends on the starting centroids. In recent years, much research effort has been devoted to developing better initialization meta-algorithms (e.g., Chen & Shixiong, 2009; El Agha, 2012; Kanungo et al., 2002). In practice, the algorithm is run multiple times with various starting conditions.

Table 1 illustrates how the k-means algorithm iteratively addresses a typical classification problem. Schematically, 18 numbers equally distributed between 108 and 992 are to be iteratively grouped into nine clusters ranging from A to I. In this example, the initial centroids set is biased as if positively skewed, and consists of nine low-value centroids: 108, 160, 212, 264, 316, 368, 420, 472, and 524. Theoretically, a skewed sample has the same probability of being drawn with the Forgy Method as any other similarly sized sample drawn from the same distribution. However, this particular starting centroids set is extreme due to the very low probability of not drawing *at least* one number greater than 524. Such a combination of data and centroids provides an opportunity to understand the algorithm's properties. Table 1 shows that stabilization is reached after eight iterations, with the following final centroid values: 108, 160, 212, 264, 316, 394, 524, 706, and 914. From initialization to final convergence, assignment-update iterations induce a global shift to higher centroid values. The shift is drastic for Clusters H and I, moderate for Clusters F and G, and null for Clusters A to E. This illustrates how the k-means algorithm minimizes the within-cluster variance conditionally to the centroids initialization, the value of *J* decreasing by 85% during the process with this particular starting

centroids set. For global minimization purposes, multiple repetitions of the algorithm with random starting centroids sets would increase the chances of obtaining a lower value of *J* (i.e., a better partition in terms of within-cluster variance).

Because the phenomena occurring at the centroids level reflect symmetrically on the response scale, the mechanics of the k-means algorithm can be transposed to judgmental issues by replacing the terms *data point* with *stimulus*, and *cluster* with *category*. Table 1 shows that, at the end of the iterative refinement process, the seventh category of the response scale, G, is assigned to the two stimuli closest to the numerical midrange, 524 and 576. Alternatively, dividing the numerical range into nine equal-length intervals would lead to the fifth category, E, being assigned to both these stimuli. From a clustering perspective, two conclusions may be drawn concerning contrast effects: Contrast is caused entirely by the skewing of the initial set of centroids, and subsequent refinement iterations contribute to reducing its magnitude.

As illustrated in the above example using evenly-spaced stimuli, the k-means properties provide an enlightening framework capable of accounting for judgmental contrast. These properties will now be examined in detail to support reinterpretation of known effects pertaining to skewed distributions. Our underlying assumption is that the stimulus distribution influences the initial choice of centroids under the form of two combined frequency and consistency principles.

Accounts of the K-Means Framework

In this section, we provide k-means accounts for judgmental contrast observed on multiple distribution shapes, and for a variety of interaction effects pertaining to the number of stimuli, the number of response categories, and the method of skewing. In the following simulations, the influence of the initial choice of centroids on judgmental contrast was

formalized by means of random sampling without replacement, based on two assumptions: a frequency-driven random sampling principle, and a distance threshold-based consistency principle. The frequency-driven random sampling principle states that the values of the starting centroids are *k'* stimuli randomly drawn from the distribution of the *n* stimuli, with each stimulus having a probability to be drawn depending directly on its frequency. The distance threshold-based consistency principle states that the initial Euclidean distance separating two neighbor centroids tends to be greater than a minimum value, referred to as the consistency threshold δ. The former principle tends to support assignment of equal quantities of stimuli to the different response categories, while the latter tends to support assignment of the same response categories to similar stimuli. To account for the observed tendency to not use all the response categories, *k'* was assumed to be lower than or equal to *k*, *k* being the total number of categories in the response scale. Practically, the centroids selection relied on the following steps:

- Step 1: Sort the *n* stimuli in random order (in the random-sorted set, the stimuli are denoted s_ε with $\varepsilon = 1$ to *n*), set the starting *k'* value to 0, and go to Step 2.
- Step 2: Select s_1 as a centroid, set the new value of *k'* to 1, set the value of ε to 2, and go to Step 3.
- Step 3: For $\varepsilon = 2$ to *n*, move down the random-sorted set of stimuli s_ε:
    - Step 3.1: If the Euclidean distances between s_ε and each previously selected centroid are all higher than or equal to δ, select s_ε as a centroid, set the new value of *k'* to *k'* + 1, set the new value of ε to $\varepsilon + 1$, and go to Step 3.3. Otherwise, go to Step 3.2.
    - Step 3.2: With probability *p*, reject s_ε, set the new value of ε to $\varepsilon + 1$, and go to Step 3.3. Otherwise, select s_ε as a centroid, set the new value of *k'* to

k' + 1, set the new value of ε to ε + 1, and go to Step 3.3.

- Step 3.3: If $\varepsilon = n$ or $k' = k$, go to Step 4. Otherwise, go back to Step 3.

- Step 4: If $k' <$ k, randomly remove k'' categories from the response scale (with $k'' = k - k'$), and go to Step 5.
- Step 5: Assign the k' centroids to the k' categories monotonically.

The rest of the algorithm followed the vanilla k-means procedure, with as many iterations as required for the centroids to reach complete stability (i.e., for the within-cluster sum of squares J to stop decreasing). The higher δ and p, the greater the weight of the consistency principle. The accounts of the k-means framework presented below were computed with 40 repetitions per condition, and fitted to empirical data by following the least squares method.

Judgmental Contrast With Multiple Distribution Shapes

Boillaud's (1997) Experiment 1 consisted of two separate factorial designs, one for each type of distribution, i.e., asymmetric and symmetric. In each factorial design, the distribution was a between-subjects factor. Participants were required to judge lines according to their length. The entire set of lines was arranged in a single column on a 21.0 x 29.7-cm sheet of paper. The lines were center-aligned and presented in random order. The stimulus set consisted of 21 lines whose length varied from 8.00 to 58.00 mm, with a width of 1 mm. Five lines were used in every condition: 8.00, 18.00, 30.00, 44.00, and 58.00 mm. No line was repeated exactly; instead, repetitions consisted of adding lines of very similar values (every line in a group was within 1 mm of another line in the same group). Two distributions were used in the asymmetric conditions: positively skewed (most of the lines had a value between 8.00 and 30.00 mm), and negatively skewed (most of the lines had a value between 30.00 and 58.00 mm). Two distributions were used in the symmetric conditions: unimodal (most of the lines had a value

between 18.00 and 44.00), and bimodal (most of the lines had a value between 8.00 and 18.00 mm, and between 44.00 and 58.00 mm). A six-response scale was used, ranging from 1 (*very small*) to 6 (*very large*).

Results are shown in Figure 1. The comparison of responses obtained in the two asymmetric conditions revealed a single-direction effect: The lines of 18.00, 30.00, and 44.00 mm all received higher ratings when presented in a positively skewed distribution than in a negatively skewed distribution. The comparison between the two symmetric conditions revealed a dual-direction effect: The line of 18.00 mm received higher ratings when presented in the bimodal distribution than in the unimodal distribution, while the line of 44.00 mm received higher ratings in the unimodal distribution than in the bimodal distribution. In summary, the response curves reflected the densities of stimuli in the distributions. Accounts of the k-means framework are in good agreement with these results, the best fits being obtained with a psychophysical transformation of the stimulus values based on a power function of exponent 0.72, and with $\delta = 1.25$ and $p = .95$ in all conditions. The response pattern obtained with the bimodal distribution is particularly interesting. The k-means algorithm is known, from a data clustering standpoint, to perform poorly on bimodal and non-convex multidimensional distributions (e.g., Sharma, Singh, & Gupta, 2013). In the case of a bimodal distribution, the large central gap in the stimulus distribution makes it impossible to draw initial centroids in the middle of the range. The frequency-driven random sampling principle pulls the initial centroids towards the endpoints of the range, but because the high densities of stimuli in those regions generate conflicts between this principle and the distance threshold-based consistency principle, Step 3.2 of the centroids selection causes k' to be lower than k. In other words, the k-means algorithm skips categories as though it was saving the missing categories for stimuli in the part

of the range that had not been sampled. Whilst not optimal in terms of variance minimization, this pattern accurately mimics what is typically observed from participants.

Parducci's frequency principle correctly accounts for judgmental contrast, but the literal interpretation of Equation 3 requires accepting the implication that any given stimulus is judged in relation to the series of stimuli as a whole (Petzold & Haubensak, 2001). This raises the question of how such a contextual integration could be plausibly managed without exceeding the limited capacity of working memory. From a k-means perspective, contrast can be interpreted as a phenomenon occurring at the early stage of scale development, a statistical consequence of rapid frequency-driven random sampling. From such a perspective, the initial choice of centroids is influenced by the distribution of the stimuli: The higher the density of stimuli in a given region of the range, the greater the probability for a stimulus of this region to be sampled as a centroid. While a frequency-driven sampling process might lead to the neglect of low-density regions at the initialization stage, subsequent iterations will converge to an acceptable partition in terms of within-category variance minimization, as long as the number of response categories remains low compared to the number of stimuli. The question of how the number of stimuli and the number of response categories affect contrast when the distributions are skewed by manipulating the frequencies is addressed in the following experiment.

Stimulus and Category Effects With Frequency-Skewed Distributions

Parducci and Wedell's (1986) Experiment 4B consisted of a factorial design involving three between-subjects factors: (a) Skewing of Distribution (positive skewing versus negative skewing), (b) Number of Stimulus Groups (five versus nine), and (c) Number of Categories (three versus nine). Participants were required to judge numbers according to their magnitude. The entire set of numbers was arranged from smallest to largest in a single column on an 8.5 x

11.0-inch sheet of paper. The stimulus set consisted of 25 numbers whose magnitude varied from 108 to 992. Five stimulus groups were used in every condition: 108, 329, 550, 771, and 992. In the nine-group conditions, four stimulus groups were added: 219, 439, 661, and 881. To avoid repeating the same number, repetitions of a group consisted of numbers of similar values (every number in a group was within one unit of any other in the same group). For example, the positively skewed, five-group set consisted of the following numbers: 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 326, 327, 328, 329, 330, 331, 332, 548, 549, 550, 551, 770, 771, 991, and 992.

Results are shown in Figure 2. The interaction between the skewing and the number of stimuli, the stimulus effect, led to a smaller difference between the two skewed conditions with five stimulus groups than with nine stimulus groups. The interaction between the skewing and the number of categories, also known as the category effect, led to a smaller difference between the two skewed conditions with nine categories than with five categories. Accounts of the k-means framework are consistent with these results, the best fits being obtained with $\delta = 5$ and $p = .50$ in all conditions. For the nine-category conditions, the effective response scale ranged from the second category to the ninth category, the first category being almost never used by the participants. Parducci and Wedell (1986) observed similar stimulus and category effects with different types of stimuli and experimental designs, most of their results being obtained in successive presentation. To explain these effects, they developed an elaborated version of RFT, the retrieval-consistency model, based on a reinterpretation of Shannon and Weaver's (1963) mathematical theory of communication. According to this model, participants tend to assign the same response category to all repetitions of the same group of stimuli (consistency principle) while assigning equal numbers of stimuli to each category (frequency principle). When there are

a large number of stimulus groups relative to the number of categories, the consistency principle can be applied without conflict with the frequency principle. When there are only a few stimulus groups relative to the number of categories, applying the consistency principle requires violating the frequency principle. To quantitatively account for both stimulus and category effects, the retrieval-consistency model states that the retrieved frequencies are flattened by means of a scale-dependent memory threshold operating on the last *t* trials. This retrieval bias leads to the percentile ranks being calculated using the retrieved stimulus frequencies instead of the real presentation frequencies. In practice, the retrieved presentations of a stimulus are not counted beyond the frequency of use of each category as dictated by the frequency principle, which precludes assignment of more than one category to each stimulus.

The plausibility of such a retrieval mechanism is questionable, as it suggests that participants manage to maximize the quantity of transmitted information by means, paradoxically, of an incidental cognitive bias. From a k-means perspective, both stimulus and category effects can be reinterpreted by emphasizing how consistency constraints rapidly affect the starting centroids in the early stage of scale development in reference to the selection loop specified in Step 3. According to the distance threshold-based consistency principle, two centroids can hardly be drawn within a subrange lower than the consistency threshold δ. In the five-group conditions, the probability of two stimuli belonging in the same group is .27. In the nine-group conditions, this probability is .16. Thus, for a given δ close to the mean within-group range, the probability of rejecting similar stimuli during the centroids selection step is higher in the five-group conditions than in the nine-group conditions. Stimuli of low-density regions are more likely to be selected as second-choice centroids in the former case than in the latter, which explains the stimulus effect. On the other hand, since consistency relies on a sequential non-

replacement procedure, the probability of rejecting a stimulus is higher when drawing more centroids, which explains the category effect. Hence, the category effect can be interpreted as a consequence of how the distance threshold-based consistency principle affects the centroids selection when the probability of two stimuli belonging in the same group is non-null. This is the case when the distributions are skewed by manipulating the frequencies. The question of how spacing-skewed distributions, wherein this probability is null, affect the category effect is addressed in the following experiment.

Absence of Category Effect With Spacing-Skewed Distributions

Parducci and Wedell's (1986) Experiment 4C used a factorial design involving two between-subjects factors: Skewing of Distribution (positive skewing versus negative skewing), and Number of Categories (three, five, nine, and 100). Participants were required to judge dot patterns according to their darkness. Each pattern consisted of solid, 1 mm-diameter black dots, scattered irregularly within a 25 mm-side square. The entire set of dot patterns was presented on an 8.5 x 11.0-inch sheet of paper in a random arrangement. The stimulus set consisted of 11 patterns ranging from 12 to 90 dots. Six patterns (with 12, 18, 27, 40, 60, and 90 dots, respectively) were common to both sets. For the positive sets, five low-density patterns were added (with 14, 15, 16, 21, and 23 dots, respectively). For the positive sets, five high-density patterns were added (with 47, 51, 70, 74, and 77 dots, respectively). Because each pattern occurred only once, skewing was manipulated by variation in the spacing of stimulus values.

Results are shown in Figure 3. In spite of some variability in the effect of skewing, the data differ from the systematic decline observed in Parducci and Wedell's (1986) Experiment 4B. The contrast magnitude proved constant with three and 100 categories, and a large effect of skewing was observed in all conditions. Accounts of the k-means framework are consistent with

these results, the best fits being obtained with a psychophysical transformation of the stimulus values based on a logarithmic function, with $\delta = 0.15$ and $p = .80$ in all conditions. To improve the slope of the response curves in the 100-category condition, k'' categories were randomly removed from the response scale as in all previous simulations, but the categories of rank 1 and 100 were always conserved. Both the retrieval-consistency model and our k-means framework explain the absence of a category effect with spacing-skewed distributions as a consequence of all stimuli being clearly perceived as different. The retrieval-consistency model states that densities retrieved in memory are biased when participants are confronted with frequency-skewed distributions, and unbiased when participants are confronted with spacing-skewed distributions, which raises questions about the plausibility of the model's assumptions. Alternatively, our k-means framework asserts that at the early stage of scale development, the tendency to sample centroids based on stimulus densities conflicts with the consistency tendency in the case of frequency-skewed distributions (as explained previously), but not in the case of spacing-skewed distributions. In the latter case, two randomly drawn centroids never belong to the same stimulus group, which mechanically prevents rejection of any previously selected centroid. In such a condition, the number of categories has no effect on how the centroids set is sampled.

In summary, the contrast and interaction effects presented above can be interpreted in reference to the initial centroids selection, which is the first of the two main features of the k-means algorithm. The contrast magnitude results from a combination of two tendencies. These are a tendency to reflect the distribution densities in the sampled centroids set, and a tendency to choose the centroids set so as to support consistent assignment of the same response categories to similar stimuli. Both these tendencies are implemented in the early stage of scale development

by means of rapid sampling operations. The following experiment demonstrates that centroids initialization, under the influence of these two combined tendencies, accounts for contrast effects obtained with stretched distributions. The second of the two main features of the k-means algorithm involves subsequent assignment-update iterations. Its properties in terms of judgmental contrast will be examined in Experiment 2.

Experiment 1: Judgmental Contrast With Stretched Sets

The objective of this experiment was to demonstrate that judgmental contrast still occurs when keeping invariant the midrange, the mean, the median, the extrema, and the frequency values, and when controlling the order of stimuli with the method of simultaneous presentation. An alternative stretching method was used to manipulate the stimulus densities so as to produce different conflict levels between the frequency and the consistency tendencies in the lower and upper regions of the range. Our k-means framework predicts that the frequency tendency should result in drawing the starting centroids in the regions of the range where it does not conflict with the consistency tendency. To test this prediction, we stretched the distributions so that the number of non-repeated stimuli (i.e., the number of stimuli that differ by more than one unit from any other) varied between the lower and upper regions of the range without affecting the central tendency indicators and the ranks of the stimuli. We expected a larger contrast effect in the regions of the range containing the larger number of non-repeated stimuli.

Method

Participants. All participants were students enrolled in human science at the University of Aix-Marseille, Aix-en-Provence. Twenty participants served in each condition.

Design and stimuli. There was one between-subjects factor (Skewing of Distribution) with two levels (positively stretched set versus negatively stretched set), and one within-subject

factor (Stimulus Value) with five levels (corresponding to the five number values used in both distributions: 108, 329, 550, 771, and 992). Each participant was presented with 25 numbers. Stimuli are shown in Table 2. In both sets, the numbers varied from 108 to 992, in order to keep the range invariant. In both sets, the median, the midrange, and the mean were equal to 550, and the *Fi* values of the five common stimuli were kept invariant (i.e., each of the five common stimuli had the same rank in both distributions). The two distributions differed only in the number of non-repeated stimuli along the range. For example, in the positively stretched distribution, there were thirteen non-repeated stimuli between 108 and 550, and three non-repeated stimuli between 550 and 992. The stimuli were presented in one single column on a sheet of paper in randomized order, and each participant received a unique order.

Procedure and instructions. Participants were each given a sheet of paper displaying instructions and numbers. Their task was to rate each number in accordance with how large or small it appeared in comparison with the other numbers. They were to write a response next to each of the numbers, consisting in a letter ranging from A (*very small*) to F (*very large*).

Results and Discussion

Participants whose responses showed one or several of the following anomalies were discarded: (a) One or several stimuli had no assigned response, (b) one or several responses fell outside of the scale, and (c) the response ranks did not monotonically follow the stimulus ranks. In each group, three participants' data were discarded, leaving the ANOVA balanced.

Figure 4 shows results for the five common numbers. These results demonstrate that the contrast effect is larger in the regions of the range containing the larger number of non-repeated stimuli. A two-way 2 (Skewing of Distribution) x 5 (Stimulus Value) mixed-design ANOVA revealed that all effects were significant: Skewing of Distribution, $F(1, 32) = 4.57$, $p < .05$;

Stimulus Value, $F(1, 32) = 445.66$, $p < .0001$; and Skewing of Distribution x Stimulus Value interaction, $F(4, 128) = 2.85$, $p < .05$. Because all three central tendency indicators were kept invariant, i.e., the median, the midrange, and the mean, the results cannot be explained by adaptation-level theory. They cannot be explained by RFT either, at least not in its canonical form, because the range and the frequency values of all five common stimuli were kept invariant. As for the consistency model, even if the participants happened to assess the stimuli sequentially (which is theoretically possible, if not probable), the control of all of the above-mentioned parameters and the randomization of the presentation order would make the internal processing sequence independent of the distribution shape. Therefore, the results cannot be explained by a central tendency that would occur early in the processing sequence.

There is an alternative approach to RFT that deserves consideration as it emphasizes the concept of similarity as a key factor in judgmental contextualization. Instead of giving equal weight to all stimuli, frequency values can be calculated by giving greater weight to stimuli of similar values. The GEMS model (Qian & Brown, 2005) is a generalization of RFT that relies on similarity-driven sampling. This model states that the judgment context for a stimulus consists primarily of similar stimuli, or, equivalently, that similar stimuli are given greater weight in the judgmental process. The specification of the GEMS model for calculating the frequency values is given in Equation 8.

. (8)

Here, F_i is the frequency value of Stimulus x_i, and γ is the similarity sampling parameter. When $\gamma < 1$, the model gives greater weight to contextual stimuli close to the stimulus being judged. Accounts of the GEMS model with the least squares method ($\gamma = 0.50$), as shown in

Figure 4, are in line with the data observed for the mid-range stimuli but deviate unacceptably from the data observed in the lower and upper regions of the scale. The poor performance of the GEMS model invalidates the assumption that the contextualization of judgments is based on similarity-driven sampling.

There is an opposite approach to rank-based contextualization that relies on the notion of discriminability. The more a stimulus is discriminable from other stimuli, the more likely it is to influence the judgmental context. The combined SIMPLE + DbS model, developed by Brown and Matthews (2011), is a particular extension of the Decision by Sampling framework (Stewart, Chater, & Brown, 2006). The Decision by Sampling framework states that the judgment of a stimulus is constructed from binary, ordinal comparisons to a sample of retrieved or experienced contextual representations (i.e., that the judgment depends on the relative ranked position of that stimulus within this particular sample). The SIMPLE + DbS model assumes that the probability of a stimulus being included in a sample for judgment depends on its discriminability. The confusability of any two stimuli in memory is a decaying exponential function of the distance between them in psychological space:

, (9)

where $\eta_{i,j}$ is the similarity between Stimulus x_i and Stimulus xj, $d_{i,j}$ the distance between them, and c a parameter acting as what we will designate here as the discriminability sampling parameter.

The retrievability of a stimulus depends on its discriminability. The discriminability of Stimulus x_i is inversely proportional to its summed similarity to every other stimulus. Specifically, the discriminability of the trace for Stimulus x_i, D_i, is given by Equation 10.

. (10)

Here, n is the number of comparison stimuli.

Discriminability is converted into predicted recall probability by taking into account the possibility of omissions. The recall probability P_i is given by Equation 11.

. (11)

Here, t is the threshold (such that if discriminability is below a threshold, a stimulus cannot be retrieved) and s determines the slope of the transforming function (effectively, how noisy the omission threshold is).

Figure 4 shows the fits of the SIMPLE + DbS model with the least squares method (c = 1.50, s = 4.0, and t = 0.50) when calculating the frequency values from the recall probabilities instead of the regular percentile ranks. These fits are in good agreement with the results, which suggests that the model captures some key phenomena in relation to the judgmental contextualization. We believe that the notion of discriminability is conceptually close to the combination of the two intertwined principles presented in this article: the frequency-driven random sampling principle and the distance threshold-based consistency principle.

According to k-means rules, within-cluster variance is minimized conditionally to the centroids initialization. As shown in Figure 4, the predictions of the k-means framework, calculated with 40 repetitions of the vanilla k-means algorithm per condition, are in strong

agreement with the results. The best fits were obtained, using the least squares method in all conditions, with $\delta = 5$ and $p = .50$, and with one iteration per repetition. The absence of subsequent assignment-update iterations in those simulations emphasizes the role of the initial centroids selection in the emergence of contrast effects. Contrast effects occur in the region of the range from which multiple centroids are drawn, leading to stimuli of this region being assigned different response categories.

While the results obtained in this experiment supply clear evidence for clustering effects in judgmental relativity, they also raise the interesting question as to why such effects have been overlooked during the past 30 years. The answer is of an epistemological nature. Their detection could have been achieved only by conducting experiments based on appropriate combinations of distribution shapes and of category scales, which has never been attempted. While Parducci was aware of the existence of the k-means algorithm as a method for automated classification (Parducci, 1983), Occam's razor led to favor RFT, because it provided a conceptual framework that was not only simpler, but also more adapted, from a psychophysical standpoint, to the measurement of persistent representations under the form of range values. In that sense, the postulate of the invariance of the range values across different stimulus distributions certainly played an important role in the success of RFT, as it proved very useful to infer scale values from category ratings for various types of stimuli, even in the absence of simple physical measures, as in the case of odor and flavor attributes (Parducci, 1982).

Experiment 2: Judgmental Contrast Across Repeated Blocks of Trials

The objective of this experiment was to demonstrate the dynamical nature of scale adjustment across repeated blocks of trials using the successive presentation method, and to interpret the obtained results in reference to the concept of assignment-update iterations. In order

to characterize the within-category variance minimization conditionally to the order of appearance of the stimuli, two sequences of presentation were used.

Method

Participants. All participants were students enrolled in introductory psychology at the University of Aix-Marseille, Aix-en-Provence. Ten participants served in each condition.

Design. There were three between-subjects factors: (a) Skewing of Distribution (positive skewing versus negative skewing), (b) Method of Skewing (equal length versus unequal length), and (c) Sequence of Presentation (Sequence 1 versus Sequence 2). Four blocks of trials were presented to each participant.

Stimuli and apparatus. Stimuli and order of appearance in the first block of trials (Trials 1-13) are shown in Table 3. The two sequences of presentation used in the first block of trials differed in the degree of correlation between the order of appearance and the rank order among the stimulus values. In Sequence 1, the Spearman's rank correlation coefficient was .47 for the positively skewed distribution and -0.47 for the negatively skewed distribution (i.e., the order of appearance of the stimuli was loosely correlated with their density within the range). In Sequence 2, the Spearman’s rank correlation was .00 for both distributions (i.e., the order of appearance of the stimuli was independent on their density within the range). In all conditions, the stimulus set consisted of 13 black lines whose length varied from 0.50 to 11.50 cm, with a width of 1 mm. The lines appeared one at a time against a white background at the center of a 23-cm monitor. In the equal-length conditions, skewing was achieved by varying the frequencies of five lines, 0.50, 2.50, 5.00, 8.00, and 11.50 cm. In the unequal-length conditions, no line was repeated exactly; instead, repetitions consisted of adding lines of very similar values (every line in a group was within 0.05 cm of another line in the same group).

Procedure. The participants were seated in front of the monitor at a distance of about 60 cm. They were asked to judge lines according to their length. The two extreme lines were shown to the participants before the start of the judgment task. A five-response scale was used, ranging from 1 (*very small*) to 5 (*very large*). Four blocks of 13 stimuli were presented for a total of 52 trials. Each line remained visible until an oral response was provided by the participant and entered by the experimenter. The next line appeared immediately after the response to the preceding one. The room was dimly lit.

Results and Discussion

Figure 5 shows results for the three common lines closest to the midrange of the stimulus range (2.50, 5.00, and 8.00 cm). These results indicate that judgmental contrast decreased for all conditions from Block 1 to Block 4. A four-way 2 (Skewing of Distribution) x 2 (Method of Skewing) x 2 (Sequence of Presentation) x 2 (Block of Trials) mixed-design ANOVA revealed that the Skewing of Distribution x Block of Trials interaction was significant, $F(1, 72) = 29.11$, $p < .0001$. The following effects were also significant: Skewing of Distribution, $F(1, 72) = 56.23$, $p < .0001$; Skewing of Distribution x Sequence of Presentation interaction, $F(1, 72) = 6.25$, $p < .05$; and Sequence of Presentation x Block of Trials interaction, $F(1, 72) = 7.79$, $p < .01$. None of the other effects was significant. If, according to RFT (Parducci, 1983), contrast is due to the tendency to use the different response categories with equal frequency, the decrease of contrast observed from Block 1 to Block 4 should occur concomitantly with a decrease of equalization. Hence, response entropy (calculated as in Equation 12) should be lower in Block 4 than in Block 1.

. (12)

Here, $H(C)$ is the response entropy, and $f(c_j)$ the frequency of category c_j, $H(c_j)$ being equal to 0 if $f(c_j) = 0$.

Response entropy remained stable over the course of the experiment: 1.81 bits in Block 1 and 1.74 bits in Block 4. As the maximum response entropy for five response categories is 2.32 bits, this represents a slight decrease of 3.20% in terms of frequency equalization, which can hardly account for the dramatic shift of the response scale. On the other hand, our results showed that 35.42% of the stimuli received different responses in Block 1 and in Block 4. These results demonstrate that contrast magnitude is not related to the degree of frequencies equalization, and that contrast decrease across repeated blocks of trials is merely caused by scale adjustment.

According to the consistency model (Haubensak, 1992a), scale adjustment is related to the forgetting parameter, f, and to how it affects the probability of applying Assumption 4 of the model, stating that if a new stimulus matches or even exceeds the highest or the lowest of the current standards in memory, participants tend to switch to a higher or lower category. This assumption ensures that the response scale spreads along the range of stimuli as the presentations proceed. If this statement is true, the response range should stretch up toward the extrema from Block 1 to Block 4. To remove the asymmetry between the two skewing conditions, we collapsed the data across the common stimuli. Our results showed that the mean response range assigned to the lines ranging from 2.50 to 8.00 cm remained unchanged, from 2.24 to 4.16 in Block 1, and from 2.18 to 4.05 in Block 4. These results are not consistent with interpreting ratings adjustment as an effect of scale expansion.

Overall, the empirical data provide evidence for a strong influence of centroids initialization in Block 1, and for significant within-category variance reduction in Block 4. As blocks of trials proceed, more complete information about the skewing of stimulus densities is

available, which enables the participants to establish the scale while reducing the within-category variance. From a k-means perspective, variance minimization is achieved, for any non-optimal set of initial centroids, by means of assignment-update iterations. The larger the number of assignment-update iterations, the lower the within-cluster sum of squares *J*. To obtain better clusters, category boundaries are progressively relocated so as to make a clearer distinction between different groups of stimuli.

Our theoretical objective is to account for the data obtained with both the successive presentation method and the simultaneous presentation method by using a single unified framework. Regarding the simultaneous presentation and the data obtained in Experiment 1, we assumed that the assignment-update iterations that are inherent to the vanilla k-means algorithm could be implemented abstractly at the series level (i.e., for all the stimuli at once). Regarding the successive presentation and the data obtained in Experiment 2, there are two modelling options capable of describing the dynamical development of the clusters: trial-by-trial centroids updating and block-level centroids updating. Both options support the concept of variance minimization over time, though at a different level of time granularity. Whilst the former is more compatible with the intuitive assumption that the participants might adjust the judgment scale after each presentation, its implementation in the framework would require adding specific trial-level calculation steps such as those that are developed in more sophisticated versions of the k-means algorithm. In some of those versions, only a limited subset of data points is processed when calculating the Euclidean distances to the centroids (e.g., Ackermann et al., 2012). In other versions, a data point is reassigned to a different cluster only if that reassignment decreases the within-cluster sum of squares *J* (e.g., Leiva & Vidal, 2011). In particular variants of the k-means algorithm devoted to the classification of moving data points, data points are updated only when

the difference between their previous value and their new value exceeds a defined threshold (Zhang, Yang, Tung, & Papadias, 2008). Those specific trial-level calculation steps would be relevant for the successive presentation but not for the simultaneous presentation, which would amount to presenting two different frameworks. On the other hand, because any given block of trials in Experiment 2 consists of the entire series of stimuli, the block-level centroids updating can be regarded as isomorphic to the series-level centroids updating used for the simultaneous presentation method. This is why the block-level modelling was preferred over the trial-by-trial modelling. Therefore, we ran one assignment-update iteration per block of presentations rather than one assignment-update iteration after each new presentation. Simulations were run with 40 repetitions in each condition.

As shown in Figure 5, the predictions of the k-means framework are consistent with the empirical data. The best fits were obtained, using the least squares method, with a psychophysical transformation of the stimulus values based on a power function of exponent 0.72, and with $\delta = 0.08$ and $p = .80$ in all conditions. Instead of randomizing the order of stimuli as stated in Step 3, centroids selection rules were run on stimuli ordered according to their respective sequence of presentation in Block 1. Our assumption was that, if a psychological principle determines how starting centroids are chosen in successive presentation, it must be dependent on the order of appearance. Both this particular assumption and the consistency model (Haubensak, 1992b) rely on the influence of the very first stimuli. While the latter highlights the role of central tendency mechanisms in judgment, the former suggests that contrast stems specifically from over-differentiating a few consecutive stimuli during the very first trials. By assuming that the initial set of centroids is influenced by the order of appearance, we emphasize

the tendency to discriminate between slightly different stimuli during the early stage of scale development.

General Discussion

In this article, our core claim is that judgment is a clustering process, and that a model incorporating within-cluster variance minimization accounts for many phenomena in the rating scale literature, for new contrast effects observed with specific stimulus distributions (Experiment 1), and for dynamical contrast reduction across repeated blocks of trials (Experiment 2). While Parducci's analysis framework and Haubensak's process model underlie two different computationalist interpretations of context effects, the present model is connectionist in essence. The concept of variance minimization has been used in multiple approaches to data clustering and unsupervised learning (e.g., He, Ji, Zhang, & Bao, 2011), and has proven to be critically important to the successful implementation of dynamical systems in the field of artificial intelligence.

Defining judgmental relativity as a product of variance minimization seems challenging in terms of psychological plausibility, as it is improbable that any conscious operations could realistically support such a sophisticated optimization, at least under the form of explicit computational processes. From a connectionist standpoint, this issue is of minor importance, as judgmental relativity can be regarded as resulting from emergent properties. From a computationalist standpoint, it is interesting to note that the concept of variance developed in this article and the concept of information transmission, which has been widely accepted since the 1950s as a powerful framework for describing various cognitive processes (Fabre, 1993), relate equally to the notion of statistical dispersion. This epistemological equivalence between the two concepts leaves little argument that the basic principle of variance minimization should be

considered as a plausible cognitive driver, even though it involves specific operations that are unknown at present.

One question remains. Do the variations of the model's parameters, δ and p, as fitted across different experimental conditions, tell something about how the consistency principle conflicts with the frequency principle? The consistency principle reflects an intuitive reluctance to apply more than a single category to similar stimuli. Table 4 shows the values of p', defined as the product of p and the probability of the Euclidean distance between any two stimuli being lower than δ. The higher p', the higher the chance of rejecting, for consistency purposes, a candidate centroid too close to a similar centroid. We found that p' is higher when the materials reflect a greater uncertainty in establishing whether certain stimuli are repeated exactly (p' varies from .13 to .28 with lines and dots), and lower when the stimuli make any switching of categories more obvious (p' varies from .07 to .11 with numbers). These variations suggest that perceptual uncertainty would tend to increase the weight of the consistency principle, and reciprocally, to reduce the weight of the frequency principle. When perceptual uncertainty causes difficulty in determining if two stimuli are identical or different, the consistency principle would help in regarding them as indistinguishable.

Starting centroids can be seen as the internal representation of the context at the outset of the task. One possible extension to this work would include systematizing backward analysis for inferring the values of the starting centroids from experimental data. In the same way as the range values can be inferred from the observed mean responses based on RFT, the values of the starting centroids could be inferred from the observed clusters based on a backward k-means calculation method. In practice, range values are not inferred at the individual participant level, but from aggregated group data for two stimulus distributions at once (most often, positively and

negatively skewed sets with antagonistic *Fi* values). Alternatively, the values of the starting centroids could be inferred directly from participant-level data, that is, from the entire set of responses assigned to a series of stimuli by any given participant. Such an approach would eliminate the need for assessing particular initialization rules when fitting the data. However, the k-means algorithm is not reversible. Multiple sets of starting centroids can lead to the same final clusters, which makes it impossible to identify the exact set of starting centroids that a given participant had in mind at the outset of the task. Because there is no known formula to determine the mean starting centroid values from a given response set, the only practical approach to backward calculation relies on brute-force equal-weight combination. The total number of initial centroids sets, S, for n stimulus and k categories, is given by equation 13.

. (13)

In the case of the example given in Table 1, where $n = 18$ and $k = 9$, which is typical of the orders of magnitude used in research on judgment, $S = 48620$. If the final clusters showed in Table 1 were explicit responses experimentally obtained from a participant, backward calculation would enable the retrieval of the complete subsets of candidate starting centroids. Brute-force execution of the k-means algorithm shows that 100 different initial centroids sets would generate the same final clusters (the initial centroids set described in Table 1 being one of those particular sets). Detailed values per quantity of iterations are shown in Table 5. Figure 6 shows the mean starting centroids for each of the nine response categories, with the assumption that each of the 100 starting centroids sets is equally probable, regardless of the number of iterations required for reaching final convergence. It is interesting to note that the standard deviation of the mean starting centroid values is reasonably low compared to the stimulus scale: 0.00 for Categories A to F, 40.38 for Category G, 83.11 for Category H, and 118.50 for Category I. The equal-weight

averaging assumption tends to smooth the resulting curve, which would contribute to capturing the main patterns in the participant-level data.

There would be several methodological advantages in systematizing backward analysis of starting centroids. First, starting centroids could be calculated for any given set of participant-level data, which would make inferential testing possible at the starting centroids level. Second, regardless of the response patterns, no psychophysical assumption would be required to transform stimulus values into perceived values, since inferred centroids could be plotted directly on the original stimulus scale. Other models, such as the SIMPLE + DbS model (Brown & Matthews, 2011), successively demonstrated that purely rank-based approaches can account not only for frequency effects, but also for apparent range effects when memory retrieval and distinctiveness are taken into account. Third, due to the equal-weight averaging assumption, between-subjects variance obtained at the starting centroids level would be lower than in the source data, which would help test small-amplitude effects. Fourth, for falsification purposes, starting centroids could be calculated per block of presentations (provided that responses are recorded for all presentations in each block). Repeated measure designs would enable per-category comparisons between starting centroids calculated in Block $t + 1$ and stimulus-response assignments observed in Block t. Fifth, for investigation purposes, starting centroids could be calculated as a function of the quantity of iterations required to reach final convergence. This would help distinguish between effects specifically occurring at the centroids initialization stage and effects occurring concomitantly with subsequent variance minimization efforts. Experimental procedures relying on attentional instructions, dual-task interference, and time pressure could help investigate the demarcation between the two types of effects.

In conclusion, we see the k-means interpretation of context effects and the backward calculation method proposed here as an important step towards unifying models that emphasize judgmental relativity, such as RFT, and models that focus on dynamical properties of judgment, such as the consistency model. We have offered results suggesting that the judgmental context can be represented under the form of centroids at the outset of the task, and presented an approach to data fitting that would help plot its evolution in time.

References

Ackermann, M. R., Martens, M., Raupach, C., Swierkot, K., Lammersen, C., & Sohler, C. (2012). StreamKM++: A clustering algorithm for data streams. *Journal of Experimental Algorithmics, 17*, 173-187. doi: 10.1145/2133803.2184450

Atkinson, R. C., & Crothers, E. J. (1963). A comparison of paired-associate learning models having different acquisition axioms. *Journal of Mathematical Psychology*, *1*, 285-315. doi: 10.1016/0022-2496(64)90005-7

Boillaud, E. (1997). *Modélisation de l'effet du contexte dans la perception* [Modelling of perceptual context effects] (Unpublished doctoral dissertation). University of Provence, Aix-en-Provence.

Brown, G. D. A., & Matthews, W. J. (2011). Decision by sampling and memory distinctiveness: Range effects from rank-based models of judgment and choice. *Frontiers in Psychology*, *2*, 299. doi: 10.3389/fpsyg.2011.00299

Chen, Z., & Shixiong, X. (2009). K-means clustering algorithm with improved initial center. *Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining*, 790-792. doi: 10.1109/WKDD.2009.210

Cogan, E., Parker, S., & Zellner, D. A. (2013). Beauty beyond compare: Effects of context extremity and categorization on hedonic contrast. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 16-22. doi: 10.1037/a0031020

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied*, *12*(3), 166–178. doi: 10.1037/1076-898X.12.3.166

El Agha, M. (2012). Efficient and fast initialization algorithm for k-means clustering. *Intelligent Systems and Applications*, *2012*, *1*, 21-31. doi: 10.5815/ijisa.2012.01.03

Fabre, J.-M. (1993). *Contexte et jugement* [Context and judgment]. Lille: Presses Universitaires de Lille.

Fasold, F., Memmert, D., & Unkelbach, C. (2013). Calibration processes in a serial talent test. *Psychology of Sport and Exercise*, *14*(4), 488–492. doi: 10.1016/j.psychsport.2013.02.001

Haubensak, G. (1992a). The consistency model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 303-309. doi: 10.1037/0096-1523.18.1.303

Haubensak, G. (1992b). The consistency model: A reply to Parducci. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 314-315. doi: 10.1037/0096-1523.18.1.314

Haubensak, G., & Petzold, P. (2003). The influence of instructions on the adjustment of scales. *Perception & Psychophysics*, *65*(2), 329-337. doi:10.3758/BF03194804

He, X., Ji, M., Zhang, C., & Bao, H. (2011). A variance minimization criterion to feature selection using Laplacian regularization. *Patterns Analysis and Machine Intelligence, 33*(10), 2013-2025. doi: 10.1109/TPAMI.2011.44

Helson, H. (1948). Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological Review*, *55*(6), 297-313. doi: 10.1037/h0056721

Kanungo, T., Mount, D. M, Netanyahu, N. S, Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *Transactions on*

Pattern Analysis and Machine Intelligence, *24*, 881-892. doi: 10.1109/TPAMI.2002.1017616

Leiva, L. A., & Vidal, E. (2011). Warped k-means: An algorithm to cluster sequentially-distributed data. *Information Sciences, 237*, 196-210. doi: 10.1016/j.ins.2013.02.042

MacQueen, J. (1966). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam, & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-297. Berkeley, University of California Press. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512992

Matthews, W. J. (2013). Relatively random: Context effects on perceived randomness and predicted outcomes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1642-1648. doi: 10.1037/a0031081

Matthews, W. J., & Stewart, N. (2009). Psychophysics and the judgment of price: Judging complex objects on a non-physical dimension elicits sequential effects like those in perceptual tasks. *Judgment and Decision Making*, *4*(1), 64-81. Retrieved from http://journal.sjdm.org/81104/jdm81104.html

Matthews, W. J., Stewart, N., & Wearden, J. H. (2011). Stimulus intensity and the perception of duration. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 303-313. doi: 10.1037/a0019961

Milhabet, I., Le Barbenchon, E., Molina, G., Cambon, L. & Steiner, D. D. (2012). Comparative optimism, so useful. *International Review of Social Psychology*, *25*(2), 5-40. doi: 10670/1.3csgwf

Molina, G., & Fabre, J.-M. (1999). Influences d'un contexte informationnel, d'une préférence et d'une norme conventionnelle sur l'évaluation du confort thermique [The influence of

context, preference and conventional norm on the estimation of comfortable temperatures]. *Revue Internationale de Psychologie sociale/International Review of Social Psychology*, *12*(1), 37-52.

Molina, G., & Fabre, J.-M. (2000). Norme et contexte : influence d'une dichotomisation du matériel et de l'évaluation sur la contextualisation de jugements catégoriels [Norm and context: Dichotomization of stimuli and of response scales influences context effects in absolute judgments]. *L'Année Psychologique*, *2000*, 37-69. doi:10.3406/psy.2000.28626

Molina, G., & Fabre, J.-M. (2001). Conflict between two dichotomies: Dichotomization of stimuli and judgments. *Current Psychology Letters: Behaviour, Brain & Cognition*, *5*, 91-103.

Parducci, A. (1956). Direction of shift in the judgment of single stimuli. *Journal of Experimental Psychology*, *51*(3), 169-178. doi: 10.1037/h0041609

Parducci, A. (1965). Category judgments: A range-frequency model. *Psychological Review*, *72*(6), 407-418. doi: 10.1037/h0022602

Parducci, A. (1982). Scale values and phenomenal experience: There is no psychological law! In H.-G Geissler, P. Petzold, H. F. Buffart, & Y. M. Zabrodin (Eds.), *Psychophysical judgment and the process of perception* (pp. 11-16). Berlin: VEB Deutscher Verlag der Wissenschaften.

Parducci, A. (1983). Category ratings and the relational character of judgment. In H.-G Geissler (Ed.), *Modern issues in perception* (pp. 262-282). Berlin: Deutscher Verlag der Wissenschaften.

Parducci, A., Calfee, R. C., Marshall, L. M., & Davidson, L. P. (1960). Context effects in judgments: Adaptation level as a function of the mean, midpoint, and median of the stimuli. *Journal of Experimental Psychology*, *60*(2), 65-77. doi: 10.1037/h0044449

Parducci, A., & Wedell, D. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(4), 496-516. doi: 10.1037/0096-1523.12.4.496

Parker, S., Moore, J. M., Bahraini, S., Gunthert, K., & Zellner, D. A. (2012). Effects of expectations on loudness and loudness difference. *Attention, Perception, & Psychophysics*, *74*(6), 1334-1342. doi: 10.3758/s13414-012-0326-8

Penney, T. B., Brown, G. D. A., & Wong, J. K. (2013). Stimulus spacing effects in duration perception are larger for auditory stimuli: Data and a model. *Acta Psychologica*, *147*, 97-104. doi: 10.1016/j.actpsy.2013.07.017

Petzold, P., & Haubensak, G. (2001). Higher order sequential effects in psychophysical judgments. *Perception & Psychophysics, 63*(6), 969-978. doi: 10.3758/BF03194516

Qian, J., & Brown, G. D. A. (2005). Similarity-based sampling: Testing a model of price psychophysics. *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*, 1785-1790. Retrieved from http://www.psych.unito.it/csc/cogsci05/frame/poster/2/p591-qian.pdf

Rashotte, M. A., & Wedell, D. H. (2012). Context effects on tempo and pleasantness judgments for Beatles songs. *Attention, Perception, & Psychophysics*, *74*(13), 575-599. doi: 10.3758/s13414-011-0255-y

Shannon, C. E., & Weaver, W. (1963). *The mathematical theory of communication.* University of Illinois Press.

Sharma, N., Singh, A. P., & Gupta, A. K. (2013). An experimental approach of k-means algorithm on the data set. *International Journal of Engineering Sciences & Emerging Technologies, 6*(1), 49-56. Retrieved from http://www.oalib.com/paper/2117562

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*(4), 881-911. doi: 10.1037/0033-295X.112.4.881

Stewart, N., Chater, N, & Brown, G. D. A (2006). Decision by sampling. *Cognitive Psychology, 53*(1), 1-26. doi:10.1016/j.cogpsych.2005.10.003

Tommasi, M. (2001). Frequency effects in distributions of stimuli with or without standards. In E. Sommerfeld, R. Kompass, & T. Lachmann (Eds.), *Proceedings of the Seventeenth Annual Meeting of the International Society for Psychophysics*, 641-646. Leipzig, Germany: The International Society for Psychophysics.

Watkinson, P., Wood, A. M, Lloyd D. M., & Brown G. D. A. (2013). Pain ratings reflect cognitive context: A range frequency model of pain perception. *Journal of the International Association for the Study of Pain*, *154*(5), 743–749. doi: 10.1016/j.pain.2013.01.016

Wedell, D. H., Parducci, A., & Roman, D. (1989). Student perceptions of fair grading: A range-frequency analysis. *American Journal of Psychology*, *102*(2), 233- 248. Retrieved from http://www.jstor.org/stable/1422955

Wedell, D. H., Santoyo, E. M., & Pettibone, J. C. (2005). The thick and the thin of it: Contextual effects in body perception. *Basic and Applied Social Psychology*, *27*(3), 213-227. doi: 10.1207/s15324834basp2703_3

Zellner, D. A., Jones, K., Morino, J., Cogan, E. S., Jennings, E. M., & Parker, S. (2010). Increased hedonic differences despite increases in hedonic range. *Attention, Perception, & Psychophysics*, *72*(5), 1261-1265. doi: 10.3758/APP.72.5.1261

Zellner, D. A, Mattingly, M. C., & Parker, S. (2009). Categorization reduces the effect of context on hedonic preference. *Attention, Perception, & Psychophysics*, *71*(6), 1228-1232. doi: 10.3758/APP.71.6.1228

Zhang, Z., Yang, Y., Tung, A., & Papadias, D. (2008). Continuous k-means monitoring over moving objects. *Knowledge and Data Engineering, 20*(9), 1205-1216. doi: 10.1109/TKDE.2008.54

Table 1

An Illustration of how the K-Means Algorithm Minimizes the Within-Cluster Variance

	Stimuli																	
	108	160	212	264	316	368	420	472	524	576	628	680	732	784	836	888	940	992
Starting centroids	(A) 108	(B) 160	(C) 212	(D) 264	(E) 316	(F) 368	(G) 420	(H) 472	(I) 524									
Iteration 1																		
Assignment	A	B	C	D	E	F	G	H	I	I	I	I	I	I	I	I	I	I
Update	(A) 108	(B) 160	(C) 212	(D) 264	(E) 316	(F) 368	(G) 420	(H) 472					(I) 758					
Iteration 2																		
Assignment	A	B	C	D	E	F	G	H	H	H	I	I	I	I	I	I	I	I
Update	(A) 108	(B) 160	(C) 212	(D) 264	(E) 316	(F) 368	(G) 420		(H) 524						(I) 810			
Iteration 3																		
Assignment	A	B	C	D	E	F	G	G	H	H	H	I	I	I	I	I	I	I
Update	(A) 108	(B) 160	(C) 212	(D) 264	(E) 316	(F) 368	(G) 446			(H) 576					(I) 836			
Iteration 4																		
Assignment	A	B	C	D	E	F	G	G	H	H	H	H	I	I	I	I	I	I
Update	(A) 108	(B) 160	(C) 212	(D) 264	(E) 316	(F) 368	(G) 446			(H) 602					(I) 862			
…																		
Iteration 8																		
Assignment	A	B	C	D	E	F	F	G	G	G	H	H	H	H	I	I	I	I
Update	(A) 108	(B) 160	(C) 212	(D) 264	(E) 316	(F) 394			(G) 524			(H) 706				(I) 914		
Final clusters	A	B	C	D	E	F	F	G	G	G	H	H	H	H	I	I	I	I

Notes. Starting and updated centroids are represented by letters in brackets. Ties are broken in favor of lower centroids.

Table 2

Stimulus Distributions Used in Experiment 1

	Stimuli																								
Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Pos. stretched	108	145	182	218	255	292	329	366	403	440	476	513	550	551	552	553	554	555	771	987	988	989	990	991	992
Neg. stretched	108	109	110	111	112	113	329	545	546	547	548	549	550	587	624	660	697	734	771	808	845	882	918	955	992

Notes. Pos. stretched = positively stretched. Neg. stretched = negatively stretched.

Table 3

Stimulus Distributions and Orders of Appearance Used in Experiment 2

	Stimuli per order of appearance in the first block of presentation (Trials 1-13)												
Set	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency													
Sequence 1													
Pos.	0.50	0.50	2.50	2.50	2.50	0.50	5.00	0.50	2.50	5.00	8.00	0.50	11.50
Neg.	11.50	11.50	8.00	8.00	8.00	11.50	5.00	11.50	8.00	5.00	2.50	11.50	0.50
Sequence 2													
Pos.	0.50	2.50	5.00	8.00	11.50	0.50	2.50	5.00	0.50	0.50	0.50	2.50	2.50
Neg.	11.50	8.00	5.00	2.50	0.50	11.50	8.00	5.00	11.50	11.50	11.50	8.00	8.00
Spacing													
Sequence 1													
Pos.	0.50	0.70	2.40	2.70	2.60	0.65	5.00	0.55	2.50	4.80	8.00	0.60	11.50
Neg.	11.50	10.90	8.15	7.70	7.85	11.00	5.00	11.30	8.00	5.20	2.50	11.10	0.50
Sequence 2													
Pos.	0.50	2.40	5.00	8.00	11.50	0.55	2.50	4.80	0.60	0.65	0.70	2.70	2.60
Neg.	11.50	8.15	5.00	2.50	0.50	11.30	8.00	5.20	11.10	11.00	10.90	7.70	7.85

Note. Pos. = positive skewing. Neg. = negative skewing. Values are in centimeters.

Table 4

Overview of the Consistency Principle in all Experimental Conditions

Experiment	Stimuli	Distribution	Model parameters		
			δ	p	p'
Boillaud's (1997) Experiment 1	Lines	Pos.	1.25	.95	.28
		Neg.	1.25	.95	.28
		Unimodal	1.25	.95	.28
		Bimodal	1.25	.95	.28
Parducci and Wedell's (1986) Experiment 4B	Numbers	Pos., 5 stimuli	5.00	.50	.11
		Neg., 5 stimuli	5.00	.50	.11
		Pos., 9 stimuli	5.00	.50	.07
		Neg., 9 stimuli	5.00	.50	.07
Parducci and Wedell's (1986) Experiment 4C	Dots	Pos.	0.15	.80	.28
		Neg.	0.15	.80	.28
Experiment 1	Numbers	Positively stretched	5.00	.50	.07
		Negatively stretched	5.00	.50	.07
Experiment 2	Lines	Pos., frequency	0.08	.80	.22
		Neg., frequency	0.08	.80	.22
		Pos., spacing	0.08	.80	.13
		Neg., spacing	0.08	.80	.13

Note. Pos. = positive skewing. Neg. = negative skewing. *p'* is defined as the product of *p* and the probability of the Euclidean distance between any two stimuli being lower than δ.

Table 5

Backward Calculation of Starting Centroids per Quantity of Iterations

		Categories								
Iterations	Occurrences	A	B	C	D	E	F	G	H	I
1	8	108.00	160.00	212.00	264.00	316.00	368.00	498.00	680.00	914.00
2	20	108.00	160.00	212.00	264.00	316.00	368.00	461.60	669.60	888.00
3	13	108.00	160.00	212.00	264.00	316.00	368.00	492.00	632.00	840.00
4	18	108.00	160.00	212.00	264.00	316.00	368.00	454.67	613.56	833.11
5	22	108.00	160.00	212.00	264.00	316.00	368.00	453.09	576.00	786.36
6	12	108.00	160.00	212.00	264.00	316.00	368.00	424.33	519.67	758.00
7	6	108.00	160.00	212.00	264.00	316.00	368.00	420.00	480.67	662.67
8	1	108.00	160.00	212.00	264.00	316.00	368.00	420.00	472.00	524.00

Notes. Ties are broken in favor of lower centroids.

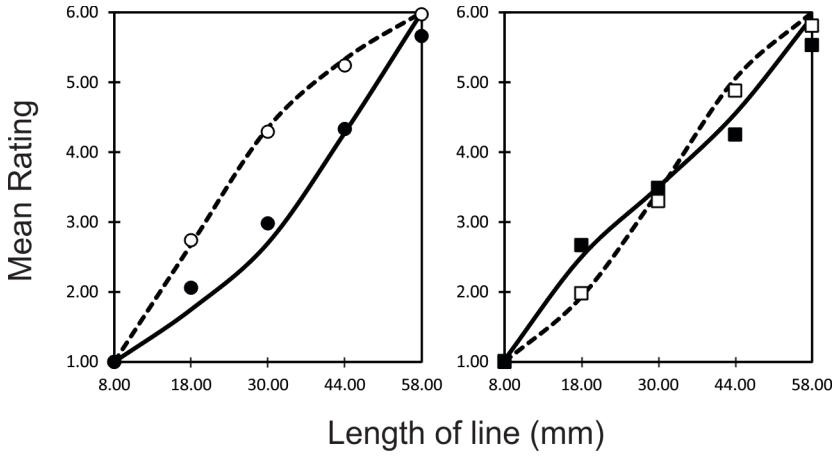

Figure 1. Effect of different distribution shapes on perceptual judgments of lengths of line. Empirical ratings are from Boillaud's (1997) Experiment 1: data for positive set shown by open points, negative set by solid points, unimodal set by open squares, bimodal set by solid squares. Theoretical fits obtained from k-means simulations are represented by lines.

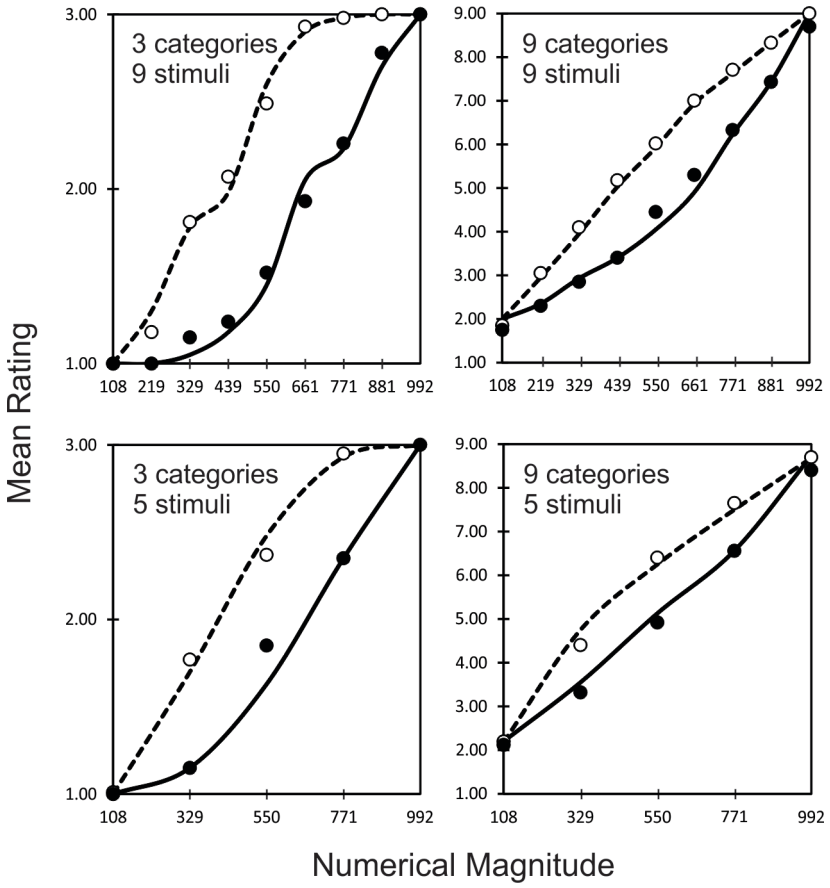

Figure 2. Stimulus and category effects with frequency-skewed distributions. Empirical ratings are from Parducci and Wedell's (1986) Experiment 4B: data for positive set shown by open points, negative set by solid points. Theoretical fits obtained from k-means simulations are represented by lines.

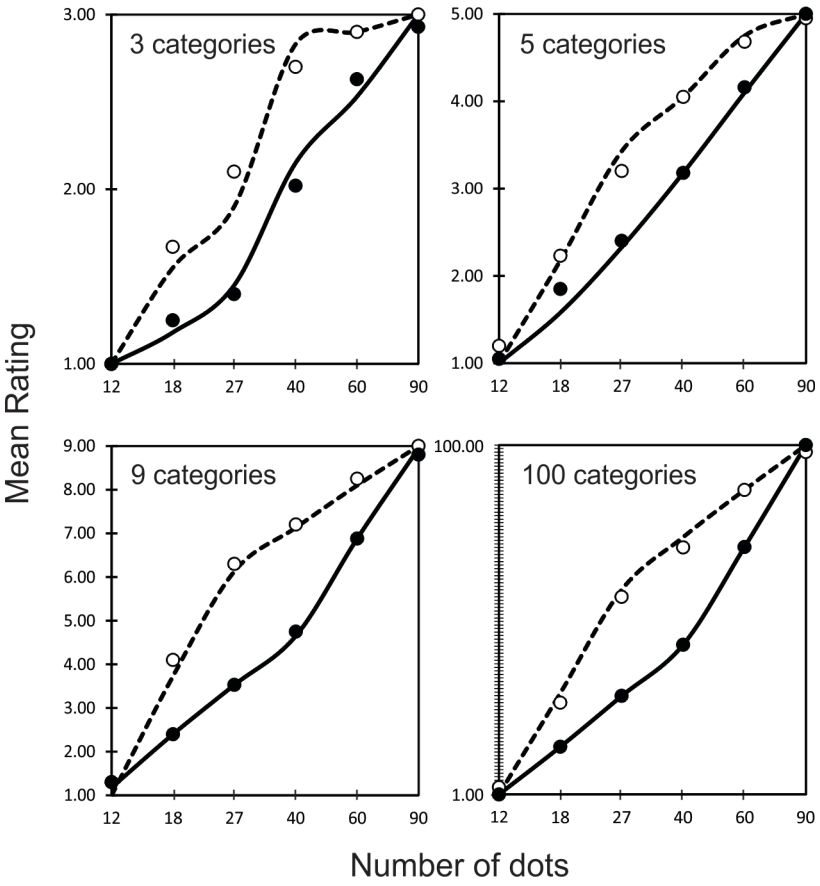

Figure 3. Absence of category effect with spacing-skewed distributions. Empirical ratings are from Parducci and Wedell's (1986) Experiment 4C: data for positive set shown by open points, negative set by solid points. Theoretical fits obtained from k-means simulations are represented by lines.

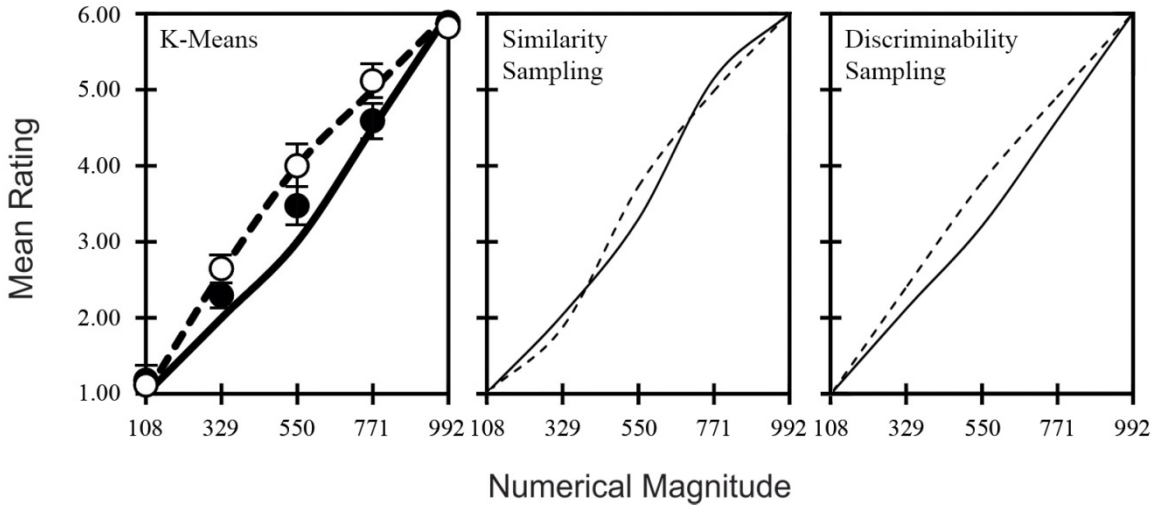

Figure 4. Judgmental contrast on two sets of numbers with identical mean, midrange, median, range, and percentile rank, and randomization of the presentation order on the page (left panel). Empirical ratings for positively stretched set are represented by open points, for negatively stretched set by solid points. Standard deviations are represented by error bars. Theoretical fits obtained from k-means simulations are represented by lines. Best fits obtained with the GEMS model are shown in the middle panel, best fits obtained with the SIMPLE + DbS model are shown in the right panel: Fits for positively stretched set are represented by dashed line, for negatively stretched set by solid line.

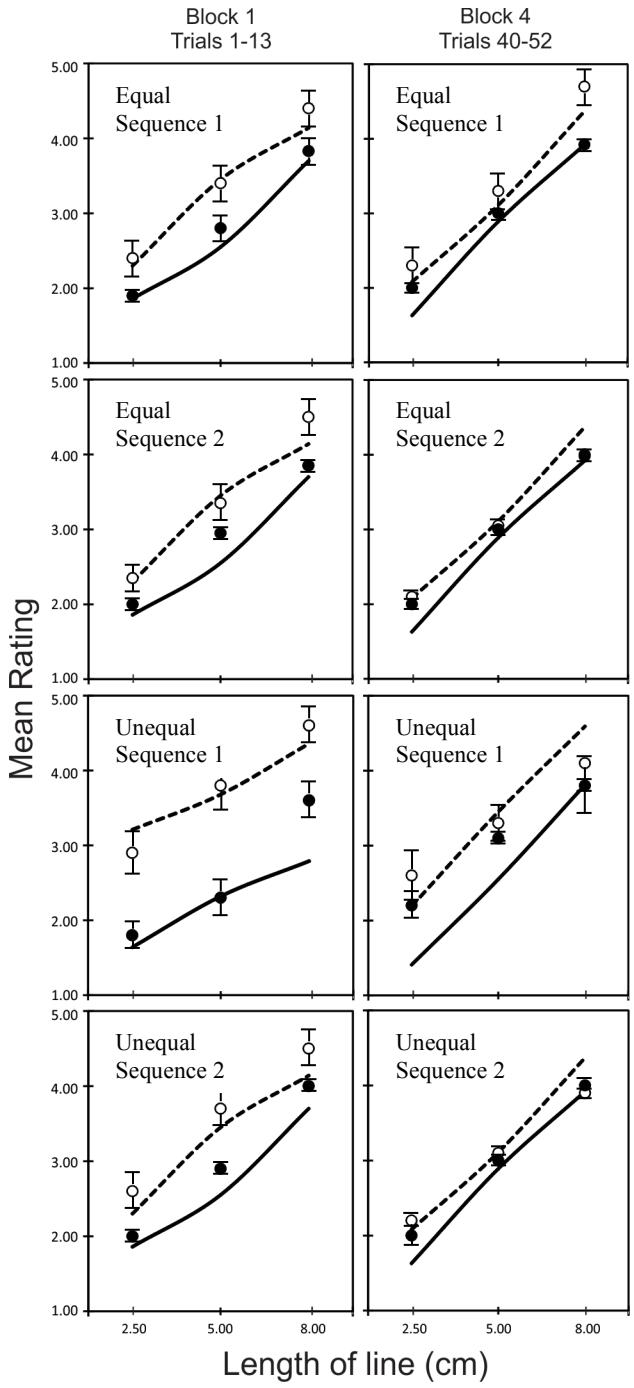

Figure 5. Dynamical contrast reduction across repeated blocks of trials, with two methods of skewing and two sequences of presentation. Empirical ratings for positive set are represented by open points, for negative set by solid points. Standard deviations are represented by error bars. Theoretical fits obtained from k-means simulations are represented by lines.

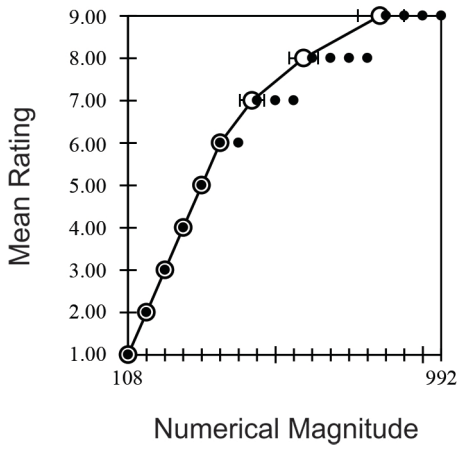

Figure 6. Backward calculation of starting centroids (equal-weight averaging assumption). Starting centroids are represented by open points, clustered data by solid points. Standard deviations are represented by horizontal error bars.