Atten Percept Psychophys (2014) 76:1847–1859
DOI 10.3758/s13414-014-0678-3

How do people order stimuli?

Simon Kemp · Randolph C. Grace

Published online: 8 May 2014

Abstract People may find it easier to construct an order after first representing stimuli on a scale or categorizing them, particularly when the number of stimuli to be ordered is large or when some of them must be remembered. Five experiments tested this hypothesis. In two of these experiments (1 and 3), we asked participants to rank line lengths or to rank photographs by artistic value. The participants provided evidence of how they performed these tasks, and this evidence indicated that they often made use of some preliminary representation—either a metric or a categorization. Two further experiments (2 and 4) indicated that people rarely produced rankings when given a choice of assessment measures for either the length of lines or the artistic value of photographs. In Experiment 5, when the number of lines was larger or lines were only visible one at a time, participants were faster at estimating line lengths as a percentage of the card covered than at rank ordering the lengths. Overall, the results indicate that ordering stimuli is not an easy or natural process when the number of stimuli is large or when the stimuli are not all perceptible at once. An implication is that the psychological measures available to individuals are not likely to be purely ordinal when many of the elements being measured must be recalled.

Keywords Measurement theory · Memory · Ordering · Ranking · Category rating

Ordering is a common human activity. Examples abound of "Top 10" lists of the best books or movies of the year; cities are ranked annually according to quality of life for their residents; and popular websites feature regular "power rankings" of the best professional clubs in a variety of sports. Common to such lists is the rank ordering of items in terms of their value on a particular dimension. Here, we focus on how such orders may be obtained, particularly in situations in which the number of items to be ordered is relatively large or the items are not simultaneously present or visible. We argue that ranking is often a difficult activity because ordinal representations are inefficient in such cases, and that people are likely to construct a preliminary metric or categorization when asked to produce a rank order.

Despite the ubiquity of ordering in human culture, relatively little previous research has investigated how people construct orders. In Rokeach (1973) scaling, there has been some question over whether values should be ranked or rated, and Alwin and Krosnick (1985) found that value ranking was more difficult for participants to do and took longer to accomplish. Within the field of computer science, researchers have devised a variety of methods to order stimuli (or in their terminology, sort elements). This work has identified a large number of different sorting algorithms but no general theory of which is most efficient in every kind of circumstance (Knuth, 1998). Chignell and Patty (1987) used computer sorting theory to suggest more efficient ordering methods in psychological research. However, none of these studies have shown that humans are naturally likely to construct orders of stimuli or ordinal scales, or investigated how people actually go about ordering stimuli when allowed to choose their own method. The relative lack of research on ordering is in curious contrast to the large number of studies investigating how people perform magnitude estimation or category scaling (e.g., Ashby, 1992; Bolanowski & Gescheider, 2013; Krueger, 1989; Laming, 1997).

The notion that ranking or ordering stimuli *is* a fundamental and natural process may be implicit in the way that psychological measurement theory has evolved. Stevens's (1946,

S. Kemp (✉) · R. C. Grace (✉)
Psychology Department, University of Canterbury, Private Bag 4800, Christchurch, New Zealand
e-mail: Simon.Kemp@canterbury.ac.nz
e-mail: Randolph.Grace@canterbury.ac.nz

1955) delineation of four scales of measurement—nominal, ordinal, interval, and ratio—is well-known in psychology, and his demonstration that the averaging of data makes little sense for nominal and ordinal scales has had far-reaching, if controversial, consequences for methodology (e.g., Davison & Sharma, 1988; Gardner, 1975; Lord, 1953; Maxwell & Delaney, 1985; Michell, 1999). One result of this debate has been that psychologists often use statistical methods that require only the assumption that the measure can be ordered (e.g., Long, 1997). On a more theoretical level, representational measurement theory frequently takes the existence of a more or less well-ordered attribute or dimension as a starting point for investigating the properties of different types of psychological scales (e.g., Krantz, Luce, Suppes, & Tversky, 1971; Luce, 1996). In measurement theory the scale types differ in terms of the admissible transformations that can be performed on them (e.g., Krantz et al., 1971, pp. 10–11; Roberts, 1979, pp. 64–65). Thus ordinal scale values can be transformed by any monotonically increasing function; interval scale values only by linear functions ($ax + b$); ratio scales only by multiplication by a constant (ax). Another scale type, absolute, cannot be transformed at all. The most common example of an absolute scale is a simple count. An ordering can be obtained from ordinal, interval, ratio, and absolute scales, but the last three also contain other information.

A rank ordering of a group of stimuli is probably the type of ordinal scale that comes most readily to mind (although it is not the only one).[1] But, although ordinality may be taken as a basic property of a mathematical system, this does not necessarily imply that ordering a collection of stimuli is a natural and easy task for people.

When an individual judges and compares stimuli in any situation including that of ordering—consider such varied tasks as estimating the relative length of lines displayed simultaneously, the loudness of sounds presented sequentially, the artistic quality of photographs, the cuteness of cats, or the excellence of student essays—then she or he is likely to be doing this on the basis of some measure or set of comparisons that is available to them. (Different individuals may use greatly different types of measure but that is another question.) We might then ask: What scale type is the measure that they use for the particular task? Interestingly, there does not seem to be a consensus answer to this question. Kemp and Grace (2012) asked practicing researchers to nominate the type of scale that individuals would use to rate the cuteness of cats, to magnitude estimate the loudness of sounds, and to assign percentage marks to a number of essays. For all three tasks, at least 20 % of the researchers (51 % for rating the cuteness of cats) believed the measure to be ordinal, and in no case did a clear consensus emerge. Thus, the issue of what scales individuals actually use seems an important topic for further theoretical and empirical enquiry.

A closer look at measurement theory provides a rationale for why ordinal scales might not be a suitable measure for individuals to use when they make relative judgments and for why ordering might not be the operation of first resort for them. An important property that distinguishes ordinal, interval, and ratio scales is their uniqueness (e.g., Luce, Krantz, Suppes, & Tversky, 1990, chs. 18, 20; Narens, 2002, ch. 5). We consider ratio scales first because these are the simplest case, and use length as an example. Many possible measures can be used for length: meters, feet, inches, cubits, and so forth. There will be relationships (automorphisms) between each pair of these measures. Now if we sample from our collection of measures of length, two of the different measures could be exactly the same. That is, for every object that we measure—for example, the heights of different people—the numbers on the two different measures might be exactly the same. The question is, How many different objects or points (apart from zero) do we need to examine to see whether the two measures are in fact identical? For length, as for any ratio scale, the answer is one. If the two measures agree at just one point, then they are the same. Formally, ratio scales are said to be one-point unique. For an interval scale (e.g., temperature), two measures are exactly the same if they agree at two different points. They are two-point unique. The concept of uniqueness corresponds to the fact that transformations of a ratio scale have one free parameter ($x' = ax$), whereas transformations of an interval scale have two ($x' = ax + b$).

A scale is n-point unique if, once the measures agree at n points, they are identical. If there is no value of n for which this is true, then the scale is said to be *∞-point unique* (e.g., Luce et al., 1990, pp. 115–116). Ordinal scales are ∞-point unique. It is fairly easy to see why this is so. Consider, for example, that we assign each person in a group a number for their height. The shortest gets 1; the next shortest 2; the tenth shortest 10; and so on. Now consider a different measure that is exactly the same, except that the tenth person gets the number 10.5. Both measures are legitimate, since order is preserved, but even though the measures agree at any n points, they are never identical.

So far, we have considered scale types solely as mathematical entities, and this mathematical theory would not necessarily relate to human functioning. However, if, as Stevens and a host of subsequent psychologists have presumed, these mathematical entities are relevant to the ways that individuals judge qualities such as line length, loudness, artistic merit, or

[1] Ordinal scales can be constructed without rank ordering. For example, take the first element encountered and arbitrarily assign it the number 100. If the next element encountered is smaller, assign a smaller arbitrary number, say 33. If the third element is somewhere in between, assign the number 82, and so on. Using this construction, the difference between the elements assigned the numbers 80 and 90 will not in any important sense be equal to the difference between the elements assigned 90 and 100. Any monotonic transformation of a scale constructed in this way would leave the ordering unaffected.

the seriousness of different crimes, then it is reasonable to consider how people might arrive at these judgments. In particular, the concept of uniqueness has implications for how people might measure or evaluate stimuli when some or all of the stimuli need to be remembered.

We can model the ordering process more formally. Let y_1, y_2, . . . , y_m be the memory representations of m exemplars along a particular ordered attribute or quality, such that $y_1 > y_2 > y_3 \ldots y_{m-1} > y_m$. Assume that an individual perceives a series of new exemplars sampled randomly, one at a time without replacement, from a set $x_1, x_2, \ldots, x_n$, such that $x_1 > x_2 > \ldots x_{n-1} > x_n$, and these are to be integrated within the existing memory representation to produce a single ordered representation of $m + n$ elements. We model this integration process, which translates perception into memory, as scale transformation. To integrate the first exemplar x_i, the individual applies a transformation $y_i = f(x_i)$, where y_i is appropriately ordered within a new memory representation of $m + 1$ elements, and this process is repeated for the m elements, producing a set $y_1, \ldots, y_{m+n}$. First, consider that $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ are measured on different ratio scales. In this case, $f(x_i) = ax_i$, so the individual would need only one additional piece of information—the proportionality factor a that equates the two ratio scales. This factor is the same for each new exemplar in the series. If $y_1, \ldots, y_m$ and $x_1, \ldots, x_n$ are assumed to be different interval scales, $f(x_i) = ax_i + b$, and two values are required to translate between the scales.

However, consider if $y_1, \ldots, y_m$ and $x_1, \ldots, x_n$ are different ordinal scales. Because ordinal scales are ∞-point unique, no valid transformation $y_i = f(x_i)$ preserves order. Instead, the transformation from perception (x_i) to memory (y_i) results in a loss of ordering information, and the individual will have to compare the new exemplar y_i with the memorial representation. To integrate y_i and produce an ordered representation with $m + 1$ elements requires on average $m/2$ individual comparisons. Thus, the total number of comparisons required to integrate the n exemplars and produce an ordered list of $m + n$ elements is

$$\frac{m}{2}\sum_{i=1}^{n-1} n = \frac{m(n^2-n)}{4}$$

Thus, if stimuli are represented in memory as ordinal scales, the number of comparisons required to integrate new exemplars with memory increases as a function of the square of stimulus set size (n). By contrast, for ratio and interval scales the integration of new exemplars depends only on the scale transformation function f and not on set size.

This reasoning may seem abstract, but is illustrated by a simple thought experiment. Consider the task of assessing the heights of people. Suppose you view one person at a time, and you remember the heights according to your personal ratio scale. A particular person may in fact be 165 cm high. To relate this measure to your personal measure of height you need only the one parameter that links the cm measure to your personal measure. If, however, you had remembered an ordinal measure of n previous person heights (this may but need not be a rank), you also need to remember n relationships between the cm measure and your ordinal measure.

In the example of ordering heights, much depends on whether all of the people are physically present and how many there are. If the number of people whose heights are to be ordered is small and they are all present, the task is fairly straightforward, especially if you are free to move them around physically. However, consider what happens when some of the people are not physically present. Suppose that yesterday you saw Anne, Brenda, and Carla and you decided that Anne is taller than Brenda and Brenda taller than Clara. Today you see Xiao. How do you insert Xiao into your ordering? Essentially the only way to include Xiao is to reconstruct the heights of Anne, Brenda, and Carla as a metric (or possibly a categorization), and then to compare Xiao's height to these reconstructed values. As was just pointed out, a separate function is needed for each height reconstruction.

It is easy to see that in this case, it would be simpler to remember some kind of information other than the ordering. For example, you might estimate the height of each person in feet and inches or centimeters, and then later sort on the basis of this information rather than use the simple higher or lower comparisons. Another, and somewhat different, alternative would be to categorize the people into groups (e.g., very tall, tall, etc.) and then later to sort within the groups to get a more precise ordering. This presumes that the categories are themselves ordered and perhaps spaced. A detailed consideration of how categories are formed and used goes well beyond the scope of the present article. However, many accounts of categorization assume, first, that the categories can be ordered and, second, that the assignment of elements to them can be described either by assuming the placement of category boundaries on an interval scale (e.g., Parducci, 1965, 1982) or by the similarity of elements to other elements or a prototype already within the category (e.g., Ashby, 1992; Nosofsky & Stanton, 2005; Petrov, 2011). Note here that when information must be remembered, to place stimuli within N categories, one needs to remember either $N - 1$ category boundaries or N

prototypes.[2] (Kemp & Grace, 2012, discussed whether a category rating established on a single dimension can be regarded as an interval-scale measurement.)

The experiments described below all tested the basic hypothesis that, when stimuli become more numerous or are not all simultaneously present, constructing an order becomes difficult, and participants are likely to use a preliminary metric or categorization. Specific hypotheses and some extra considerations behind the different experimental design choices are given in the introductions to some of the experiments below.

In overview, one independent variable in all the experiments, with the exception of Experiment 3, was whether the stimuli were all visible at once (open) or only one stimulus could be viewed at a time (closed). The numbers of stimuli to be assessed were varied in Experiments 1, 2, and 5, which all used lines of different lengths as their stimuli. We had a number of reasons for examining the dimension of line length. First, physical line length is a ratio scale, and good evidence suggests that perceived or remembered line length is a ratio scale as well (e.g., Narens, 1996; Steingrimsson & Luce, 2007). Second, the perception of line length, at least in the experiments outlined below, is very likely to be unidimensional, depending almost exclusively on the physical length of the line. Thus, line length experiments avoid the issue of how orderings in different dimensions might be conjoined (Kemp & Grace, 2010). Third, both perceived and remembered line length generally relate to actual length with a Stevens's law exponent close to 1 (e.g., Kerst & Howard, 1978). Thus, it was natural to use a linear spacing of different line stimuli and to reduce the possibility of contextual effects (e.g., Laming, 1997, ch. 11; Poulton, 1989).

In Experiment 1, participants were asked to construct a rank ordering of the stimuli, and the key dependent variable was the method that they used to do this. We hypothesized that the participants would be more likely to use a metric or categorization as an intermediate measure rather than simply ordering when the number of stimuli to be ordered was large, or in the closed conditions. In Experiment 2, participants chose the final line length measure themselves. We predicted that they would be more likely to report metrics or categories than orders when the number of stimuli was large or in the closed conditions.

In Experiments 3 and 4, the participants assessed the artistic quality of photographs. In Experiment 3 they were required to produce a rank ordering, and in Experiment 4 they were free to choose their own measure. Thus, Experiments 3 and 4 paralleled Experiments 1 and 2 and addressed similar hypotheses, but with a very different quality.

In Experiment 5, the time taken to assess the stimuli was the dependent variable, and we contrasted the reaction times taken to rank order line lengths and those taken to assess the lengths using a percentage (metric) measure under different conditions.

Experiment 1

In both Experiments 1 and 2, the key independent variables were the numbers of lines presented in a set and whether all of the lines in a set were simultaneously visible (the open condition) or were only visible one at a time (the closed condition). In Experiment 1, respondents were required to end up with a rank ordering of the line lengths. We hypothesized that as the task required greater demands on memory—that is, in the closed as opposed to the open conditions of the experiments and for larger (25) rather than smaller (10) sets—participants would tend not to rank directly, but instead establish an order after first estimating a metric or categorization of line length. We hypothesized that direct ranking would be less frequent when the lines were only visible one at a time or when many lines were to be ordered.

Method

A total of 24 paid participants (17 female, seven male), with a median age of 24 years and a range from 18 to 33, were recruited from around the university. They were tested individually. All were asked to rank order the lengths of lines displayed on cards in both an open and a closed condition. Each participant ordered a set of ten lines and a set of 25 lines. Whether the participants ordered a ten-line set in the open condition and a 25-line set in the closed condition was counterbalanced across participants. Similarly counterbalanced were the order of doing the two tasks and the actual sets used.

The stimuli consisted of two sets of 25 cards and two sets of ten cards. Each card had the university logo on the back. On the front was pasted a 100-mm-long label featuring a single solid black line on an otherwise white background. The standard label and card size also gave the participant an idea of the possible stimulus range from the outset. The line was always 6 points in width. The two different sets of 25 line lengths were

[2] To illustrate why ordering may be more efficient after prior categorization, consider sorting a deck of playing cards into a (Bridge) order from the ace of spades down to the two of spades, then from the ace of hearts to the two, and so on to the two of clubs. One way to do this would be to first sort (categorize) all of the cards by suit and then to sort within suits. This procedure entails fewer judgments than the more intuitive selection sort (cf. Knuth, 1998), which requires first finding the ace of spades, then the king of spades, and so on. The selection sort requires on average of 26 judgments to find the ace of spades, 25.5 to find the king, and so forth. This adds up to 689 judgments [$1/2 * n(n + 1)/2$]. The category sort requires 52 judgments to put all of the cards into suits; then, for each suit an average of 6.5 judgments are required to find the ace, 6 to find the king, and so on. This adds up to 234 judgments $\{n + 4(1/2)[(n/4)(n/4 + 1)/2]\}$. Of course, there is a "setup cost" of knowing the suits, and examining each card to see whether it is greater or less than the previous highest card may or may not be easier than placing a card on the spades pile.

chosen to approximate an even distribution between 0.1 and 100 mm. For each set, five lengths were randomly chosen without replacement from rectangular distributions of 0.1–20, 20.1–40, 40.1–60, 60.1–80, and 80.1–100 mm. The ten-line sets were similarly constructed, but with only two lines chosen from each of the five length categories. Each card also had a small pair of letters in the top left-hand corner. The first letter denoted the set (a–d) and the second the line in that set. The second letters were randomly assigned to the different lines. The different sets were assigned to the different experimental conditions using a counterbalanced design.

In all cases, the lines were arrayed in rows of five on a table in front of the participant and ordered a to e, f to j, and so forth, according to the second letters displayed on the labels. Thus, although in the closed conditions the participant had to remember the length of line, he or she did not have to remember the position of the particular lettered card on the table.

The key instructions for the open condition were:

> You will see (10/25) cards face up in front of you on the table. Your goal is to rank order the cards in terms of the lengths of the lines on them. You write the rank orders—from 1 shortest to (10/25) longest—on the assessment sheet alongside the letters. The letters have been randomly allocated to the lines. You are not allowed to physically rearrange the cards.

For the closed condition, these read:

> You will see (10/25) cards face down in front of you on the table. Your goal is to rank order the cards in terms of the lengths of the lines that are on the other side of them. You write the rank orders—from 1 shortest to (10/25) longest—on the assessment sheet alongside the letters. The letters have been randomly allocated to the lines. You are not allowed to physically rearrange the cards. You may only turn over each card once, and you may only have one card showing a line at one time.

The assessment sheets listed the appropriate letters for the card sets, with a space alongside for the ranking. In addition, participants also received a worksheet for each condition. The worksheets reflected the layout of the cards on the table and featured blank spaces laid out in rows of five, with the letters in each corner. Participants were instructed as follows:

> We also provide you with a worksheet with a small area for each card, as well as the sheet for making your final ranking on. We are interested in the processes you use to come up with ordering, so we would like to encourage you to use the worksheet as much as possible to record information for your decision-making. We also think you might find the worksheet useful to keep track of what is going on. We ask you to leave the worksheet (as well as the ranking sheet) with us at the end of the experiment.

The experimenter timed how long the participant took to perform each task. After each task the participant was asked to outline the process used to order the lines, and this description was recorded.

This and the succeeding experiments were all conducted after obtaining procedural approval from the University of Canterbury Human Ethics Committee. A key consideration in piloting for this and subsequent experiments was limiting session length in order to maintain high levels of participant interest.

Results

Two independent coders classified the predominant strategy used by each participant for each task into one of the following: *ranking*, *categorization*, *metric*, and *mixed*. This coding system was devised after viewing protocols from an earlier, unreported experiment. [3] Although coding decisions were made from all the material available—the descriptions provided by the participants and the worksheets—in practice a few features usually proved decisive. *Ranking* strategies were identified by the substantial absence of any numerical or categorical information other than the ranks themselves. Respondents often reported starting with the shortest (or longest) and then moving to the next shortest and so on. For the open conditions, the worksheets were often unused. In closed conditions, the worksheets often contained evidence of comparisons of pairs of line lengths. Occasionally, participants ranked from both the shortest and the longest line. *Categorization* strategies were identified by the participant stating that they had first categorized the lines (often using verbal labels such as "long," "very long," "short," etc.) and then ranking within the categories. The worksheets generally showed the categories too. A minimum of three categories was necessary for this code (since simple categorization into long and short is very similar to the two-ended ranking strategy). The number of categories had to be less than the number of lines. In *metric* strategies, the participant used a measure of length and then ranked from the measure. The different measures included actual length estimations (e.g., 6.5 cm), estimation of the percentage or proportion or fraction of the length of the label covered by the line, or drawings of the lines themselves. Such measures were easily visible in the worksheets. The coders tried to avoid the *mixed* strategy code. This strategy could arise if the participant changed strategy during the experiment and stated that he or she had ranked some lines with

[3] This experiment was identical to Experiment 1, except that participants were permitted repeat viewing of the lines in the closed conditions. The results were in line with our hypotheses, but the coding reliability (77 %) was rather low, mainly reflecting strategy changes.

Table 1 Numbers of participants using ranking or other (metric or categorization) strategies in each condition of Experiment 1

	10-Line Set		25-Line Set	
Measure	Closed	Open	Closed	Open
Ranking	0	9	0	9
Other	9	3	11	2

one strategy and some with another. A *mixed* strategy was also indicated if a substantial proportion, but not all, of the lines had categories or metrics assigned to them on the worksheets. Originally an *other* code had been included, but this proved unnecessary during piloting. Overall, the amalgamation of different types of measures into ranking, metric, and categorization was motivated by theoretical considerations, but the coding was not constrained by these considerations, since other codings remained possible, although they were unused. Coding was carried out blind as to whether the condition was open or closed (although, of course, not to the number of lines to be ordered). The coders achieved initial agreement of 88 % and resolved differences by discussion.

Of the 48 separate tasks performed by the participants, 18 were undertaken with a ranking strategy, 23 with a metric strategy, two with categorization, and five with a mixed strategy. As Table 1 shows, ranking strategies were clearly predominant in the two open conditions, whereas other strategies (one categorization and 19 metric) dominated in the closed conditions. Two-tailed tests of proportions showed significant differences in the use of strategies with the ten-card sets ($p < .001$), 25-card sets ($p < .001$), and combined ($p < .0001$). *No* participant used a ranking strategy in either closed condition.

Table 2 shows the results relating to the accuracy and time taken in the experiment. As a measure of the accuracy of sorting, we took the Spearman correlation coefficient between the final order arrived at and the correct order for each participant and task. (As a rule of thumb for interpreting the numbers, the misordering of one pair of adjacent line lengths in a set of 25 would lower rho from 1.0 to .999). The principal results were that participants took longer to order 25 line lengths than to order 10 [$F(1, 21) = 138.7$, $p < .001$], and they were less accurate in the former case [$F(1, 22) = 8.21$, $p < .01$].

Table 2 Average accuracies (Spearman's rho) and times taken to complete the sort in the four conditions of Experiment 1

	10-Line Set	25-Line Set
Accuracy (ρ)		
Open	.998	.982
Closed	.984	.963
Time taken (s)		
Open	129	813
Closed	422	1,103

Table 3 Numbers of participants who used different assessment measures of 10- and 40-line sets in Experiment 2

	10-Line Set			40-Line Set		
Measure	Closed	Open	Both	Closed	Open	Both
Rank	9	16	25	6	6	12
Categories	4	3	7	10	9	19
Metric	6	1	7	4	5	9
Mixed	1	0	1	0	0	0
Totals	20	20	40	20	20	40

Experiment 2

In Experiment 1, the participants were required to produce a rank ordering. In Experiment 2 we asked a different question: If the participants themselves can choose the final assessment measure, what measure do they provide? We hypothesized that they would choose not to provide rank orders when the stimuli were not all visible at once. We also hypothesized that they might choose to provide some other measure when the number of stimuli was relatively large.

Method

Two sets, one of ten and one of 40 cards, were prepared, each showing one line. Lines were between 0 and 100 mm in length and placed on one side of a card as before. Line lengths were uniformly distributed in the sets. Each participant, recruited as in Experiment 1, assessed each set of cards once. Assessments were made in two conditions, an open condition in which all the cards were visible, and a closed one in which all cards were initially face down and only one at a time could be (repeatedly) turned over. Half of the 40 participants (23 female, 17 male) performed the open condition, and half the closed condition, first. Similarly, half performed in the open card condition with the ten-card set and half with the 40-card set. Participants were issued a response sheet for each card set featuring only randomized card labels and a space for the assessment. The open assessment instructions were:

> You will see (10/40) cards face up in front of you on the table. Your task is to assess the length of the line shown on each card. You can choose for yourself the measure of line length that you use. For example, you may choose to categorize the lines as very long through middling to very short; you may choose to rank the length of the lines from shortest to longest; you may

> choose to assign a mark or a grade to each length (as happens in tests); you may choose some other measure. Please write your final measures on the response sheet alongside the letters.

The closed assessment instructions were similar, except for reminding participants that they must turn cards face down after viewing and only have one face-up card showing at once. There was no restriction on how many times each card could be turned over.

After each task, the participants were asked to describe the process that they had used in the task.

Results

The different assessment methods were independently categorized by two coders, who achieved 93 % initial agreement. In this experiment, the main determinant of the code assigned (ranking, categorization, metric, or mixed) was the final assessments themselves. The six discrepancies were resolved by discussion. Only one discrepancy concerned whether ranking was used, and this was resolved in the statistically conservative direction as a rank.

As Table 3 shows, ranks were less frequently used for the 40-line than for the ten-line set (sign test, $z = 3.33$, $p < .001$). In 13 instances, a participant used ranks to assess the ten-line set but not the 40-line set, and in no case did anyone do the reverse. Whether the assessment was done under open or closed conditions made no difference to the choice of ranking or other methods for the 40-card sets (Mann–Whitney U, $p > .05$), but ranking the ten-card set was more frequent under open conditions (Mann–Whitney U, $z = 2.24$, $p < .05$).

In brief, the results showed that the ranking measure was provided by the majority of participants only when the number of stimuli was relatively small and all of the stimuli were visible at once. Otherwise, either a category or a metrical measure was preferred.

Experiment 3

Could similar results be obtained when the stimuli to be judged were more complex than line lengths? In Experiments 3 and 4, the participants assessed the artistic quality of photographs. We chose artistic quality because, as a continuum to be judged, it is very different from line length. Whereas line length has an objective physical scale, artistic quality is notoriously subjective and it is unclear how people evaluate it. Assessments of artistic quality are complex and likely depend on multiple attributes that differ across individuals, and not based on any single, measurable physical dimension. (For implications of multidimensionality for ordering, see Kemp & Grace, 2010.) One could even ask whether it makes sense to think about a dimension of artistic quality at all. (For previous research on artistic judgments and how they might be formed, see, e.g., Dutton, 2009; Leder, Belke, Oeberst, & Augustin, 2004; Lindauer & Long, 1986.) A rather different consideration is that, although asking people to assess line length is an artificial task—why not simply measure with a ruler and record the answers?—people often do assess complex dimensions like artistic quality, sometimes ending up with a rank ordering, and sometimes not. Consider, for example, reviewers rating the excellence of films or restaurants, a professor grading essays, or judges assessing photographs submitted to a competition.

In Experiment 3, participants provided a rank ordering of the perceived artistic merits of photographs, which were viewed singly and only once. Our prediction was that some participants would use either a metric or categorization of the photographs before ordering.

Method

A total of 20 photographs that had been entered into an artistic competition were downloaded from the Web. The respondents, who were tested individually, were seated in front of a computer screen and instructed as follows:

> In this experiment, you will see 20 photographs on the computer screen. Your goal is to rank order them in terms of artistic quality.
> The photographs are displayed one at a time. You may only view each photograph once. When you press the computer space bar, a new one will appear. There is no time limit on how long you can view them, but we suggest you spend no more than a minute on each one.
> We also provide you with a worksheet with a small area for each photograph, and a sheet for making your final ranking on. We are interested in the processes you use to come up with ordering, so we would like to encourage you to use the worksheet as much as possible to record information for your decision-making. We also think you might find the worksheet useful to keep track of what is going on. We ask you to leave the worksheet (as well as the ranking sheet) with us at the end of the experiment.

The photographs were labeled A to T, and the ranking sheet simply listed the numbers 1 to 20.

When the respondents had finished the ranking, they were asked to comment in detail on how they had gone about the task, and an experimenter recorded the details. The experimenter also recorded the time taken from viewing the pictures to finalizing the ranking for 20 of the 26 participants.

Eight of the participants were male. The minimum, modal (n = 14), and median ages were all 21 years; the oldest participant was 38.

Results

The strategy-coding scheme was similar to that used in Experiment 1, with the main exception being that representations of the photographs were disregarded as evidence for a metric strategy. The initial coder classified eight participants as using ranking, 12 as categorizing, four as metric, and two as mixed or unclear. A second coder provided similar results (eight ranking, 12 categorizing, four metric, and two mixed or unclear), but the two coders disagreed about four participants: In all four cases of disagreement, a participant classified as categorizing by one coder was classified as using a mixed or unclear strategy by the other. All 26 participants used their worksheets to record verbal and/or pictorial details of the pictures themselves. Many also included quality keywords (e.g., cool, boring, stylish), and all four of those using metric strategies included scores out of 10 in their worksheets. When participants categorized, they later ordered within the categories. Coding judgments of categorization were largely based on participants' descriptions of how they had proceeded (rather than the worksheets). Although participants could have categorized the photographs on the basis of type or content, in fact all their categorizations were based on quality. Coding disagreements arose because some participants used a combination of processes. Nonetheless, the key result is that the majority of the participants chose to employ either categorization or (less frequently) a metric before rank ordering the photographs.

The mean time taken to view the photographs averaged 712 s (with a range from 309 to 1,278 s), and the mean time taken to complete the rankings after viewing the photographs was 329 s (ranging from 120 to 661 s). An analysis of variance showed no significant differences in either time with the four-way classification (regardless of who did it). We also investigated whether a simple ranking versus other classification system was related to the timing. We observed a suggestive tendency [$t(18) = 2.05$, $p = .051$] for viewing times to be longer (average: 848 vs. 652 s) and for sorting times to be shorter [$t(18) = 1.90$, $p = .074$] with the ranking process (average: 257 vs. 367 s).

Experiment 4

Experiment 4 resembled Experiment 2 in allowing participants to choose their own form of final assessment, but the participants assessed the artistic merit of photographs. We predicted that participants would often choose not to provide a ranking, particularly when the stimuli were not all visible at once.

Method

All respondents were asked to assess the artistic quality of the same 20 photographs used in Experiment 3, but were free to choose their measure of artistic quality. The wording of the key instructions was:

> You can choose for yourself the measure of artistic quality that you use. For example, you may choose to categorize the photographs as very good through average to very poor; you may choose to rank the quality of each of the photographs from first through to 20th; you may choose to assign a mark or a grade to each (as happens in tests); you may choose some other measure. Please write your final measures on the response sheet provided.

The final response sheet consisted of a page headed up "Final photograph measure" and listing the letters A to T. Respondents were also provided with a completely blank worksheet and asked to nominate the assessment they used at the very end of the experiment.

Respondents were alternately assigned to one of two viewing conditions. In the *serial* condition (similar to the closed condition in the line experiments), the photographs were presented in random orders one at a time for as long as the respondent chose, with no opportunity given for repeat viewing. These viewing conditions were identical to those of Experiment 3, and respondents were instructed similarly.

In the *array* condition (similar to open), respondents were presented with all 20 pictures simultaneously. The pictures were arranged in four rows with five photographs each (as thumbnail sketches). The respondent could enlarge any one picture by double-clicking on it. No time limits and no restriction were imposed on how many times each photograph could be enlarged.

Twenty naïve respondents were recruited and paid for each condition. Overall, 13 of the respondents were male, and the ages ranged from 19 to 60 years, with a median age of 22 years.

Results

Table 4 shows the assessment measures for each condition as nominated by the respondents. The obvious result is that ranks were rarely used, even in the array presentation condition. Of the two respondents who did report ranks, one did so after first categorizing the photographs into three groups by artistic quality. The other employed a "limited ranking" measure. The photographs were first categorized into six types of photographs (e.g., artistic, outdoor) and then ranked within each category. (Thus, there was no attempt to compare the artistic quality of photographs of different types.) No other participant attempted to categorize according to types of photographs: All other categorizations were based on artistic quality.

Table 4 Numbers of participants who used different final photograph measures for the two conditions of Experiment 4

	Serial	Array
Mark out of 10	13	14
Mark out of 20	2	1
Five or fewer grades	5	2
Complex letter grading (A+, A, etc.)		1
Rank or limited rank		2
Mean number of categories used	6.3	7.6

Of the (seven) respondents who chose a measure when five or fewer grades were nominated, three reported in verbal terms, one as three letter grades, two as numbers (from 1 to 5), and one in both numbers and words. We also analyzed the number of different responses actually given. A respondent who stated that she had marked out of 10 did not necessarily use all of the numbers between 1 (or 0) and 10. In fact, of the 27 respondents who gave marks out of 10, none gave a 0, only three gave any 1 s, only five gave any 10s, and none gave both a 1 and a 10. On the other hand, several of the respondents used half marks. Taken over the 38 respondents who did not use a final rank measure, the number of response categories actually used ranged from 3 to 19 (the next highest was 11), with a median of 7 and an average of 6.9. Although the number of categories used was slightly less in the serial condition (see Table 4) the difference was not statistically significant [$t(36) = 1.44$, n.s.]. The respondent who produced 19 different responses clearly could have used these as the basis for a near total ordering, but he chose not to do so: As was clear from his worksheet, the numbers that he produced (e.g., 6.75 and 3.3) arose from his grading each picture on four attributes ("novelty," "use of photographic devices," etc.) on a simple 0–10 scale and then averaging. No other respondent explicitly used a multiattribute procedure.

Experiment 5

The dependent variables in Experiments 1–4 were either the measure chosen or the method used to obtain a ranking. In Experiment 5, the measure used was an independent variable—either participants ranked line lengths or they estimated the length as a percentage of the label covered. The main dependent variable was the time taken to evaluate the lengths of lines. The chief hypothesis was that we would find interactive effects of the measure chosen with the number of lines to be estimated and whether the lines were visible only one at a time or were continuously visible. That is, as the number of cards increased, there would be markedly more slowing with the ranking than the percentage measure, and ranking would be slower in the closed conditions. This follows from the argument outlined in the introduction, because of the extra memory load required by the transformation functions.

Method

Thirty-two paid participants (22 female, ten male) with median ages in the range 21–25 years were recruited around the university. All were naïve to line (and other) sorting experiments and were tested individually. Each took part in two separate sessions run on different days. In each session, the participant took part in four conditions. In one session, he or she was instructed to construct a rank order of the line lengths; in the other he or she estimated the percentage of the length of the label on the card (running from 0 % to 100 %) that the line stretched across. In each session, the four conditions were ten cards open, ten cards closed, 25 cards open, and 25 cards closed. The order of these conditions and the session task were counterbalanced. The actual cards and the instructions in the rank-ordering session were identical to those used in Experiment 1, and a similar record was made of the methods used by the participant. The key instructions in the percentage estimation (open) condition read:

> You will see (10/25) cards face up in front of you on the table. Your goal is to assess the lengths of the lines on them. The line lengths are to be measured as the percentage of the width of the label on the card that is covered by the line. It can range from 0 % (no line at all) to 100 % (the line stretches to the full width of the label—not the card). You write the percentages on the assessment sheet alongside the letters. The letters have been randomly allocated to the lines. You are not allowed to physically rearrange the cards.

Participants were informed that they were being timed and were asked to do the tasks as quickly as possible.

Results

Two independent coders classified the dominant strategy used by each participant in performing each rank-ordering task as simply *ranking*, *categorizing*, *percentage (or fractional) measure*, and *other metric* (usually, the participant would simply reproduce the line on the worksheet and then compare all of the recorded lines; occasionally, participants recorded a line length in estimated centimeters or millimeters). [4] The two coders achieved 92 % agreement on the initial coding, and differences were resolved after discussion. Of the ten

[4] The change in coding scheme was motivated partly by wanting to dispense with the mixed strategy code, and partly in an attempt to separate out the particular metric measure (percentage or proportion) used in the percentage estimation task.

discrepancies, four concerned whether an ordering had been achieved by simple ranking or via some other, intermediate estimation.

Table 5 shows the strategies used to perform the rank ordering in the four conditions. The obvious result is that simple ranking predominated in the open conditions, but some intermediate estimate was employed in the two closed conditions. An analysis of variance based on a two-point classification of the strategies into "ranking" versus "other" showed a suggestive effect of the number of cards [$F(1, 131) = 3.9$, $p = .057$], a significant difference between the closed and open conditions [$F(1, 31) = 119.1$, $p < .001$], and no significant interaction [$F(1, 31) = 1.3$, n.s.]. These results were very similar to those of Experiment 1, as expected.

Table 6 shows the average completion times for the different conditions of the experiment. Consistent with our prediction, we found substantial interactive effects [task, $F(1, 31) = 125.5$, $p < .001$; number of cards, $F(1, 31) = 159.9$, $p < .001$; open/closed, $F(1, 31) = 30.4$, $p < .001$; Task × Number of Cards, $F(109.4) = p < .001$; Task × Open/Closed, $F(1, 31) = 29.8$, $p < .001$; Number of Cards × Open/Closed, $F(1, 31) = 14.7$, $p < .001$; three-way interaction, $F(1, 31) = 19.3$, $p < .001$]. One way to understand the pattern of results would be to consider first the results for the percentage estimation task. As the table shows, it made little difference whether the cards were all visible at once (open) or only one at a time (closed), but if the number of cards was increased, the task took longer. Rank ordering of ten cards in the open condition took about as long as percentage estimation, but when only one card at a time was visible (ten card, closed), the task was lengthened considerably. Increasing the number of cards lengthened the ordering task, and indeed did so disproportionately. Many participants reported finding rank ordering in the 25-card closed condition difficult.

Including session order as a variable produced no significant interaction effects. We compared whether using simple ranking or other strategies affected completion times in each condition of the rank-ordering tasks, but no significant ($p < .05$) or suggestive ($p < .1$) effects were found. (Note, however, that the power of these analyses was low.)

Overall accuracy, as measured by Spearman rho correlation coefficients, was quite high (average $\rho = .98$, $SD = .025$], but no significant differences emerged between tasks, conditions, or the interactions. Effectively, the participants maintained a high standard of accuracy in the different tasks and conditions, and appear to have taken whatever time they needed to maintain that standard.[5]

Table 5 Percentages of Experiment 5 participants ($n = 32$) who used each of four different strategies to perform rank ordering under the four different conditions

Condition	Rank	Categorization	Percentage	Metric
10-card open	94	0	6	0
10-card closed	16	16	25	44
25-card open	81	9	9	0
25-card closed	13	19	31	38

Table 6 Average completion times in seconds (with *SD*s in parentheses) for the four different conditions and two different tasks of Experiment 5

Condition	Percentage Estimation	Rank Ordering
10-card open	65.8 (26.8)	81.2 (35.5)
10-card closed	77.8 (30.2)	171.9 (60.8)
25-card open	190.9 (76.3)	525.9 (211.9)
25-card closed	191.7 (83.1)	984.3 (329.4)

General discussion

Experiments 1 and 3 showed that, when people were asked either to order lines according to their lengths or photographs according to their artistic quality, they often chose first to construct either a metrical measure or a set of categories, and then they used this measure as a basis for their rankings. This was particularly true when the stimuli were not all simultaneously visible. In Experiment 1, when each line length could only be viewed once, all the participants in the closed conditions used a metrical or category measure as an intermediate step. In general, straightforward ranking was the preferred method only when all the stimuli were simultaneously visible and no remembering was necessary.

In Experiments 2 and 4, in which the participants themselves chose the type of assessment they would deliver, the majority chose a ranking only when ten simultaneously visible lines were to be assessed. By contrast, no participants chose ranking when the photographs were presented one at a time in Experiment 4.

The qualities of line length and artistic quality are very different. However, whether an intermediate measure was used for a ranking or whether people chose to use a ranking measure was resolved fairly similarly for both qualities. Given that our theoretical expectations about how and when orders might be obtained were not linked to any particular dimension or kind of dimension, this was the expected result. One way to think of our results is that they are taken from extreme ends of a continuum of kinds of quality. Length appears to be a simple quality, in that it is not apparently made up from other qualities. Moreover, commonly accepted ways of measuring

[5] In total, five participants had $\rho < .9$ in at least one of the eight conditions of the experiment. No participant had $\rho < .9$ in more than two conditions. When these five participants were excluded from the analysis of variance of the completion times, neither the pattern of the results nor the statistical significance levels changed.

length are available. By contrast, artistic quality does not appear at all simple, and it is notoriously difficult even to know what accurately estimating this quality might mean (e.g., Dutton, 2009). It is tempting to believe that similar results would be found for a variety of qualities of intermediate complexity. However, clearly, this will be a matter for further research.

Given the relative lack of prior studies, we thought it important to demonstrate that the basic hypothesis would hold not only with more than one dimension, but also with more than one type of dependent measure. This was especially true because the most direct test of the hypothesis, which was employed in Experiments 1 and 3, relied both on participants' ability to introspect on the process they used to construct the rank orders and on the ability of the coders to characterize that process from the evidence provided. If, as was quite possible, a participant had no conscious access to the process he or she was using, or if the process represented on the worksheets and described to the experimenter was not actually the process they employed, or if the coders misinterpreted this evidence, then the results would not be accurate (e.g., Nisbett & Wilson, 1977). On the other hand, in both experiments, at least the first two sources of inaccuracy should have biased the results away from the basic hypothesis, since one would expect failures to report on an intermediary measure that actually had been used rather than reports of using a process that actually had not been. The introspection issue was less critical in Experiments 2 and 4, in which the crucial measure was a behavioral choice. Moreover, in Experiment 4, no coding was necessary at all. An additional cross-check was supplied by Experiment 5, which relied entirely on behavioral measures, and showed that it took longer to rank order the length of lines than to provide a simple measure of their length when the lines were not all visible and as the number of lines increased. The results of all of the experiments were in line with the basic hypothesis, even though very different types of dependent variables were employed.

Thus, our results confirm the hypothesis posed in the introduction—that ordering is often not a natural and easy task for humans, and that in specifiable situations constructing an order is facilitated by prior categorization or by use of a metric. This conclusion holds not only for a dimension such as line length, in which a ratio scale is clearly available, but also for the much less well-defined measure of artistic quality. It is also worth remarking that our basic finding is in line with that of Alwin and Krosnick (1985) for values estimation, although their work did not seek to investigate particular conditions under which ordering might be difficult.

In Experiments 1 and 5, the task of distinguishing the line lengths was not easy when 25 or 40 different lengths had to be ordered, and even less easy in the memory conditions, in which two line lengths could not be directly compared. However, the results showed that almost all of the participants in these experiments chose to maintain a high degree of accuracy, and took the time necessary to do this. These results are quite different from those obtained from perceptual absolute-identification experiments, in which accuracy is typically low (e.g., Laming, 1997, ch. 10; Stewart, Brown, & Chater, 2005; Ward & Lockhead, 1971). The obvious explanation for the different results is that the procedures used in our experiments, particularly the labeling of the stimuli and the provision of the worksheets, removed much of the difficulty present in the absolute-identification research. Similarly, our procedures enabled memories that are usually only privately accessible (consider, e.g., a professional assessor of a photograph trying to decide on its artistic quality) to be at least partly visible to observers. However, our procedures make it difficult to compare this work with previous work on memory limitations. Nonetheless, it is notable that in Experiment 4 the number of different categories or grades actually used fitted nicely into the seven plus or minus two that might be predicted from a consideration of working memory (see, e.g., Miller, 1956).

One theoretical connection to previous thinking about the psychophysics of sensation is worth remarking. Individual comparisons of some kind appear to underlie all ordering, and a number of researchers have stressed that judgments of qualities such as loudness are made in context and depend on comparisons with other stimuli (e.g., Laming, 1984, 1997; Stewart et al., 2005; Ward & Lockhead, 1971). The results above generally support this position. However, our focus was somewhat different. Most importantly, much of the earlier work seems to have concerned the form of comparisons between different stimuli, some present at the time of comparison and some remembered. For example, Laming (1984) suggested that a comparison between a line just seen and one now in view might be simply expressed by the previous one being much shorter, a little shorter, about the same, a little longer, or much longer. By contrast, in this article we have been concerned to identify the kind of measure that might be available in memory, and that could subsequently be used to make these comparisons and orderings. That is, what information might the participant use to determine that the previous line was "much shorter" or "a little longer"?[6]

[6] An important question that we have not addressed in this article is how the remembered quality, whether ranking, metric, or categorization, gets formed in the first place? For instance, if a participant in Experiment 4 decided to allot a mark out of 10 for artistic quality, how might she set up her basic grading system? She might give the first photograph encountered a mark that was referenced to some photographs that she has seen in the past, and then use differences from this baseline, but many other alternatives were possible. Rather different types of experiments would be needed to answer such questions (cf. Laming, 1997).

Although both theory and experimental results have indicated that either a metric or a categorical measure would be preferable to remembering simple ordinal information under the prescribed conditions, they do not provide a strong indication as to which of these might be preferred. Ideally, one could envisage a theory that could predict what the optimal measure to employ might be in different situations, but such a theory goes well beyond what we have attempted here.

On the empirical side, when line lengths were ordered (Exps. 1 and 5), metric measures were used much more often than categorization. By contrast, in Experiment 2 more participants chose to report a categorization than a metric measure for the 40-line sets. In Experiment 3, participants more frequently chose to categorize than to construct a metrical measure when they ordered the photographs according to artistic quality. Overall, the safest conclusion from the present results is that, although there do appear to be experimental conditions that favor the use of categorization over a metric measure, or vice versa, the key principles favoring the use of one or the other are not clear, and remain a matter for further research.

R. Steingrimsson (personal communication, December 16, 2013) pointed out that it is not clear whether the metrics used by the participants in these experiments constituted interval or ratio scales. For Experiments 1, 2, and 5, where the metrics used were judgments of actual length (e.g., "5 cm"), percentages of the label covered, or attempts to reproduce the line itself, it seems likely that they were ratio scales, although probably not perfectly accurate ones (Steingrimsson & Luce, 2007). However, the natures of the metrics employed to assess artistic quality in Experiments 3 and 4 are not clear. We note, first, that the metrics used by different individuals are likely to have been quite different: If A and B did both employ interval scales to assess artistic quality, there is no reason why they should have used interval scales that were in any way similar. For example, A might have assessed how happy each picture made him feel, and B some measure of how balanced the compositions appeared to her. If an individual were to employ an ordinal scale as a metric, this would be subject to exactly the same uniqueness complication as a rank order. By default, then, it seems to us likely that individuals who, for example, graded the pictures out of 10 actually did use idiosyncratic interval scales.

Our research shows that, particularly when stimuli must be recalled because they are not physically present, people will often choose to order the stimuli on some dimension after first constructing a categorical or metric measure of the dimension. In addition, they often prefer to provide assessments of the stimuli that are categories or metric measures rather than orderings. These results are relevant for understanding how people construct orders.

Nonetheless, our results are not unexpected. Clearly, if people order stimuli, they must be doing it on the basis of comparisons of some quality of the stimuli. Nor is it particularly surprising that, as the number of stimuli increases or in circumstances in which many of them must be remembered, ordering them becomes more difficult. On the other hand, it is important that these findings have been demonstrated, because they have an important implication: Our results provide strong evidence that the internal measures that people use in situations in which they must evaluate both present and absent stimuli are unlikely to be ordinal scales. Instead, memory constraints make it more effective to use metrics or categorizations. In the introduction, we showed that hypothesized difficulties in simply ordering stimuli were related to the ∞-point uniqueness of all ordinal scales. The experimental finding that ordering is not straightforward in circumstances in which some stimuli are absent thus implies that using any internal ordinal scale in these circumstances would be similarly difficult. Thus, if people do have internal measures of such qualities as line length and artistic quality, it seems unlikely that these representations could usefully be orderings or ordinal scales. It is much more likely that the representations are ratio scales, interval scales, or categories. Thus, our finding contradicts the apparent belief of many researchers that individuals might employ internal representations that are ordinal scales or orderings. Depending on the nature of the measurement situation, they also provide a counterargument to the frequent recommendation to use ordinal scales for data analysis when one is uncertain about the underlying metric.

Author note We are grateful to Sarah Beggs, May Chan, Anna Clark, and Jessica Richardson for their help in collecting data and running the experiments, and to Conor Dolan, Donald Laming, Ragnar Steingrimsson, and an anonymous reviewer for commenting constructively on earlier drafts.

References

Alwin, D. F., & Krosnick, J. A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly, 49,* 535–552.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Erlbaum.

Bolanowski, S. J., & Gescheider, G. A. (Eds.). (2013). *Ratio scaling of psychological magnitude: In honor of the memory of S. S. Stevens*. London, UK: Taylor & Francis.

Chignell, M. H., & Patty, B. W. (1987). Unidimensional scaling with efficient ranking methods. *Psychological Bulletin, 101,* 304–311. doi:10.1037/0033-2909.101.2.304

Davison, M. L., & Sharma, A. R. (1988). Parametric statistics and levels of measurement. *Psychological Bulletin, 104,* 137–144. doi:10.1037/0033-2909.104.1.137

Dutton, D. (2009). *The art instinct: Beauty, pleasure, and human evolution*. Oxford, UK: Oxford University Press.

Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research, 45,* 43–57.

Kemp, S., & Grace, R. C. (2010). When can information from ordinal scale variables be integrated? *Psychological Methods, 15,* 398–412. doi:10.1037/a0021462

Kemp, S., & Grace, R. C. (2012, 21 March). *Ordinal scales in psychology*. University of Canterbury Working Paper. Retrieved from www.psyc.canterbury.ac.nz/people/kemp.shtml

Kerst, S. M., & Howard, J. H. (1978). Memory psychophysics for visual area and length. *Memory & Cognition, 6,* 327–335. doi:10.3758/BF03197463

Knuth, D. E. (1998). *The art of computer programming: Vol. 3. Sorting and searching* (2nd ed.). Reading MA: Addison-Wesley.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: Vol. I. Additive and polynomial representations*. Mineola, NY: Dover.

Krueger, L. E. (1989). Reconciling Fechner and Stevens: Towards a unified psychophysical law. *Behavioral and Brain Sciences, 12,* 251–320.

Laming, D. (1984). The relativity of "absolute" judgements. *British Journal of Mathematical and Statistical Psychology, 37,* 152–183.

Laming, D. (1997). *The measurement of sensation*. Oxford, UK: Oxford University Press.

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology, 95,* 489–508.

Lindauer, M. S., & Long, D. A. (1986). The criteria used to judge art: Marketplace and academic comparisons. *Empirical Studies of the Arts, 4,* 163–174.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. London, UK: Sage.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750–751.

Luce, R. D. (1996). The ongoing dialog between empirical science and measurement theory. *Journal of Mathematical Psychology, 40,* 78–98. doi:10.1006/jmps.1996.0005

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement: Vol. III. Representation, axiomatization, and invariance*. Mineola, NY: Dover.

Maxwell, S. E., & Delaney, H. D. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97,* 85–93. doi:10.1037/0033-2909.97.1.85

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Narens, L. (1996). A theory of ratio magnitude estimation. *Journal of Mathematical Psychology, 40,* 109–129. doi:10.1006/jmps.1996.0011

Narens, L. (2002). *Theories of meaningfulness*. Mahwah, NJ: Erlbaum.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259. doi:10.1037/0033-295X.84.3.231

Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 608–629. doi:10.1037/0096-1523.31.3.608

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

Parducci, A. (1982). Category ratings: Still more contextual effects. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 89–105). Hillsdale, NJ: Erlbaum.

Petrov, A. A. (2011). Category rating is based on prototypes and not instances: Evidence from feedback-dependent context effects. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 336–356. doi:10.1037/a0021436

Poulton, E. C. (1989). *Bias in quantifying judgements*. Hove, UK: Erlbaum.

Roberts, F. S. (1979). *Measurement theory*. Reading, MA: Addison-Wesley.

Rokeach, M. (1973). *The nature of human values*. New York, NY: The Free Press.

Steingrimsson, R., & Luce, R. D. (2007). Empirical evaluation of a model of global psychophysical judgments IV: Forms for the weighting function. *Journal of Mathematical Psychology, 51,* 29–44.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680. doi:10.1126/science.103.2684.677

Stevens, S. S. (1955). On the averaging of data. *Science, 121,* 113–116.

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

Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics, 9,* 73–78. doi:10.3758/BF03213031