The language of magnitude comparison

William J. Matthews

Alexandra S. Dylman

University of Essex

Published as: Matthews, W.J., & Dylman, A.S. (2014). The language of magnitude comparison. *Journal of Experimental Psychology: General, 143*, 510-520.

Correspondence Address:

William J. Matthews
Dept. of Psychology
University of Essex
Colchester
CO4 3SQ
United Kingdom

Tel.: +44 1206 873818
Fax: +44 1206 873801
E-mail: will@essex.ac.uk

Abstract

When two objects differ in magnitude, their relation can be described with a "smaller" comparative (e.g., "less", "shorter", "lower") or a "larger" comparative (e.g., "more", "taller", "higher"). We show that, across multiple dimensions and tasks, English speakers preferentially use the latter. In sentence-completion tasks, this higher use of larger comparatives (HULC) effect is more pronounced when the larger item is presented on the left (for simultaneous presentation) or second (for sequential presentation). The HULC effect is not diminished by making the two items more similar, but it is somewhat lessened when both objects are of low magnitude. These results illuminate the processes underlying the judgment and representation of relative magnitudes.

Keywords: Comparative judgment; Language; Magnitude comparison

The language of magnitude comparison

It is often argued that there is no "absolute judgment" and that all judgments are comparisons of one thing with another (e.g., Laming, 2003; Matthews, 2013; Mussweiler & Epstude, 2009; Stewart, Brown, & Chater, 2005). Correspondingly, much of our speech and writing describes the relative magnitudes of perceptual, social, and economic quantities (e.g., "one light is brighter than the other"; "he is less powerful than she"; "losses loom larger than gains"). The language that people use when comparing magnitudes is important because language both reflects and shapes our representations of the world (e.g., Boroditsky, 2000; Winawer et al., 2007), and understanding the language of magnitude comparison will clarify how people construct and construe a fundamental aspect of their experiences.

The present work investigates one key aspect of this issue, namely the *direction* of comparative adjectives used by English speakers. For virtually all dimensions, a difference in magnitude may be described by a word implying greater magnitude or by a word implying lesser magnitude. That is, there is a set of "larger" comparatives (including "more", "bigger", "greater", "taller", "higher", "longer", "thicker", "wider", "heavier") and a set of "smaller" comparatives (including "less", "fewer", "shorter", "lower", "thinner", "lighter"). Although there is not a perfect one-to-one pairing between the "larger" words and the "smaller" ones, the relation between two objects can almost always be equally well-described with either type of word. For example, the magnitude relations captured by "the redhead is taller than the blonde" and "there are more rats than people" are logically equivalent to "the blonde is shorter than the redhead" and "there are fewer people than rats".

Given that the same magnitude relation can be expressed in different ways, a basic question arises: Are there systematic patterns in people's choice of comparative adjective? Do they, for example, use "smaller" and "larger" comparatives equally often when describing a given pair of objects? Or is there a preference for one form over the other? And is the choice affected by other factors, such as the spatial or temporal structure of the items or the size of the magnitude difference between them?

Some of these issues have been discussed by linguists (e.g., Allan, 1986; Cruse, 1976), but the analysis is typically formal rather than empirical. Similarly, word frequencies provide limited information because there is semantic overlap between comparatives (e.g., "larger", "bigger", and "greater" may all be used to describe the same relation), and because some comparatives are homographs (e.g., "lower" may mean "more low", "to make more low", or "to look threatening"). More importantly, neither abstract analysis nor examination of word frequencies permits manipulation of the to-be described objects. In this paper we report nine experiments which

examined people's use of "smaller" and "larger" comparatives when they described pairs of objects whose temporal, spatial, and magnitude relations were systematically manipulated.

The importance of comparative language

The way in which magnitudes are described can have profound effects on judgment and preference. One famous example comes from framing effects on risky choice: When faced with a disease that threatens 600 people, participants largely prefer to save 200 lives for sure rather than taking a one-third chance of saving everyone and a two-thirds chance of saving no-one. However, when the former option is framed as losing 400 lives for sure, people's preferences switch to favouring the risky option (Kahneman and Tversky, 1984). Thus, whether a given end-state magnitude is framed as a gain or a loss has a large influence on judgment. Similarly, when choosing between a smaller reward in the near future and a larger reward in the more distant future, participants' choices become more impatient when the delays are described in terms of lengths of time (e.g., "8 weeks from today") rather than equivalent calendar dates (e.g., "May 16th") (Read et al., 2005). In short, the language used to describe a given magnitude – be it a length of time, an amount of money, or a death toll – can powerfully influence the way that people compare different magnitudes.

Other research has explicitly examined how the choice of comparative adjective can influence judgments. If people are asked to compare a target quantity (e.g., the length of the river Elbe) with an arbitrary "anchor" value (e.g., 550 km), their subsequent best-estimate of the target quantity is heavily biased towards the anchor, irrespective of its relevance and the participant's expertise and motivation (e.g., Englich, Mussweiler, & Strack, 2006; Matthews, 2011). Importantly, Mussweiler and Strack (1999) found that when the initial comparison employed a "larger" adjective (e.g., "Is the river Elbe longer than 550 km?"), participant's subsequent best-estimates were greater than when the comparative question used a "smaller" adjective (e.g., "Is the river Elbe shorter than 550km?"). They interpreted this as evidence that people adopt a hypothesis-consistent testing strategy when answering the comparative question (Chapman & Johnson, 1999; Klayman & Ha, 1987; Mussweiler & Strack, 1999). For example, when the comparison question includes a "smaller" comparative, participants retrieve knowledge consistent with the idea that the target is less than the anchor. From this perspective, the choice of adjective shapes the semantic information that is activated when forming a comparative judgment, with subsequent effects on people's best estimates of the target quantity. Similarly, Choplin (2010) has found that the choice of comparative adjective ("fatter" vs.. "thinner") can affect people's memory for women's body size.

The framing of comparisons is also important in psychophysics. For example, Schneider and Komlos (2008) found that the effect of attentional cues on perceived visual contrast depends on

whether participants are making a comparative judgment ("which target has higher contrast?") or an equality judgment ("are the two targets equal in contrast?").

Finally, the language of comparison has been extensively studied by scientists interested in relational reasoning. For example, researchers have studied the time taken to solve transitive inference problems such as: A is better than B; B is better than C; who is best? The choice of adjective influences the speed with which such problems are solved (e.g., Clark, 1969) and these effects have contributed to debates about the role of linguistic processing, spatial imagery, and deductive logic in solving such problems (e.g., Clark, 1969; De Soto, London, & Handel, 1965; Sternberg, 1980; see Goodwin & Johnson-Laird, 2005, for a review).

In short, the linguistic description of magnitude relations can have powerful effects on judgment and choice across a wide variety of tasks and domains. Understanding how people spontaneously describe magnitude relations, and the factors that modulate these choices, is therefore a topic of considerable general importance. Whereas the foregoing studies examined how experimenter-provided language can influence people's judgements, choices, and reasoning, the current work examines the language that people spontaneously produce when describing the relations between items, and the factors that influence their choice of language.

Theoretical possibilities

In the current work, we took a novel experimental approach in which participants were presented with pairs of objects and asked to describe the relation between them in the language that they would naturally use. By manipulating the objects and their spatial, temporal, and magnitude relations, we examined the following hypotheses about the language of magnitude comparison:

The ambivalence hypothesis. People may use both "larger" and "smaller" comparatives equally often, reflecting their logical equivalence (e.g., Hunter, 1957). We refer to this as the ambivalence hypothesis. On the other hand, reaction time studies show that people are faster to verify sentences such as “Dogs are bigger than cats” than sentences such as “Cats are smaller than dogs” (e.g., Holyoak, Dumais, and Moyer, 1979), suggesting that “larger” comparatives may be more accessible and, therefore, widely used when people are spontaneously describing objects in front of them. To anticipate a key result of the current paper, we find this to be the case across a wide range of dimensions, tasks, and presentation formats.

The sequential processing hypothesis. The order in which items are processed has pronounced effects on perceptual, social, and economic judgments (e.g. Damisch, Mussweiler, & Plessner, 2006; Matthews & Stewart, 2009a, 2009b). In particular, people seem to be particularly sensitive to the *direction* of change from one item to the next, using the first item as a standard when judging the subsequent item (e.g., Laming, 1995; Stewart et al., 2005). Correspondingly, the

direction of change may well determine how people construe and describe the relation between two items, with a small-to-large transition being labelled "larger" and a large-to-small sequence being labelled "smaller". That is, if items are presented sequentially, the change in magnitude (increasing or decreasing) may dictate the choice of comparative adjective ("larger" or "smaller", respectively).

The same idea can be extended to simultaneously-presented items, but the specific predictions will depend on the assumptions one makes about the order in which the items in a pair are processed. One possibility is that the tendency of English speakers to process from left to right (e.g., Dickinson & Intraub, 2009) will mean that the leftmost item will be processed first and serve as the reference point, implying that a small-large spatial arrangement will lead to the use of a "larger" comparative whereas a large-small layout will lead to a "smaller" comparative. We refer to this as the *left-right-sequential hypothesis*.

Alternatively, magnitude itself may define the processing sequence. In the case of objects that differ in physical size, there is ceterus paribus greater likelihood of initially fixating the larger item, and the same may apply for symbolic magnitudes. Such preferential initial fixation would mean a higher proportion of large-to-small than small-to-large processing and, correspondingly, an overall tendency to use "smaller" comparatives more often than "larger" ones. We label this the *size-based sequential hypothesis*.

The location-matching hypothesis. Comparative sentences often have the form: "A is [comparative adjective] than B", where the to-be-compared object (A) comes at the start/left of the sentence and the reference object (B) comes at the end/right. If people match these word locations to the spatial locations of the objects (presented simultaneously), then the right-most item in a pair will serve as the reference point. Objects with a small-large arrangement will then evoke a "smaller" comparative while a large-small arrangement will be labelled "larger".

The similarity hypothesis. Any difference in the use of "smaller" and "larger" comparatives might be driven by the size of the difference between the objects, such that the relation is more likely to be labelled "larger" as the difference in magnitudes increases. Hummel and Holyoak (2001) incorporated this idea in a computational model of transitive inference wherein objects are mentally mapped onto a symbolic spatial array. They suggested that a "smaller" comparative implies that both objects are low magnitude and correspondingly '"crowded" together near the bottom of the array' (p. 10), whereas "larger" comparatives do not entail such crowding. An alternative justification for the same prediction is that people may tend to use "larger" to indicate that "the difference between the items is large" and "smaller" to indicate that the "difference is small".

The magnitude hypothesis. The choice of comparative may depend on the items' position on the focal dimension. Linguists often classify "larger" comparatives as unmarked, meaning that they

can "be used in a neutral way to compare the relative degrees of two items on a scale" (Goodwin & Johnson-Laird, 2005, p. 471) whereas "smaller" comparatives are "marked" and are reserved for items at the lower end of the scale (e.g., Clark, 1969). Thus, for two high-magnitude items, "largeness" may seem most relevant so that a "larger" comparative is preferred, and vice-versa for low-magnitude items. Consistent with this, there is a semantic congruity effect in comparison response times such that participants 'can more quickly select the smaller of "mouse" and "flea," whereas they can more quickly select the larger of "hippo" and "moose"' (Holyoak & Mah, 1981, p. 197).

These ideas and predictions are summarized in Table 1. Note that these possibilities are not all mutually exclusive. It may be, for example, that the temporal sequence of processing and the absolute magnitudes of the stimuli both shape people's choice of comparative language, or that people have a tendency both to match word order to the spatial layout of the objects and to use "larger" when the difference between the two objects is more pronounced.

General Method

Most studies were run on-line via Amazon's Mechanical Turk using distinct samples of native English speakers from the USA, Canada, and Australia (see Supplementary Materials for additional methodological information). In each study, participants saw 8 or 9 pairs of items and indicated the word that they would naturally use to describe the relationship between them. Each pair differed on a particular dimension. Examples are presented in Figure 1. Each participant saw just one example of each pair, with the order of the pairs randomized unless otherwise noted. In the interests of generality, both the set of dimensions and the pairs of items used for each dimension varied across studies.

Most experiments used a sentence completion task with sentences of the form "One X is _____ than the other". (E.g., "One weight is ____ than the other"). Participants' free responses were classified into three categories: "smaller" responses, where the comparative adjective implied a decrease in magnitude (e.g., "smaller", "fewer", "less", "lower", "lighter", "shorter"), "larger" responses, where the comparative adjective implied an increase in magnitude (e.g., "bigger, "larger", "greater", "more", "heavier", "higher", "longer", "taller"); and cases where the response was irrelevant/unclassifiable and excluded from analysis (4% of all responses). Two raters coded the responses; agreement was >99% with disagreement resolved through discussion.

Experiments 1a and 1b

These experiments used sequential presentation of the items in each pair.

Methods

The stimuli are summarized in Table 2. Each trial began with a "get ready" message for 1 second, followed by the two items, one after the other in the same spatial location for 2 seconds each, and then the sentence completion task (with no time limit) in which participants indicated the word that they would naturally use to complete sentences such as "One line is _____ than the other". Experiments 1a (n=212) and 1b (n=378) differed only in the pairs of items shown to the participants (Table 2).

Results and Discussion

The top row of Figure 2 shows the mean proportion of "larger" responses for each experiment, organized by temporal order (Large then Small or Small then Large). Here we have calculated the response proportions for each participant and averaged them; full analyses with the data for each stimulus dimension examined separately are presented in the Supplementary Materials and yield the same patterns as the aggregated analysis for this and all subsequent experiments.

The first key finding is that, across dimensions, experiments, and temporal arrangements, participants show a strong tendency to favour "larger" responses: All of the bars in Figure 2 are above the 50% line, and one-sample t-tests show this effect to be highly significant in these and all subsequent experiments (all ps << .001). Thus, participants show a strong tendency to use "larger" comparatives, even when a "smaller" comparative would be logically equivalent. This core finding replicates in all of the current studies, and we refer to this higher use of larger comparatives as the HULC effect. The HULC effect argues against the idea that people use smaller and larger comparatives equally often (the ambivalence hypothesis), and applies across dimensions including number, length, area, money, probability, time, and height. (The proportion of "larger" responses was more than 50% for all 36 dimension x temporal order combinations in these two experiments, significantly so in 32 cases; see Supplementary Materials. Similar cross-dimensional generality was also found in the subsequent experiments.)

The second finding is that the HULC effect is modulated by temporal order; "larger" responses were more common when the smaller member of the pair was shown first. This is consistent with the sequential processing hypothesis: The first item to be encoded acts as a reference point, such that the direction of change from this point shapes the language used to describe the comparison. That is, going from small to big entails conceptualization of the relationship as "larger" whilst going from big to small is represented as "smaller". Paired-sample t-tests show this effect to be significant for both Experiments 1a and 1b [$t(208) = 4.57$, $p < .001$, d =

0.48 and $t(369) = 5.46$, $p<.001$, $d = 0.43$ respectively[1]. Here and throughout, the results of the t-tests were mirrored by the results of non-parametric Wilcoxon tests. All Cohen's *d* values are reported unsigned and were computed using pooled variance from the two conditions (Dunlap, Cortina, Vaslow, & Burke, 1996)].

Although this temporal order effect is large, it is not sufficient to overwhelm participants' strong preference for "larger" comparatives, which remains pronounced even when the direction of change implies a decreasing magnitude. That is, sequential processing shapes but does not dictate people's use of comparative language.

Experiments 2a and 2b

Experiments 2a and 2b used simultaneous presentation of the two items.

Methods

The stimuli are summarized in Table 2. As before, participants indicated the word that they would naturally use to complete sentences such as "One line is _____ than the other". Experiment 2a (n=203) was conducted on-line with each stimulus pair on a separate web-page. The two items in each pair were presented side-by-side with their left-right arrangement randomized on each trial. Experiment 2b (n=135) was conducted using a pen-and-paper questionnaire in the University of Essex, UK; four versions of the questionnaire varied the question order and left-right arrangement of the items in the pair.

Results

The middle row of Figure 2 plots the proportion of "larger" responses for each stimulus pair in each experiment. The data are organized according to whether the larger item was on the left (white bars) or on the right (gray bars). As before, there is a robust higher use of larger comparatives. The HULC effect therefore generalizes to simultaneous presentation of the to-be-compared items and to a pen-and-paper version of the sentence completion task.

More importantly, the HULC effect is modulated by spatial arrangement; the proportion of "larger" responses is always greater when the larger member of the pair is on the left [the white bars are above the gray ones; for Experiment 2a, $t(200) = 5.24$, $p<.001$, $d = 0.55$; for Experiment 2b,

[1] The mean response proportions shown in Figure 2 are based on all classifiable responses. When comparing specific pairs of conditions, a handful of participants were excluded because they only had data from one of the relevant conditions due to the trial-by-trial randomization of stimulus presentation.

$t(134) = 5.22$, $p<.001$, $d = 0.61$][2]. Experiments 4a-c below also randomized left-right arrangement and replicated this pattern in every case.

In Experiments 1a and 1b, the order in which the items in a pair were processed influenced the way their magnitude relation was described. One obvious possibility for simultaneously-presented items is that participants process the items left-to-right, consistent with the direction of reading. As noted above, this left-right sequential hypothesis predicts greater use of larger comparatives for a small-large spatial arrangement, the opposite of what was found. Similarly, the data argue against a tendency to process the larger item in the pair first and to use it as the reference point when defining the relation between the two objects (the size-based sequential hypothesis); such a strategy would entail large-to-small processing and a corresponding tendency to respond "smaller", rather than the robust tendency to respond "larger" that was actually observed. The data therefore argue against the left-right and size-based sequential processing hypotheses.

The effect of spatial arrangement is, however, consistent with a tendency to match word order to object order (the location-matching hypothesis), such that small-large spatial layouts encourage the response "One object is smaller than the other" whereas large-small layouts encourage "One object is larger than the other" (see Table 1). The next two experiments further tested this explanation for the effect of spatial layout by using simultaneous object presentation and a choice task rather than a sentence-completion task, with no implicit or explicit instruction to phrase the comparison as "One X is ____ than the other".

Experiments 3a and 3b

Methods

Experiment 3a (n=362) used the same stimuli as Experiment 2a. The two items in each pair were presented simultaneously and participants were instructed to: "Compare these two [squares, lines, etc]. Which word best described the relationship between them?" Below this were two options based on the modal "smaller" and "larger" responses from Experiment 2a (see Supplementary

[2] A reviewer asked whether the spatial and temporal order effects of Experiments 1a-2b held when only the first response from each participant was analysed. For Experiment 1a the proportions of "larger" responses using first-trial data were 78.1% and 81.9% for the large-then-small and small-then-large sequences; for Experiment 1b the proportions were 86.6% and 79.5%. Chi-square tests showed that neither difference is significant [$\chi^2(1) = 0.43$, $p = .511$ and $\chi^2(1) = 3.13$, $p = .077$, respectively; note that single-trial data are noisy and the analysis low in power]. For Experiment 2a, the proportions of "larger" responses were 85.4% and 95.8% for the small-large and large-small arrangements, $\chi^2(1) = 5.92$, $p = .015$, and for Experiment 2b the proportions were 63.0% and 90.2%, $\chi^2(1) = 13.20$, $p < .001$, mirroring the effects in the main analysis.

Material), one above the other (with position randomized). The left-right arrangement of the objects in each pair was randomized. Experiment 3b (n=379) was similar, except different stimulus pairs and response options were used (see Table 3).

Results and Discussion

The results are shown in the bottom row of Figure 2. Participants were more likely to use "larger" adjectives for both spatial arrangements in both experiments. Thus, the HULC effect generalized to a choice task. However, unlike the previous experiments, there was no effect of spatial order [for Experiment 3a, $t(358) = 0.90$, $p = .371$, $d = 0.08$; for Experiment 3b, $t(373) = 1.34$, $p = .180$, $d = 0.12$; note that the non-significant effect is in the opposite direction in the two studies], bolstering the idea that the influence of spatial order in the sentence completion studies reflects a tendency to match word order to object order.

Experiments 4a-4c

These experiments manipulated the magnitudes of the to-be-judged items to test the similarity hypothesis (that "larger" responses are more likely when the difference between the items is more pronounced) and the magnitude hypothesis (that "larger" responses are more likely when the items are of higher magnitude).

Method

Experiments 4a (n=300), 4b (n=376), and 4c (n=443) differed only in the stimuli used (see Table 3). For each dimension, four items of increasing magnitude were constructed and labelled S1, S2, L1, and L2. S1 and S2 were both small (e.g., for money, \$10 and \$14) and L1 and L2 were both large (e.g., \$1000 and \$1400). These four items were used to create three magnitude conditions: a "small pair", where both items were low magnitude (S1 and S2); a "large pair", where both items were high magnitude (L1 and L2); and a "big jump", where there was a big difference in the magnitude of the two items (S1 and L2).

Each participant saw one pair for each dimension, randomly selected from these three types. The items in the pair were presented simultaneously, with left-right arrangement randomized. Participants indicated how they would describe the relationship between the two items via a sentence completion task.

Results and Discussion

The proportion of “larger” responses for each magnitude condition (small pair, large pair, and big jump) is shown in Figure 3 and is greater than 50% in all cases, replicating the HULC effect. More important is the comparison between the three magnitude conditions. The similarity hypothesis predicts that “larger” responses are more likely when the difference between the items is greater. That is, we would expect more “larger” responses in the big-jump condition (where the items are very different) than in the small-pair and large-pair conditions (where the stimuli are more similar). The data do not support this prediction: The difference between the small-pair and big-jump conditions is not significant for any experiment [Expt 4a: $t(268) = 0.52$, $p = 0.601$, $d = 0.04$; Expt 4b: $t(344) = 0.80$, $p = 0.425$, $d = 0.05$; Expt 4c: $t(408) = 1.22$, $p = .222$, $d = 0.08$]. Likewise, there is no difference between the large-pair and big-jump conditions for Experiments 4b and 4c [$t(353) = .18$, $p = .859$, $d = 0.01$ and $t(399) = 1.70$, $p = .096$, $d = 0.10$, respectively], and for Experiment 4a there is a significant effect in the wrong direction, with the large-pair condition producing more “larger” responses than the big-jump stimuli, $t(270) = 2.31$, $p = .022$, $d = 0.18$.

There is, however, some support for the idea that people are more likely to respond “larger” when the absolute magnitude of the two items being compared is greater (the magnitude hypothesis). In all three experiments, the HULC effect is more pronounced for the large-pair than the small-pair, and this difference is significant for Experiments 4a and 4c but not 4b [$t(277) = 3.01$, $p = .003$, $d = 0.24$; $t(393) = 2.86$, $p = .004$, $d = 0.18$; $t(348) = 1.19$, $p = .234$, $d = 0.08$ respectively). Analysing each stimulus dimension separately yields the same conclusions regarding both the similarity and magnitude hypotheses (see Supplementary Materials).

General Discussion

Across multiple dimensions, objects, and tasks, English speakers preferentially used "larger" comparatives to describe the relationship between two magnitudes. This HULC effect was modulated by temporal order, left-right spatial arrangement, and stimulus magnitude.

The HULC effect

As noted above, linguists and psycholinguists classify many "larger" comparatives as *unmarked* and many "smaller" comparatives as *marked*. That is,"the senses of certain "positive" adjectives, like *good* and *long*, are stored in memory in a less complex form than the senses of their opposites" (Clark, 1969, p. 389), with the presumed consequence that comparisons such as "A is shorter than B" are harder to process than "B is longer than A". Whereas previous support for this idea has come from analysis of the acceptability of certain sentence structures and from studies of the time taken to solve reasoning problems or to rapidly decide which of two objects is larger, the

robust HULC effect described here demonstrates a widespread preference for "larger" comparatives when people actually describe the relations between objects that are in front of them. In addition, the current studies show that the language of comparison depends on the spatio-temporal arrangement of the objects. Thus, our results are consistent with greater accessibility of "larger" (unmarked) comparatives, but also show that the use of these comparatives is modulated by other aspects of the judgment situation.

Why does the HULC effect arise? One possibility is that it reflects a general preoccupation with "bigness" or “positivity”, perhaps arising from the evolutionary or ontogenetic importance of identifying the largest item in a set. For example, Silvera, Josephs, and Giesler (2002) found that adults and children tend to favour the larger to two abstract stimuli when making aesthetic judgments. If the larger of two items is typically the more important or valuable then it is likely to be referred to more frequently, and this may well include reference to its relative magnitude. (For example, children rarely cry out for a small(er) ice cream.) The HULC effect may arise because such tendencies have generalized to dimensions and object pairs for which there is no clear reason for preferring the larger item, like the pairs of squares, lines, or numbers of dots used in some of the current experiments. More generally, the cross-dimensional robustness of the HULC effect may lend weight to the idea that a common system represents many different magnitudes (e.g., Walsh, 2003; see Matthews, Stewart, & Wearden, 2011 for an alternative perspective).

Establishing the generality of the HULC effect will provide one important test of these speculations, and future work should ask whether the higher use of larger comparatives extend to languages other than English, whether it is modulated by writing direction, and whether it extends to cultures which place greater value on smaller items.

Spatial and temporal layout

The HULC effect is modulated by the spatial and temporal arrangement of the to-be-compared items. With sequential presentation, the change in magnitude shapes the choice of comparative such that a small-then-large sequence is more likely to be labelled "larger" than a large-then-small ordering. This support for the sequential processing hypothesis fits with the general principle, noted in the Introduction, that people frequently use the first item in a pair as the reference point when judging the second (e.g., Stewart et al., 2005). Importantly, however, this tendency modulates but does not over-ride the HULC effect: Even with a large-then-small sequence, participants were more likely to label the relation "larger" than "smaller".

With simultaneous object presentation we found no indication that language use is shaped by a tendency to use the left-most or largest item as a reference point. (Given the effects of sequential presentation noted above, a left-to-right processing sequence would predict a greater

tendency to respond "larger" when the large item is on the right, and a large-to-small processing sequence would predict an overall tendency to say "smaller" irrespective of left-right layout; see Table 1.) Rather, the effect of spatial arrangement seems to reflect a desire to match word order and object locations such that, when people construct sentences of the form "One item is [comparative adjective] than the other", they use the leftmost object as the "one item" and the rightmost as "the other". This explanation of the spatial layout effects is supported by the fact that the left-right arrangement no longer affected responses when a choice task was used (Experiments 3a and 3b). It also predicts that simultaneous presentation coupled with a forced-choice task in which people choose a word to complete a sentence would yield a return of the location-matching effect seen in Experiments 2a and 2b.

Why the difference between sequential and simultaneous presentation? The tendency to use the first-encountered item as a reference point may well still be present with simultaneous presentation, but in this situation the order of processing is much less constrained than with strict sequential presentation: People can fixate the left- or right-hand member of the pair randomly from trial to trial, and can rapidly flip between them in the early stages of viewing. It would therefore be instructive to record eye-movements to see whether the choice of comparative relates to the sequence of object inspection on a trial-by-trial basis. Similarly, the tendency to use the layout of the objects as a form of "template" for the construction of a comparative sentence may not be completely absent with sequential presentation. That is, people may wish to match the order of the items in time to their spatial order in the sentence, but this is likely to be weaker than the strictly spatial location-matching afforded by a sentence completion task with simultaneous presentation. It would therefore be interesting to see whether location-matching effects emerge when the items are presented sequentially but at different left/right locations, or when the participant produces a temporally-unfolding output (e.g., verbally describing what they have seen) where there might be a tendency to match the temporal order of the objects to the temporal order of the output. In short, we conjecture that the difference between simultaneous and sequential presentation is likely a question of the balance between competing tendencies, with order effects more prominent when the items occur in strict sequence and location (or temporal) matching effects more prominent when the stimulus structure readily maps onto the output structure.

The effects of magnitude and similarity

Experiments 4a-4c found no support for the idea that "larger" comparatives would be more common when the difference between the two items was greater (or, equivalently, that "smaller" comparatives would be more common when the two items were more similar; Holyoak & Hummel, 2001). The lack of a similarity effect in our experiments accords with results from Choplin and

Hummel (2002), who found no evidence that using a "larger" comparative leads people to regard the difference between two items as greater in magnitude.

We did, however, find some indication that the HULC effect depends on the absolute magnitudes of the stimuli being compared: Participants were more likely to use "larger" comparatives when both members of the pair were high-magnitude. This fits with other research showing that people associate "larger" comparatives with higher-magnitude items. For example, Choplin and Hummel (2002) found that describing the relation between two VCR warranties with the comparative "longer" led participants to construe *both* warranties as longer (when compared to the average VCR warranty) than when the relation between them was described as "shorter". Similarly, as noted in the Introduction, there is a semantic congruity effect such that people are quicker to identify the smaller of two small items and the larger of two large items (e.g., Holyoak & Mah, 1981).

Taken together, these results suggest that "smaller" (marked) comparatives are more likely to be used when the items are low in magnitude. However, two points are worth noting. Firstly, whereas the estimation and semantic-congruity studies mentioned above examined how experimenter-provided comparatives affect people's judgments, the current work examined the effects of magnitude on language production. Secondly, and perhaps relatedly, the effect in the current experiments was small: Across Experiments 4a-4c, increasing the magnitude of the to-be compared items only increased the proportion of "larger" responses by 6.6%, despite the fact that in many cases the Small Pair and Large Pair items differed greatly in magnitude. And, even when the two items were very small, people made higher use of "larger" comparatives. Thus, the effect of magnitude is perhaps rather less than assumed/implied by discussions of markedness (e.g., Clark, 1969).

One explanation for the weak magnitude effect is that, when a single pair is presented in isolation, people have little sense of how “large” or “small” they are (although set against this is the idea that people can use their extensive experience of stimuli in the “real world” to gain an immediate sense of where a given item comes on a particular dimension such – see e.g., Brown & Matthews, 2011; Stewart et al., 2006). In a preliminary investigation, we examined whether adding a reference point that highlights the magnitudes of the to-be-compared items would increase the magnitude effect. Participants compared pairs of items (e.g., two circles) which were both small or both large (like the small-pair and large-pair conditions of Experiments 4a-4c). In between the to-be-compared items was a medium-sized object (e.g., a medium-sized square) which was bigger than both members of the small pair and smaller than both members of the large pair. The middle object therefore provided a reference point to emphasize the “smallness” of both members of the small pair and the “largeness” of both members of the large pair. Interestingly, provision of such a

reference point did not increase the magnitude effect – across 5 dimensions, the mean proportion of "larger" responses was 76% for the small pair and 81% for the large pair – a difference comparable to that in Experiments 4a-4c.

Implications and future questions

The current work suggests several questions for future enquiry. First, does the effect generalize across languages and cultures? Similarly, the generalization to multi-object arrays and to other forms of magnitude comparison (for example, value judgments such as "better" and "worse", and comparisons for which the "larger" comparative implies a *lower* valuation, such as "fatter" vs "thinner") will be important, as will examination of more sophisticated linguistic constructions such as those that include negation.

Second, are there individual differences in the HULC effect? For example, is there a sub-population of individuals who preferentially use "smaller" comparatives? As a first exploration of this issue, we calculated the overall proportion of "larger" responses from each participant. The distribution of these proportions, collapsed across all experiments, is plotted in Figure 4. There is no indication of a bimodal distribution; only 6 (out of 2783) participants exclusively used "smaller" comparatives, whereas 681 exclusively used "larger" comparatives. Nonetheless, the heterogeneity visible in Figure 4 suggests that there may be important individual differences in the HULC effect, perhaps arising from differences in personality or processing style. We had participants make just one response to each of a handful of object pairs. Future work grounded in the psycholinguistic tradition might elicit more responses from each participant and apply multi-level analysis to give a more sensitive assessment of key manipulations and individual differences.

Finally, how does the HULC effect relate to choices and behaviour? As described in the Introduction, the way that a magnitude relation is described can influence estimation, preference, and reasoning. One implication of the current work is that explicitly labelling a comparison as "larger" should have less of an effect than labelling it as "smaller", because the latter contradicts the "default" representation whereas the former does not. Another implication is that manipulating the spatial and temporal order of presentation will affect people's construal of a given relationship and, correspondingly, their judgments and choices. For example, whether people say "Option A has a lower probability than Option B" or "Option B has a higher probability than Option A" will depend on the spatio-temporal arrangement of the options, potentially allowing option layout to shape preference and choice via the mediating influence of language.

References

Allan, K. (1986). Interpreting English comparatives. *Journal of Semantics*, *5*, 1-50.

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition*, *75*, 1-28.

Brown, G.D.A., & Matthews, W.J. *Frontiers in Psychology*, *2*, article 299, 1-4, doi: 10.3389/fpsyg.2011.00299

Chapman, G.B., & Johnson, E.J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes*, *79*, 115-153.

Choplin, J.M. (2010). I am "fatter" than she is: Language-expressible body-size comparisons bias judgments of body size. *Journal of Language and Social Psychology*, *29*, 55-74.

Choplin, J.M., & Hummel, J.E. (2002). Magnitude comparisons distort mental representations of magnitude. *Journal of Experimental Psychology: General*, *131*, 270-286.

Clark, H.H. (1969). Linguistic processes in deductive reasoning. *Psychological Review, 76*, 387-404.

Cruse, D.A. (1976). Three classes of antonym in English. *Lingua*, *38*, 281-292.

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*, 166-178.

De Soto, C.B., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology, 2*, 513-521.

Dickinson, C.A., & Intraub, H. (2009). Spatial asymmetries in viewing and remembering scenes: Consequences of an attentional bias? *Attention, Perception, & Psychophysics*, *71*, 1251-1262.

Dunlap, W.P., Cortina, J.M., Vaslow, J.B., & Burke, M.J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170-177.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*, 188-200.

Goodwin, G.P., & Johnson-Laird, P.N. (2005). Reasoning about relations. *Psychological Review*, *112*, 468-493.

Holyoak, K.J., Dumais, S.T., & Moyer, R.S. (1979). Semantic association effects in a mental comparison task. *Memory & Cognition*, *7*, 303-313.

Holyoak, K.J., & Mah, W.A. (1981). Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. *Memory & Cognition*, *9*, 197-204.

Hummel, J.E., & Holyoak, K.J. (2001). A process model of human transitive inference. In M.Gattis (Ed.), *Spatial schemas in abstract thought* (pp. 279-305). Cambridge MA: MIT Press.

Hunter, I.M.L. (1957). The solving of three-term series problems. *British Journal of Psychology*, *48*, 286-298.

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

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211-228.

Laming, D. (1995). Screening cervical smears. *British Journal of Psychology*, *86*, 507-516.

Laming, D. (2003). *Human judgment: The eye of the beholder*. London: Thomson.

Matthews, W.J. (2011). What might judgment and decision making research be like if we took a Bayesian approach to hypothesis testing? *Judgment and Decision Making*, *6*, 843-856.

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

Matthews, W. J., & Stewart, N. (2009a). 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*, 64-81.

Matthews, W. J., & Stewart, N. (2009b). The effect of inter-stimulus interval on sequential effects in absolute identification. *Quarterly Journal of Experimental Psychology, 62*, 2014-2029.

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,* 303-313.

Mussweiler, T. & Epstude, K. (2009). Relatively fast! Efficiency advantages of comparative thinking. *Journal of Experimental Psychology: General*, *138*, 1-21.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*, 136-164.

Read, D., Frederick, S., Orsel, B., & Rahman, J. (2005). Four score and seven years from now: The date/delay effect in temporal discounting. *Management Science, 51,* 1326-1335.

Schneider, K.A., & Komlos, M. (2008). Attention biases decisions but does not alter appearance. *Journal of Vision, 8*, 1-10.

Silvera, D.H., Josephs, R.A., & Giesler, R.B. (2002). Bigger is better: The influence of physical size on aesthetic preference judgments. *Journal of Behavioral Decision Making*, *15*, 189-202.

Sternberg, R.J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General*, *109*, 119-159.

Stewart, N., Brown, G.D.A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*, 881-911.

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

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space, and quantity. *Trends in Cognitive Sciences, 7*, 483-488.

Winawer, J., Witthoft, N., Frank, M.C., Wu, L., Wade, A.R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Science*, *104*, 7780-7785.

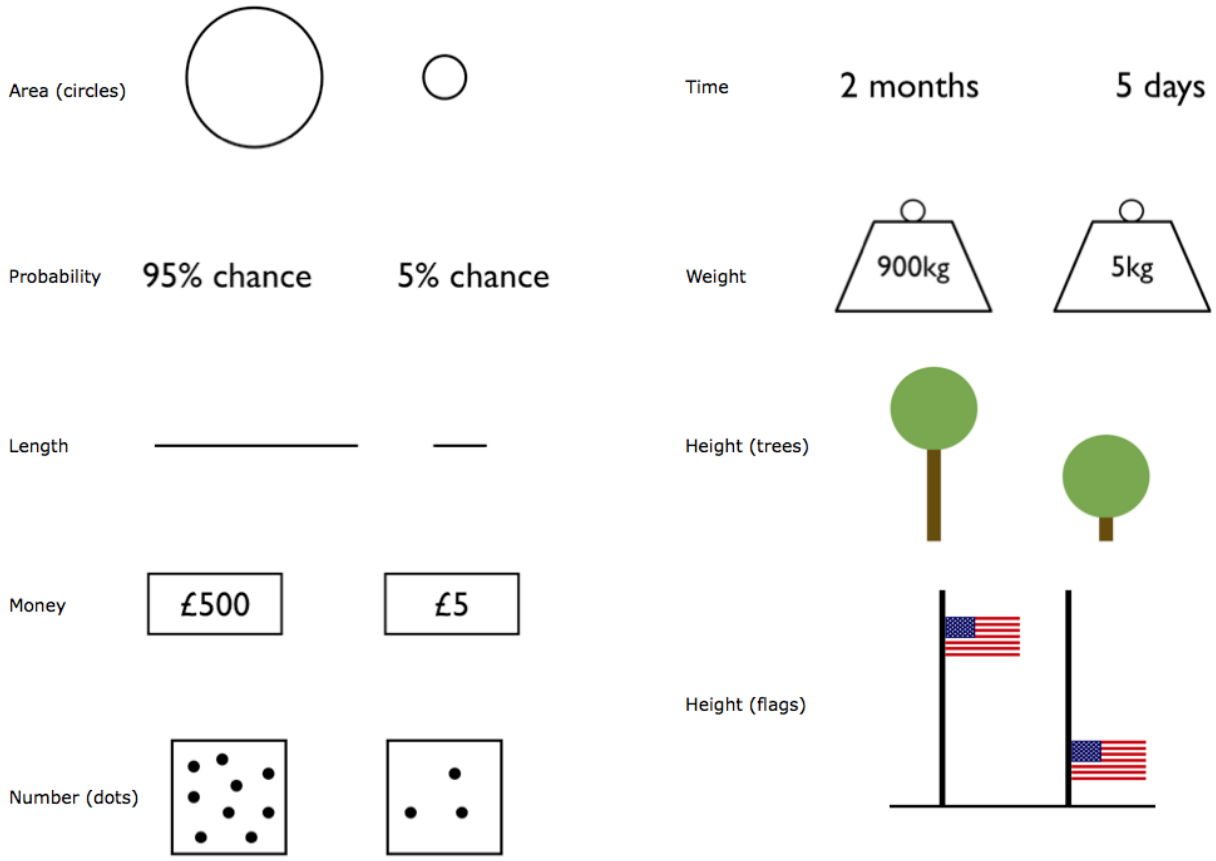

Figure 1. Example stimulus pairs (not shown actual size). The specific magnitudes and details of the items varied across experiments; these are from Experiment 2b, except for the flags which are from Experiment 3b (see Tables 2 and 3 for details). The stimulus pairs shown here have a large-small spatial arrangement; the small-large layout was identical but with the larger object on the right. Each participant saw one object pair for each dimension and indicated the comparative adjective that they would naturally use to describe the relation between the items.

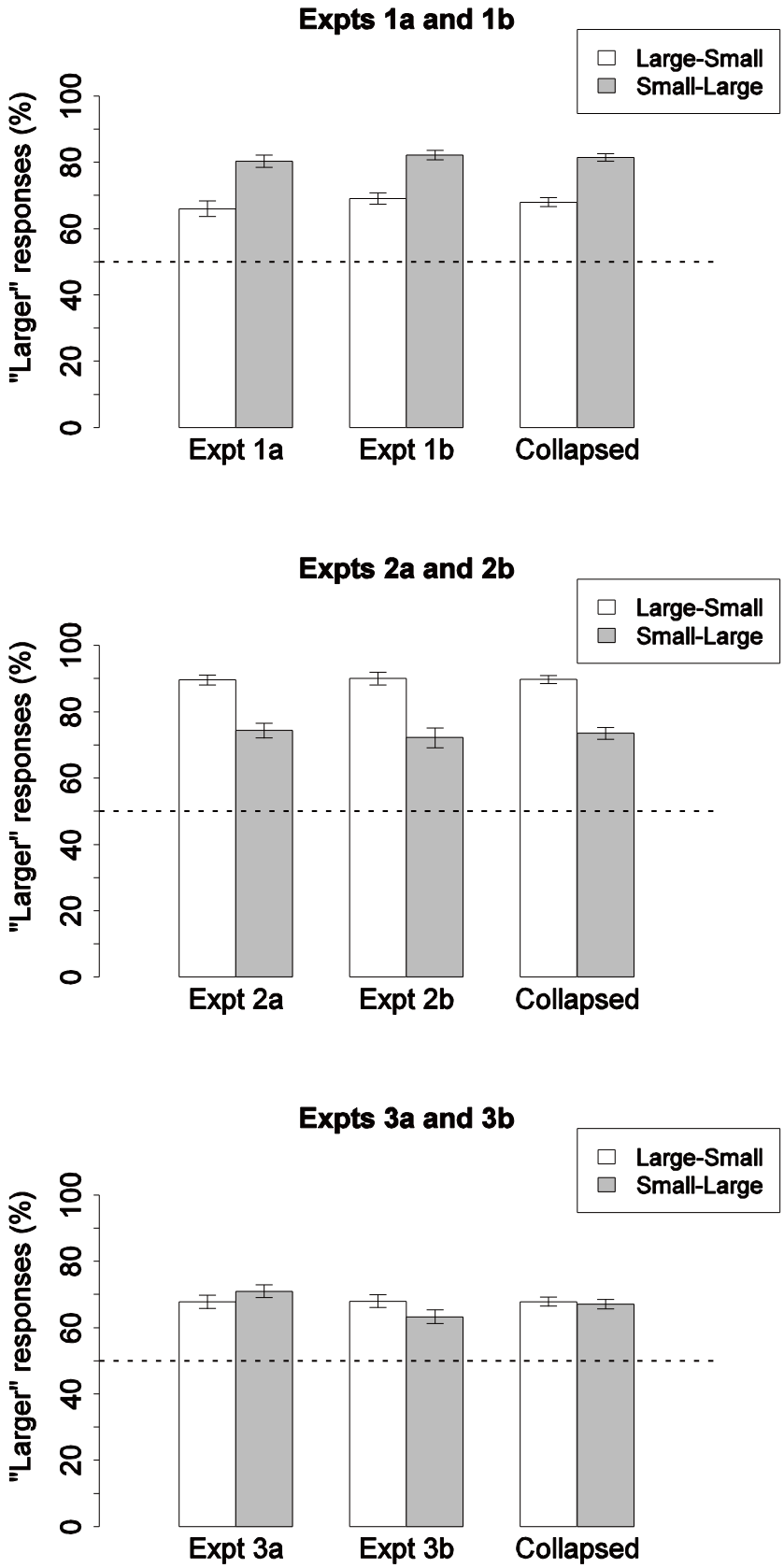

Figure 2. Results from Experiments 1a-3b. The top row shows the mean response proportions from Experiments1a and 1b and collapsed across both experiments. The items in each pair were presented sequentially, either large-then-small (white bars) or small-then-large (grey bars), and participants undertook a sentence-completion task. The middle row shows the results from

Experiments 2a and 2b, which used a sentence-completion task and simultaneous presentation of the objects in each pair, either with the large item on the left (white bars) or the right (grey bars). The bottom row shows the results of Experiments 3a and 3b, which were similar to Experiments 2a and 2b but which used a choice task rather than sentence completion; there is no longer any effect of spatial arrangement. The error bars show plus/minus one standard error (SE), calculated separately for each data point. (Note that these SEs cannot be used to make inferences about the significance of differences between pairs of means.)

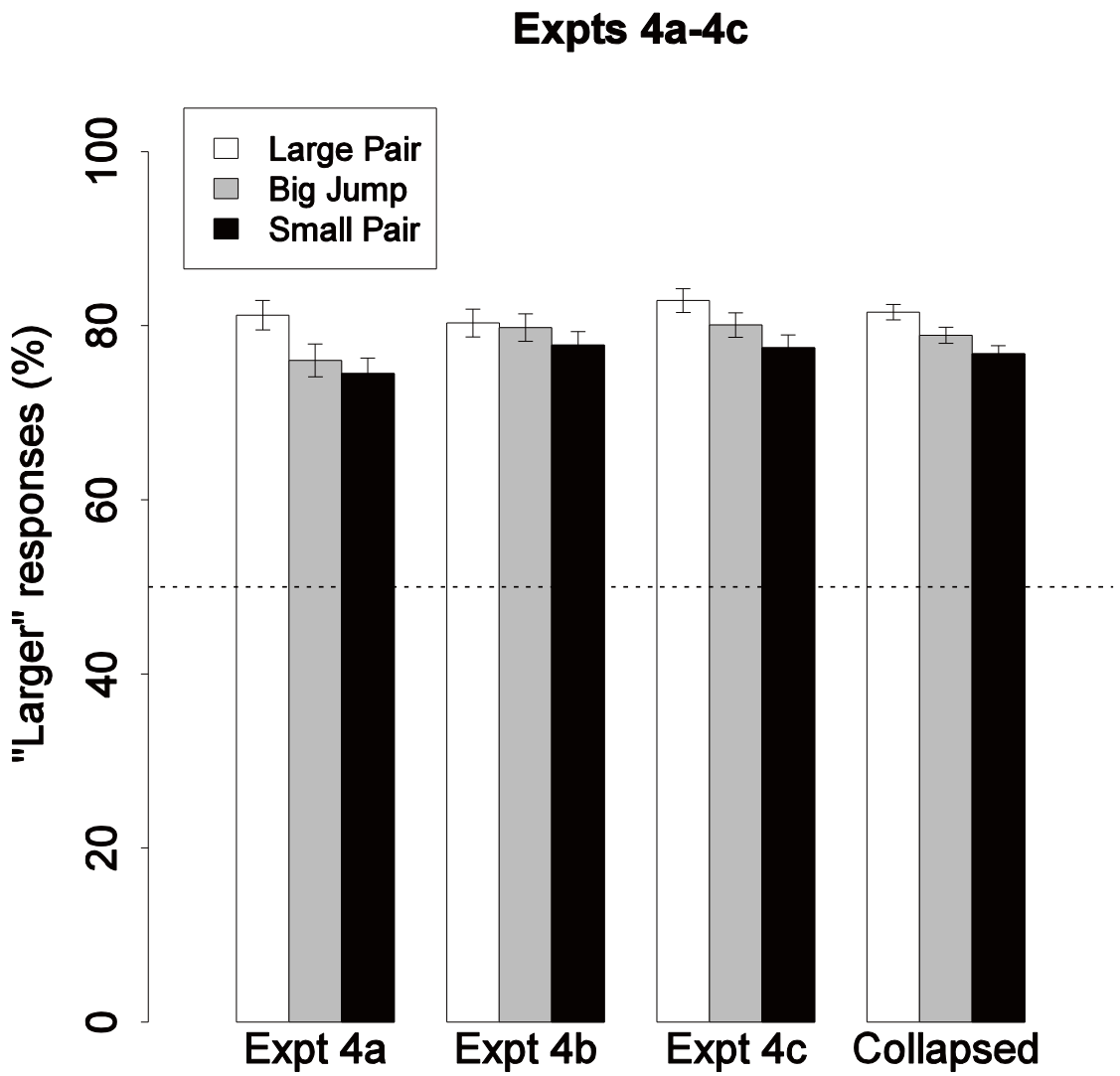

Figure 3. Results from Experiments 4a-4c. Error bars are as for Figure 2.

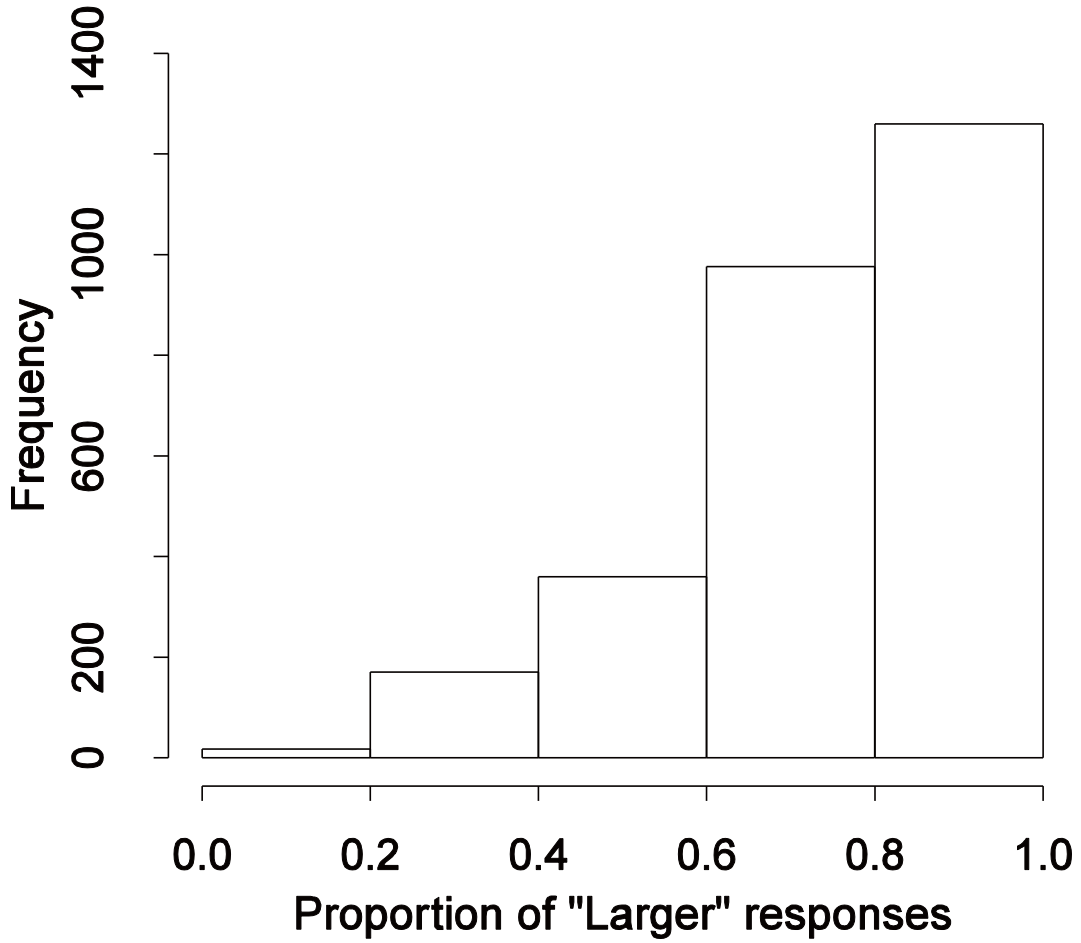

Figure 4. Distribution of response proportions across all experiments. The bars show the number of participants making different proportions of "larger" responses. A large number of people exclusively used "larger" comparatives; only a handful exclusively used "smaller" comparatives.

Table 1. Summary of theoretical ideas explored in the current studies.

Hypothesis	Core idea	Response prediction
Ambivalence	“Smaller” and “larger” comparatives are logically equivalent	Equal use of “smaller” and “larger” comparatives
Sequential processing	Direction of magnitude change dictates choice of comparative	"Larger" responses when small item processed first "Smaller" responses when large item processed first
Left-right sequential	Processing tends to be left-to-right	“Larger” responses more likely when large item is on right “Smaller” responses more likely when large item is on left
Size-based sequential	Processing tends to be large-to-small	Overall tendency to use “smaller” comparatives
Location matching	Word locations matched to object locations, such that constructions of the form "One item is [comparative] than the other" encourage use of right-hand object as reference point	“Larger” responses more likely when large item is on left “Smaller” responses more likely when large item is on right
Similarity	Choice of comparative depends on item similarity	“Larger” responses more likely when difference in magnitude is more pronounced
Magnitude	Choice of comparative depends on absolute magnitude	“Larger” responses more likely when both items are large

Table 2: Stimuli from Experiments 1a, 1b, 2a, and 2b

Expt	Dimension	Small	Large
1a	Area (squares)	57 px	171 px
	Height (flags)	84 px	298 px
	Height (trees)	213 px	396 px
	Length	124 px	518 px
	Money (rewards)	$40 Reward	$100 Reward
	Number (dots)	7	35
	Probability	1% chance	95% chance
	Time	3 minutes	12 minutes
	Weight	1 lb	20 lbs
1b	Area (circles)	18 px	198 px
	Height (flags)	130 px	242 px
	Height (trees)	175 px	295 px
	Length	24 px	224 px
	Money	$3	$500
	Number (squares)	5	17
	Probability	7% chance	70% chance
	Time	3 days	3 years
	Weight	1 kg	100 kg
2a	Area (squares)	79 px	304 px
	Height (trees)	178 px	378 px
	Length	44 px	353 px
	Money (prizes)	Win $10	Win $1000
	Number (dots)	6	19
	Probability	A 2% chance	A 98% chance
	Time (delays)	A 2 week delay	A 6 month delay
	Weight	1 kg	10 kg
2b	Area (circles)	1.5 cm	4.5 cm
	Height (trees)	4.0 cm	6.6 cm
	Length	1.8 cm	6.9 cm
	Money	£5	£500
	Number (dots)	3	9
	Probability	5% chance	95% chance
	Time	5 days	2 months
	Weight	5 kg	900 kg

Note: px = pixels. The nature of the objects presented varied somewhat across experiments. For example, the two shapes that differed in area were sometimes squares and sometimes circles (the

dimensions in the table indicate diameters). Similarly, monetary amounts were sometimes presented in the abstract and sometimes as a prize or reward; times were sometimes stated as abstract durations or as delays. Length was manipulated by presenting two horizontal lines. Number refers to a box circumscribing small circles ("dots"), squares, or stars. Weights were depicted as circles or cartoon weights with written labels. Height (trees) can be thought of as "tallness" and refers to two cartoon trees differing in height; the height in the table gives the distance from the base to the highest point. Height (flags) is closer in meaning to "altitude" and refers to two flags of equal size differing in their distance up two identical flagpoles. The measurement gives the distance from the base of the pole to the bottom of the flag.

Table 3. Stimuli from Experiments 4a-4c.

Expt	Dimension	S1	S2	L1	L2
4a & 3b	Area (squares)	19 px	27 px	119 px	141 px
	Height	42 px	52 px	202 px	252 px
	Length	14 px	20 px	122 px	182 px
	Money	$10	$14	$1000	$1400
	Number (stars)	3	4	15	20
	Probability	A 4% chance	A 5% chance	A 72% chance	A 90% chance
	Time	5 seconds	6 seconds	5 years	6 years
	Weight	4 grams	7 grams	4 tons	7 tons
4b	Area (circles)	20 px	23 px	179 px	207 px
	Height (trees)	80 px	96 px	200 px	240 px
	Length	20 px	26 px	146 px	194 px
	Money	$5	$6	$5000	$6000
	Number (stars)	2	4	10	20
	Probability	6% chance	7% chance	84% chance	98% chance
	Time	2 days	3 days	200 days	300 days
	Weight	4 grams	5 grams	40 tons	50 tons
4c	Area (circles)	9 px	12 px	179 px	233 px
	Height (trees)	88 px	97 px	318 px	350 px
	Length	11 px	15 px	161 px	225 px
	Money	$4	$5	$8000	$10000
	Number (squares)	2	3	14	21
	Probability	5% chance	7% chance	70% chance	98% chance
	Time	3 seconds	4 seconds	300 years	400 years
	Weight	3 grams	5 grams	300 tons	500 tons

Note: px = pixels. The S1 and S2 stimuli form the "small pair" condition; the L1 and L2 stimuli form the "large pair" condition; the S1 and L2 stimuli together form the "big jump" condition. The stimuli were chosen such that the ratio of S1 to S2 on the focal dimension was approximately the same as the ratio of L1 to L2 (although minor drawing errors meant occasional departures from this). For example, the monetary amounts in Experiment 4a are in the ratio 1:1.4 for both S1:S2 and L1:L2, and the ratio of areas in Experiment 4b is 1:1.3 for both the S1:S2 and L1:L2 pairs. Experiment 3b used the S1 and L2 stimuli of Experiment 4a.

Supplementary Material

The following includes additional information about data collection and analysis.

Data Collection and Coding

The on-line studies were run on Amazon Mechanical Turk using the Crowdflower crowdsourcing service (www.crowdflower.com) with participants recruited from the US, Canada, and Australia. The on-line studies were run using the Qualtrics survey software (www.qualtrics.com[3]).

To be included in the final sample for a given study, participants had to meet the following requirements. (1) Their IP address had not previously appeared in the same or a previous study (either from the series reported here or from related experiments). That is, only the first occurrence of an IP address was used; in the case where the same IP appeared at overlapping times, both sets of responses were discarded. (2) They reported an age of 16 or greater. (3) They answered "yes" to the question: "Is English your first language ("mother tongue")? (4) They answered all questions. (The web program required a response to each question before the participant could progress, so participants without a full response set must have left the task early). Some of the studies included questions asking whether all of the images had displayed properly, and either participants or individual responses were excluded in cases where participants indicated a problem. For Studies 1a and 1b, individual responses were excluded if the display time of the "Get Ready" message or either of the two stimuli was out by more than 0.5 seconds. Experiment 2b was a pen-and-paper task; sixty five out of 200 participants who indicated a first language other than English were excluded. For all

[3] We noted some instability with this widely-used platform. After testing was complete, we ran through the experimental programs and discovered that the appearance of the stimulus was sometimes briefly preceded by the name of the corresponding image file (e.g., "10_1000" for the "Win $10 Win $1000" pair with the small item on the left in Experiment 2a, or "light weight" for the 20 lb weight of Experiment 1a). This error was sporadic, only happening for some runs/trials/web browsers, and did not seem to happen when we first ran the experiments (it may have been a consequence of the image library becoming overloaded, or a change made to the Qualtrics platform). Experiments 1b, 2b, 3b, and 4c were unaffected by this issue because the image files were linked to the Qualtrics software in a different way (and Experiment 2b used a pen-and-paper task). Given that the results of these experiments are identical to those of the potentially-affected experiments, we do not regard the stimulus display problem as having had an important effect on our findings. However, a useful lesson for other researchers using the Qualtrics platform is that image display seems to be much more reliable when the image files are hosted on a local server and linked to via a URL rather than uploaded to the Qualtrics library.

experiments, the n-values reported in the main text are after these exclusions; the data from excluded participants were not analyzed.

For the on-line studies, the sentence-completion tasks presented the two objects and, beneath them, a sentence such as: "One circle is _____ than the other". Below this were instructions: "In the space below, type the word that you would naturally use to fill the gap in this sentence", followed by a text box into which participants could type their response. In the pen-and-paper version, participants simply wrote their response in the blank space in the middle of the sentence. In Experiments 1a and 1b, where the two objects appeared one after the other, the sentence completion task appeared after the second object and was reworded to be past tense (e.g., "One square was _____ than the other").

For the sentence-completion tasks, responses were coded as "smaller", "larger", or "unclassifiable/irrelevant", with the latter type excluded from analysis on a case-by-case basis. General coding principles included: (1) the response had to include a comparative adjective. That is, if the response is X, one can say "X than..."; non-comparative adjectives (such as "big") were excluded. (2) The adjective must be appropriate for, and clearly refer to, the focal dimension (e.g., a response of "darker" for the area stimuli would be excluded). (3) Modifiers (e.g., "very") were ignored when deciding on the category of a response, as were spelling or grammatical errors and extraneous words (e.g., where the participant typed the whole of the to-be-completed sentence rather than just the missing word). (4) Unusual responses were acceptable provided they could reasonably be taken to refer to the dimension of interest and could be classified as "smaller" or "larger". (5) Affective or value judgments (e.g., "better"), contradictory responses, and ambiguous responses were excluded. Ambiguous responses included ones which were incomprehensible and ones where classification as "smaller" or "larger" was problematic (e.g., "more small").

As noted in the main text, only a small proportion of responses were excluded and inter-rater agreement was excellent. A full copy of the responses from all experiments (and their categorization as "smaller", "larger", and "unclassifiable") is available from the authors.

For the two choice experiments (Experiments 3a and 3b), participants selected which of two words best described the relationship between the items in the stimulus pair. In Experiment 3a the options were the modal "smaller" and "larger" responses for each dimension taken from Experiment 2a (which used identical stimuli). In Experiment 3b the options were: Area: smaller, larger; Height (flags): lower, higher; Length: shorter, longer; Money: less, more; Number: fewer, more; Probability: lower, higher; Time: shorter, longer; Weight: lighter, heavier.

Data Analysis

For the sake of brevity, the main text shows the proportion of "larger" responses collapsed over items. That is, we calculated, for each participant, the proportion of classifiable responses that were coded as "larger" for each experimental condition. (As we note in the main text, each participant only saw one instance of each stimulus pair and randomization/exclusion of unclassifiable responses meant that some participants do not provide data for all conditions.)

Here were present a more complete analysis in which the data for each dimension are shown separately. Figures S1-S3 show the results for each dimension (area, length, height, etc) for Experiments 1a-3b. Each bar shows the proportion of participants in a given condition who responded "larger". These proportions were calculated after excluding the small number of unclassifiable/irrelevant responses; thus, the proportion of "smaller" responses is simply one minus the proportion of larger responses. White bars show the results when the larger item of the pair was presented first (Experiments 1a, 1b) or on the left (Experiments 2a-3b); grey bars show the results when the larger item was presented second or on the right.

The results mirror those in the main text: Across dimensions, there is a robust tendency to favour "larger" comparatives (the HULC effect). In Experiments 1a and 1b, the HULC effect is modulated by the temporal order of the stimuli: "larger" responses were more common when the smaller member of the pair was shown first for all 18 comparisons (10 significant). In Experiments 2a and 2b, the choice of comparative adjective is influenced by the spatial arrangement of the items: The proportion of "larger" responses is always greater when the larger member of the pair is on the left (the white bars are above the grey ones) and this difference is significant for 12/16 comparisons. [Experiments 4a-4c also randomized left-right arrangement and replicated this pattern for 24/24 comparisons (22 significant).] Experiments 3a and 3b suggest that the effect of spatial order is specific to the sentence completion task; only two of the 16 spatial order effects were significant, and overall there is little indication of a systematic effect of spatial order on people's responses in these choice tasks, bolstering the idea that the influence of spatial order in the sentence completion studies reflects a tendency to match word order to object order.

Figure S4 shows the proportion of participants who responded "larger" for each dimension in each condition of Experiments 4a-4c. As before, there is a robust HULC effect across multiple dimensions and stimulus values. More important are the comparisons between magnitude conditions. According to the similarity hypothesis (see main text), there will be more "larger" responses in the big-jump condition (where the items are very different) than in the small-pair and large-pair conditions (where the stimuli are more similar). The top two panels of Figure S5 plot the relevant contrasts (arranged in ascending order in each panel, to clarify the overall pattern). The similarity hypothesis predicts positive differences; there is little indication of this. The big jump vs.

small pair comparison (top panel) has 15/24 contrasts in the predicted direction (2 significant) and 9 in the wrong direction (2 significant); the big jump vs. large pair comparison has just 8/24 contrasts in the predicted direction (1 significant) and 16 in the wrong direction (4 significant). Thus, increasing the similarity of the items does not seem to ameliorate the HULC effect.

By contrast, there is some support for the magnitude hypothesis. The bottom panel of Figure S5 shows that in 18/24 cases participants were more likely to say “larger” when both items were large than when both were small; this difference was significant in 10 cases. Taken together, these studies suggest that the magnitudes of the items, rather than the difference between them, moderates people’s use of comparative adjectives. In all three panels, it is the pair with the highest mean magnitude which is more likely to elicit a “larger” response.

Tables S1-S5 give the absolute number of "smaller" and "larger" responses for each experiment, organized by stimulus dimension and condition (e.g., left-right arrangement). Note that randomization of conditions meant that the total number of responses in each condition was not constant. The tables show the chi-square tests used to establish whether “larger” and “smaller” responses are equiprobable and whether the choice of comparative adjective depends on experimental condition. In all cases, there is one degree of freedom and the critical values are: 3.841 (p <.05), 6.635 (p<.01), and 10.828 (p<.001).

Tables S6 to S8 give the modal responses for each dimension in each experiment. The tables show both the modal "smaller" and "larger" responses, and the overall modal responses.

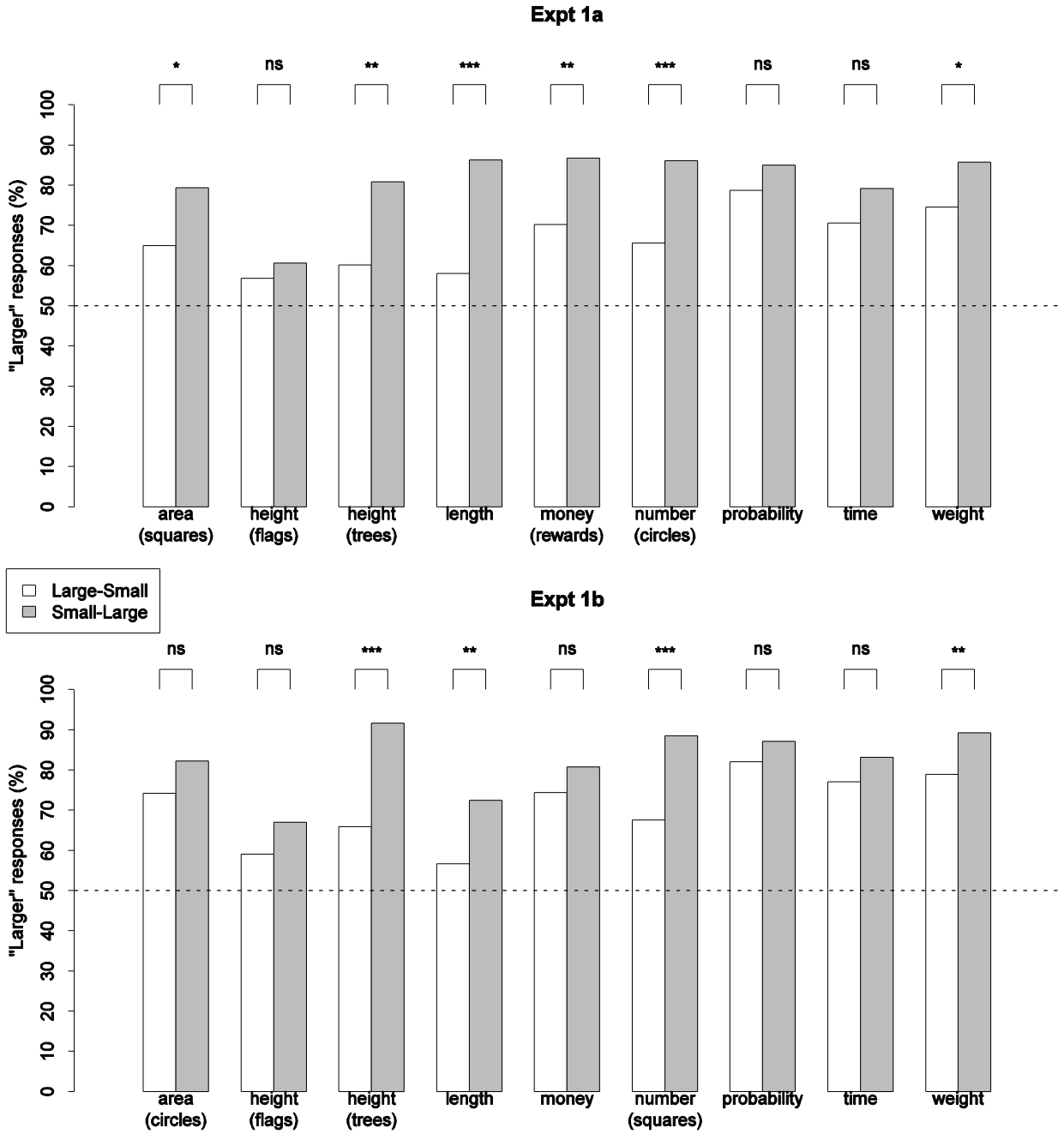

Figure S1. Results from Experiments 1a and 1b. These studies used a sentence-completion task and varied the temporal sequence of the stimuli. The white bars show the proportion of "larger" responses when the larger member of the pair was shown first; gray bars show the proportion of "larger" responses when the larger member of the pair was shown second. Note that all bars are above the 50% line, indicating a robust preference for "larger" responses. Significance markers indicate the results of a chi-squared test for association between response and temporal order: ns = p >.05; * = p <.05; ** = p <.01; *** = p <.001.

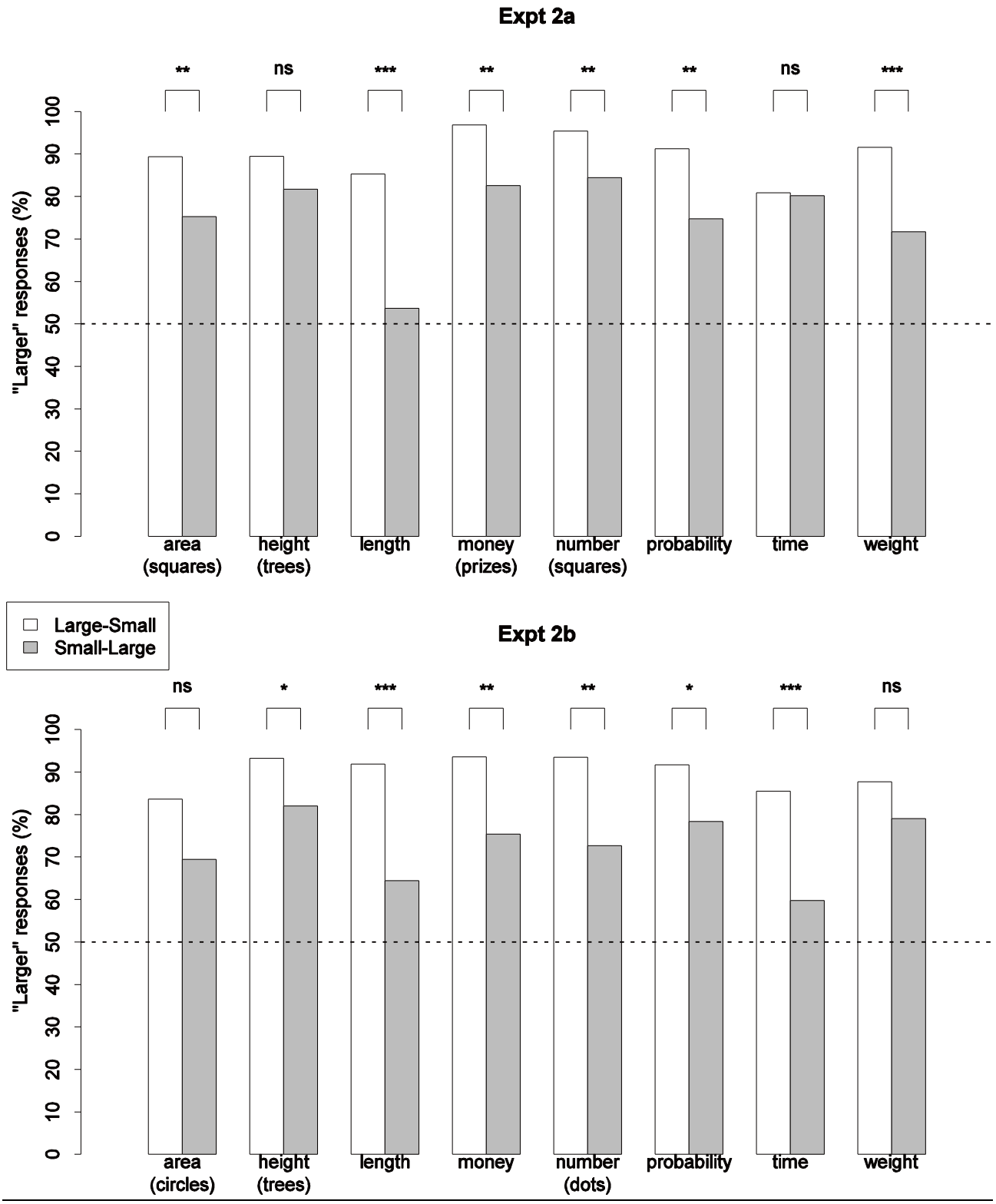

Figure S2. Results from Experiments 2a and 2b. These studies used a sentence completion task and varied the spatial arrangement of the items in each pair. The white bars show the proportion of "larger" responses when the larger member of the pair was on the left; gray bars show the proportion of "larger" responses when the larger member of the pair was on the right. The significance markers indicate the results of a chi-squared test for association between response and spatial arrangement: ns = $p > .05$; * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

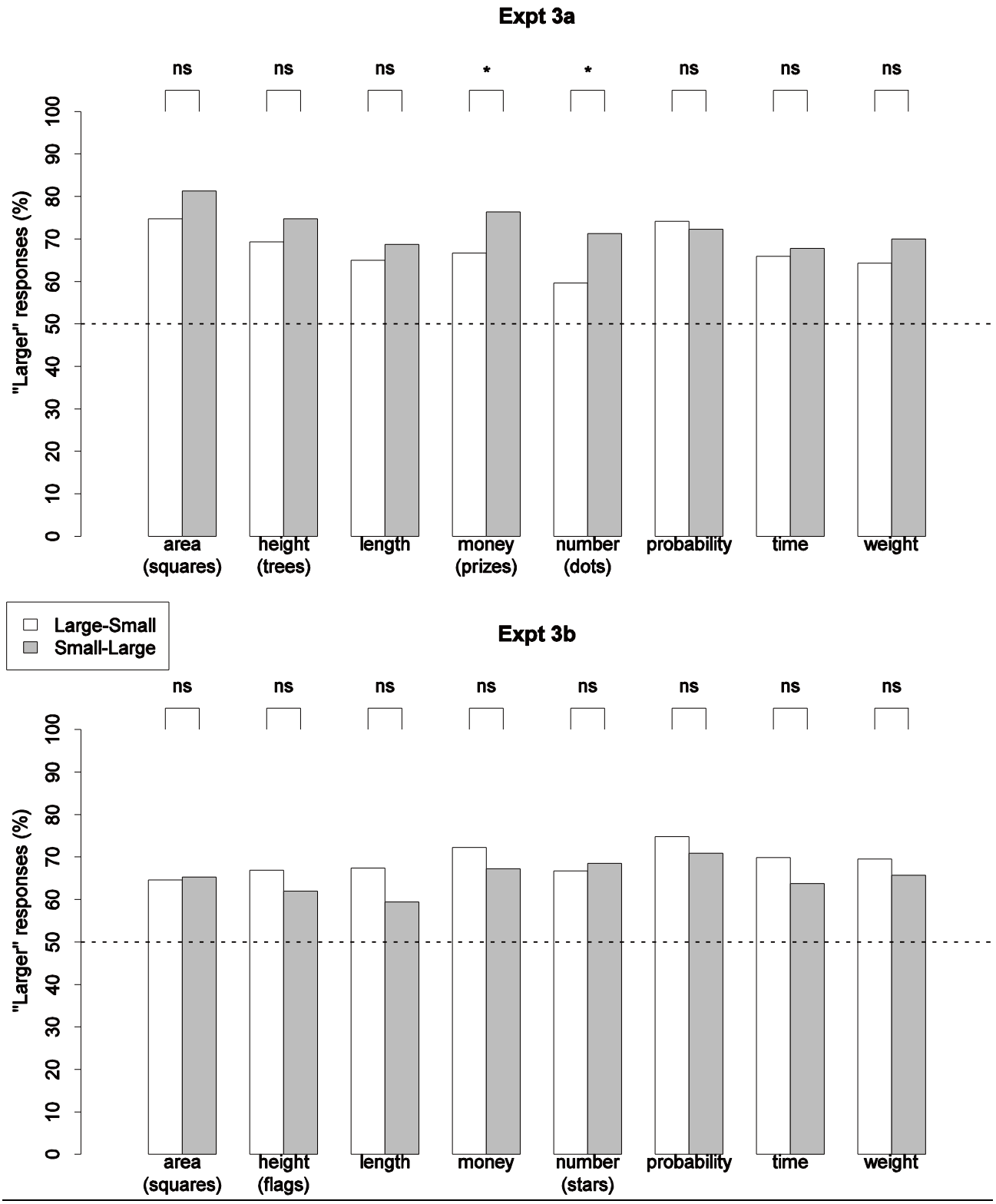

Figure S3. Results from Experiments 3a and 3b. These studies used a two-alternative forced choice task. The white bars show the proportion of "larger" responses when the larger member of the pair was on the left; gray bars show the proportion of "larger" responses when the larger member of the pair was on the right. Significance markers indicate the results of a chi-squared test for association between response and spatial arrangement: ns = p >.05; * = p <.05; ** = p <.01; *** = p <.001.

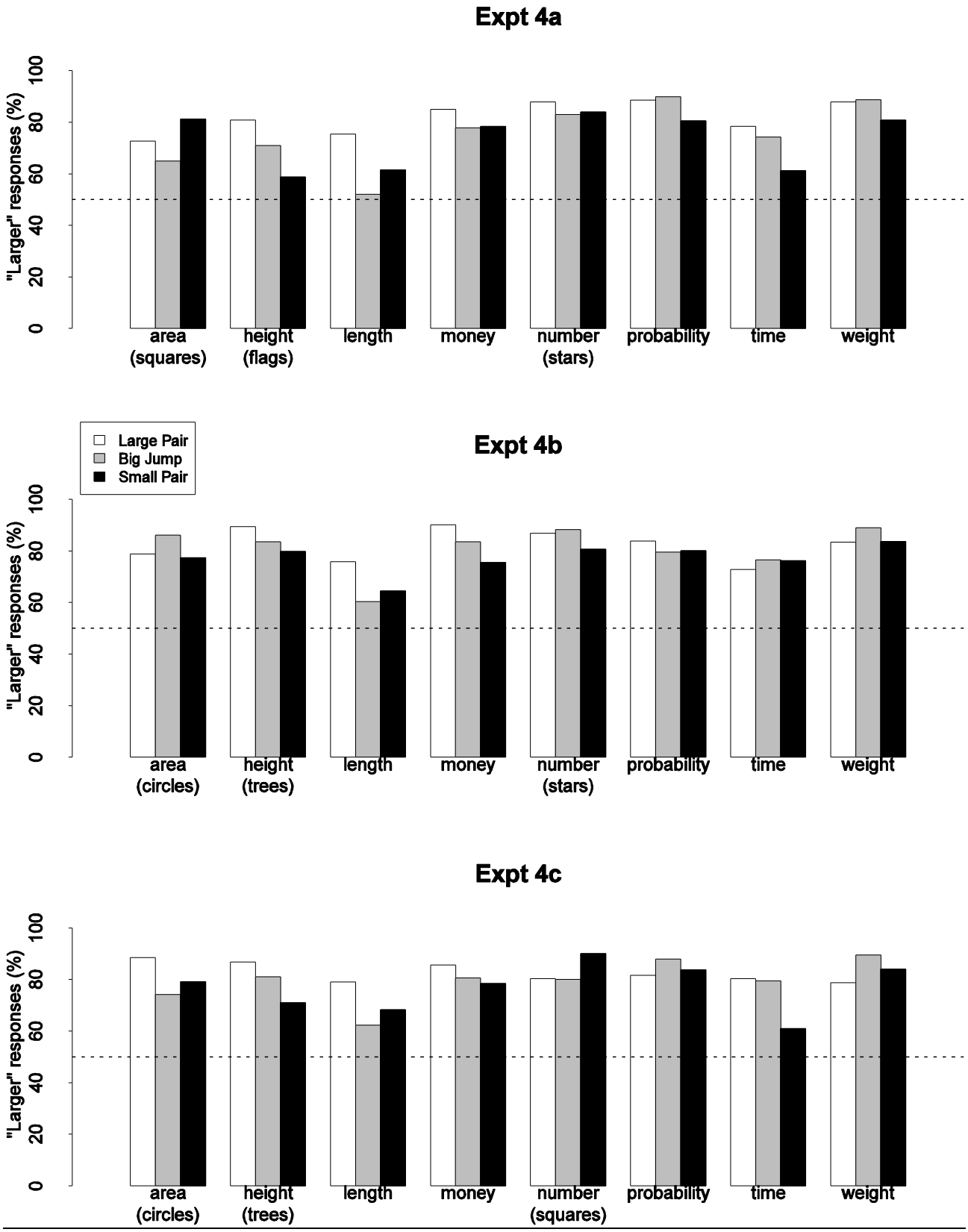

Figure S4. Results from Experiments 4a-4c. White bars show the proportion of "larger" responses when both objects were large (the "large pair"); black bars show the proportion when both objects were small ("small pair"); gray bars show the proportion when there was a big difference between the small and large member of each pair ("big jump").

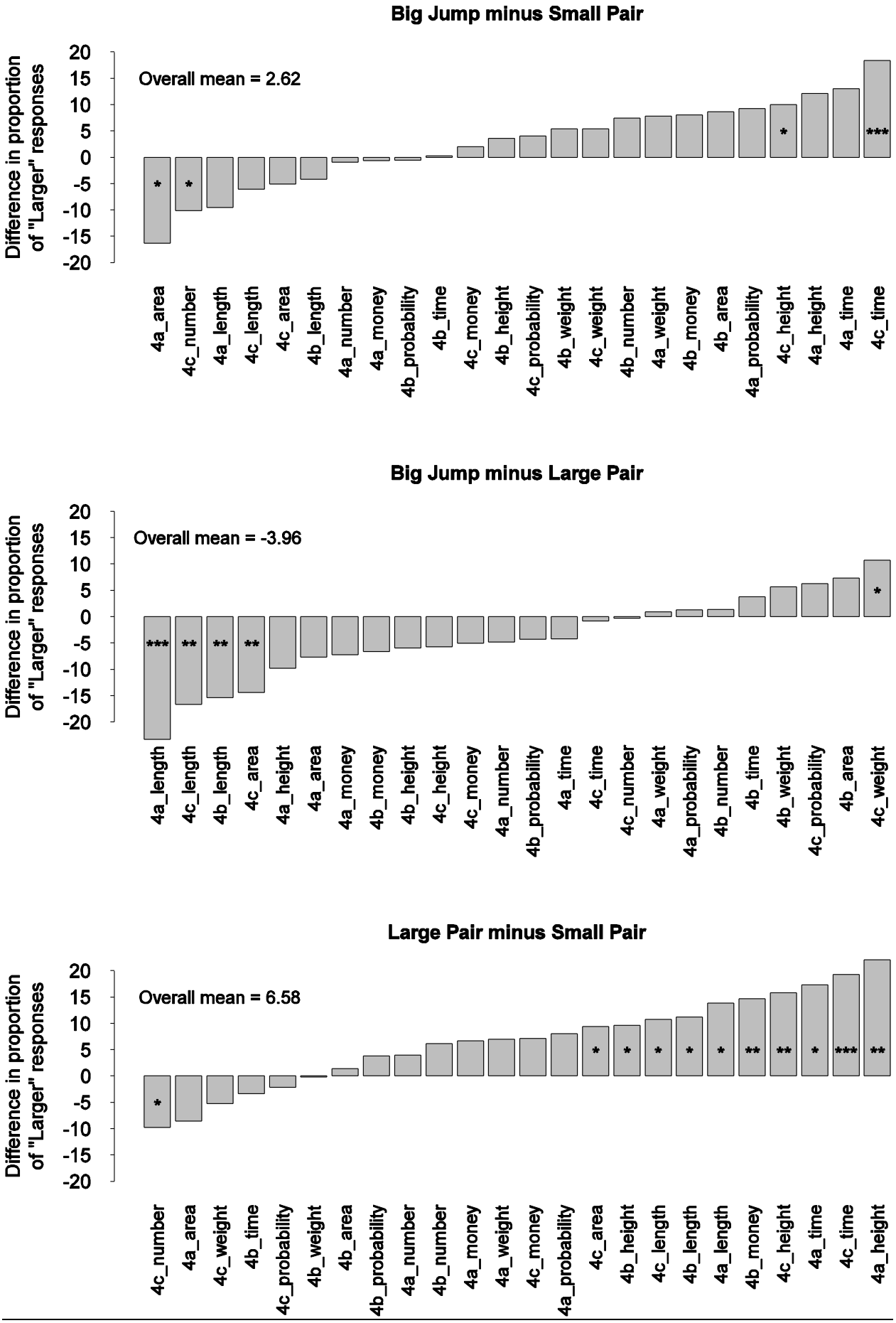

Figure S5. Relevant contrasts for Experiments 4a-4c. The top two panels show the difference in the proportion of "larger" responses between the big-jump condition and the small-pair and large-pair

conditions, respectively. The bottom panel shows the difference between the proportion of “larger” responses for the large-pair and small-pair conditions. In all panels, differences are arranged in order of increasing positivity to clarify the overall pattern. The bar labels indicate the experiment (e.g., 4a) and dimension in question. The significance markers indicate the results of chi-squared tests for association between response type and condition: * = p <.05; ** = p <.01; *** = p <.001.

Table S1. Number of "smaller" and "larger" responses for Experiments 1a and 1b.

		Small-Large			Large-Small			Temporal order
Expt	Dimension	n_S	n_L	χ^2	n_S	n_L	χ^2	χ^2
1a	Area (squares)	21	81	35.29	35	65	9.00	5.23
	Height (flags)	39	60	4.45	41	54	1.78	0.28
	Height (trees)	22	93	43.83	35	53	3.68	10.52
	Length	13	82	50.12	44	61	2.75	19.49
	Money (rewards)	12	79	49.33	33	78	18.24	7.90
	Number (dots)	15	93	56.33	32	61	9.04	11.74
	Probability	15	85	49.00	20	74	31.02	1.29
	Time	22	84	36.26	25	60	14.41	1.91
	Weight	14	84	50.00	28	82	26.51	4.01
1b	Area (circles)	33	153	77.42	47	135	42.55	3.53
	Height (flags)	57	116	20.12	70	101	5.62	2.36
	Height (trees)	15	166	125.97	61	118	18.15	35.94
	Length	49	129	35.96	81	106	3.34	9.91
	Money	35	147	68.92	46	134	43.02	2.08
	Number (squares)	20	154	103.20	60	125	22.84	22.70
	Probability	21	142	89.82	30	137	68.56	1.63
	Time	29	143	75.56	42	141	53.56	2.06
	Weight	20	166	114.60	38	142	60.09	7.36

Note: Small-Large and Large-Small refer to the temporal order of the to-be-compared objects. n_S and n_L are the number of "smaller" and "larger" responses, respectively; the corresponding chi-square values test whether these two types of response occurred equally often. The final column gives the chi-square value for a test of association between response type ("smaller" vs. "larger") and temporal order.

Table S2. Number of "smaller" and "larger" responses for Experiments 2a and 2b.

		Small-Large			Large-Small			Spatial order
Expt	Dimension	n_S	n_L	χ^2	n_S	n_L	χ^2	χ^2
2a	Area (squares)	27	82	27.75	10	84	58.26	6.76
	Height (trees)	19	85	41.88	10	85	59.21	2.39
	Length	50	58	0.59	14	81	47.25	23.32
	Money (prizes)	15	71	36.47	3	92	83.38	10.28
	Number (dots)	14	76	42.71	5	104	89.92	6.87
	Probability	24	71	23.25	8	83	61.81	8.85
	Time (delays)	20	81	36.84	18	76	35.79	0.01
	Weight	30	76	19.96	8	87	65.69	12.92
2b	Area (circles)	19	43	9.29	12	61	32.89	3.83
	Height (trees)	11	50	24.93	5	68	54.37	3.95
	Length	26	47	6.04	5	56	42.64	14.05
	Money	18	55	18.75	4	58	47.03	8.15
	Number (dots)	20	53	14.92	4	57	46.05	9.82
	Probability	13	47	19.27	6	66	50.00	4.72
	Time	29	43	2.72	9	53	31.23	10.88
	Weight	13	49	20.90	9	64	41.44	1.83

Note: Small-Large and Large-Small refer to the left-right arrangement of the to-be-compared objects. n_S and n_L are the number of "smaller" and "larger" responses, respectively; the corresponding chi-square values test whether these two types of response occurred equally often. The final column gives the chi-square value for a test of association between response type ("smaller" vs. "larger") and spatial arrangement.

Table S3. Number of "smaller" and "larger" responses for Experiments 3a and 3b.

		Small-Large			Large-Small			Spatial order
Expt	Dimension	n_S	n_L	χ^2	n_S	n_L	χ^2	χ^2
3a	Area (squares)	33	143	68.75	47	139	45.51	2.23
	Height (trees)	47	139	45.51	54	122	26.27	1.32
	Length	57	125	25.41	63	117	16.20	0.55
	Money (prizes)	43	139	50.64	60	120	20.00	4.19
	Number (dots)	52	129	32.76	73	108	6.77	5.39
	Probability	51	133	36.54	46	132	41.55	0.16
	Time (delays)	57	120	22.42	63	122	18.82	0.14
	Weight	54	126	28.80	65	117	14.86	1.34
3b	Area (squares)	66	124	17.71	67	122	16.01	0.02
	Height (flags)	72	117	10.71	63	127	21.56	1.01
	Length	78	114	6.75	61	126	22.59	2.61
	Money	63	129	22.69	52	135	36.84	1.12
	Number (stars)	59	128	25.46	64	128	21.33	0.14
	Probability	54	131	32.05	49	145	47.51	0.74
	Time	69	121	14.23	57	132	29.76	1.62
	Weight	66	126	18.75	57	130	28.50	0.66

Note: Columns as are in Table S2.

Table S4. Number of "smaller" and "larger" responses in Experiments 4a-4c.

Expt	Dimension	S1S2		S2S1		L1L2		L2L1		S1L2		L2S1	
		n_S	n_L	n_S	n_L	n_S	n_L	n_S	n_L	n_S	n_L	n_S	n_L
4a	Area (squares)	12	27	6	51	19	37	10	40	23	26	11	37
	Height (flags)	22	25	18	32	8	34	7	29	16	27	9	34
	Length	27	27	15	40	17	32	5	35	30	15	18	37
	Money	18	43	6	44	10	44	5	41	10	30	8	33
	Number (stars)	7	37	7	36	10	50	3	44	14	43	3	40
	Probability	13	32	4	38	10	45	2	47	5	30	3	40
	Time	18	23	10	21	17	47	7	40	18	28	5	38
	Weight	12	35	7	45	10	33	2	53	8	37	3	49
4b	Area (circles)	19	38	7	51	20	49	7	51	17	51	1	60
	Height (trees)	15	41	8	50	7	52	7	66	17	49	4	57
	Length	33	32	11	48	27	39	7	67	29	29	15	38
	Money	19	36	10	53	7	49	5	60	16	43	6	68
	Number (stars)	21	41	1	51	9	50	8	62	12	64	3	48
	Probability	18	46	5	46	12	44	8	59	20	38	3	51
	Time	18	36	8	47	25	41	8	47	24	46	4	45
	Weight	15	48	7	64	16	35	3	60	11	58	3	55
4c	Area (circles)	23	50	8	68	8	55	8	69	27	48	11	61
	Height (trees)	22	50	18	48	13	61	6	64	18	59	11	65
	Length	28	36	16	59	19	53	11	60	44	33	14	63
	Money	25	48	7	69	14	56	7	69	21	46	6	66
	Number (squares)	8	76	6	51	20	55	8	59	22	62	8	58
	Probability	15	64	9	60	16	52	9	59	11	52	5	64
	Time	29	40	22	40	18	47	7	55	28	57	3	63
	Weight	18	56	7	76	21	56	8	52	12	57	3	71

Note: Column headings indicate stimulus pair and left-right arrangement (e.g., S1S2 means S1 was on the left and S2 was on the right). Thus, S1S2 and S2S1 refer to the two left-right arrangements of the "small pair"; L1L2 and L2L1 refer to the two arrangements of the "large pair"; and S1L2 and L2S1 are the two left-right orientations of the "big jump" stimuli.

Table S5. Chi-square values for Experiments 4a-4c.

Expt	Dimension	SL	LS	Spatial order	Small Pair vs Big Jump	Large Pair vs Big Jump	Small Pair vs Large Pair
4a	Area (squares)	9.00	65.81	15.24	6.51	1.40	2.09
	Height (flags)	12.12	28.84	2.21	2.95	2.15	9.72
	Length	0.00	36.51	19.32	1.91	10.94	4.27
	Money	40.26	71.54	5.25	0.01	1.57	1.53
	Number (stars)	60.88	86.08	5.14	0.03	0.98	0.62
	Probability	46.23	100.42	11.15	2.76	0.07	2.36
	Time	13.41	49.00	9.63	3.13	0.49	6.40
	Weight	41.67	114.62	12.84	2.33	0.04	1.79
4b	Area (circles)	34.66	122.08	24.87	3.08	2.36	0.06
	Height (trees)	58.61	123.52	9.63	0.53	1.95	4.38
	Length	0.64	77.42	36.79	0.43	6.81	3.96
	Money	43.51	126.73	13.44	2.49	2.40	9.03
	Number (stars)	64.82	128.33	15.29	2.59	0.11	1.68
	Probability	34.18	113.95	20.18	0.01	0.72	0.56
	Time	16.51	89.06	23.80	0.00	0.44	0.35
	Weight	53.56	143.52	19.60	1.60	1.62	0.00
4c	Area (circles)	42.77	129.96	16.64	1.05	9.76	4.66
	Height (trees)	61.39	95.11	3.55	4.04	1.82	10.62
	Length	4.51	89.15	30.57	1.16	9.90	4.15
	Money	38.57	151.14	27.81	0.19	1.29	2.52
	Number (squares)	84.15	112.19	6.23	5.74	0.00	5.36
	Probability	75.60	124.27	6.16	0.96	2.03	0.23
	Time	21.74	83.56	15.95	11.52	0.03	11.49
	Weight	63.29	150.97	18.21	1.91	6.02	1.34

Note: Column SL gives the chi-square values testing whether "smaller" and "larger" responses were equally likely when the large item was on the right (small-large layout). In all cases bar one there is a significant preference for use of "larger" comparatives. Column LS gives the same values for the large-small layout. Larger comparatives are significantly more likely in almost every case. The Spatial order column gives the chi-square test for an association between response and left-right arrangement. The last three columns give the chi-square tests for association used to see whether the proportion of "smaller" and "larger" responses depended on the magnitudes of the presented objects.

Table S6. Modal responses in Experiments 1a and 1b.

Expt		"Smaller"	"Larger"	Overall	*n*
	Area (squares)	smaller	bigger	bigger	202
	Height (flags)	lower	higher	higher	194
	Height (trees)	shorter	taller	taller	203
	Length	shorter	longer	longer	200
1a	Money (rewards)	less	larger	larger	202
	Number (dots)	fewer	more	more	201
	Probability	smaller	greater	greater	194
	Time	shorter	longer	longer	191
	Weight	lighter	heavier	heavier	208
	Area (circles)	smaller	bigger	bigger	368
	Height (flags)	lower	higher	higher	344
	Height (trees)	shorter	taller	taller	360
	Length	shorter	longer	longer	365
1b	Money	less	more	more	362
	Number (squares)	fewer	more	more	359
	Probability	smaller	greater	greater	330
	Time	shorter	longer	longer	355
	Weight	lighter	heavier	heavier	366

Note: The "Smaller" and "Larger" columns give the most common responses among those classified as "smaller" and "larger"; the Overall column gives the most common response ignoring category. The *n* values are the sample sizes for each dimension after removing unclassifiable responses.

Table S7. Modal responses in Experiments 2a and 2b.

Expt		"Smaller"	"Larger"	Overall	*n*
2a	Area (squares)	smaller	larger	larger	203
	Height (trees)	shorter	taller	taller	199
	Length	shorter	longer	longer	203
	Money (prizes)	smaller	bigger	bigger	181
	Number (dots)	fewer	more	more	199
	Probability	less	greater	greater	186
	Time (delays)	shorter	longer	longer	195
	Weight	lighter	heavier	heavier	201
2b	Area (circles)	smaller	bigger	bigger	135
	Height (trees)	shorter	taller	taller	134
	Length	shorter	longer	longer	134
	Money	less	greater	greater	135
	Number (dots)	less	more	more	134
	Probability	less	greater	greater	132
	Time	shorter	longer	longer	134
	Weight	less	heavier	heavier	135

Table S8. Modal responses in experiments 4a-4c.

Expt		"Smaller"	"Larger"	Overall	n
4a	Area (squares)	smaller	bigger	bigger	299
	Height (flags)	lower	higher	higher	261
	Length	shorter	longer	longer	298
	Money	less	more	more	292
	Number (stars)	less	more	more	294
	Probability	less	greater	greater	269
	Time	shorter	longer	longer	272
	Weight	less	heavier	heavier	294
4b	Area (circles)	smaller	bigger	bigger	371
	Height (trees)	shorter	taller	taller	373
	Length	shorter	longer	longer	375
	Money	less	larger	larger	372
	Number (stars)	less	more	more	370
	Probability	less	greater	greater	350
	Time	shorter	longer	longer	349
	Weight	less	heavier	heavier	375
4c	Area (circles)	smaller	larger	larger	436
	Height (trees)	shorter	taller	taller	435
	Length	shorter	longer	longer	436
	Money	less	more	more	434
	Number (squares)	less	more	more	433
	Probability	less	greater	greater	416
	Time	shorter	longer	longer	409
	Weight	lighter	heavier	heavier	437