Cognition 109 (2008) 251–266

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Judgments of discrete and continuous quantity: An illusory Stroop effect

Hilary C. Barth

Department of Psychology, Wesleyan University, 207 High Street, Middletown, CT 06459, USA

ARTICLE INFO

Article history:
Received 4 July 2007
Revised 5 September 2008
Accepted 14 September 2008

Keywords:
Numerical cognition
Discrete quantity
Continuous quantity

ABSTRACT

Evidence from human cognitive neuroscience, animal neurophysiology, and behavioral research demonstrates that human adults, infants, and children share a common nonverbal quantity processing system with nonhuman animals. This system appears to represent both discrete and continuous quantity, but the proper characterization of the relationship between judgments of discrete and continuous quantity remains controversial. Some researchers have suggested that both continuous and discrete quantity may be automatically extracted from a scene and represented internally, and that competition between these representations leads to Stroop interference. Here, four experiments provide evidence for a different explanation of adults' performance on the types of tasks that have been said to demonstrate Stroop interference between representations of discrete and continuous quantity. Our well-established tendency to underestimate individual two-dimensional areas can provide an alternative explanation (introduced here as the "illusory-Stroop" hypothesis). Though these experiments were constructed like Stroop tasks, and they produce patterns of performance that initially appear consistent with Stroop interference, Stroop interference effects are not involved. Implications for models of the construction of cumulative area representations and for theories of discrete and continuous quantity processing in large sets are discussed.

1. Introduction

There is now a large body of evidence from human cognitive neuroscience, animal neurophysiology, and behavioral research demonstrating that humans of all ages and nonhuman animals share a common ability to represent approximate numerical magnitudes (e.g. Brannon, 2006; Dehaene, 1997; Gallistel & Gelman, 2000; Nieder, 2005). This is striking in part because numerical magnitude is a relatively abstract property of a set: it cannot be determined by assessing a single perceptual property of a stimulus. For example, 20 plums will take up much more space than 40 grapes, so 40 grapes cannot be judged as "more" based on a perceptual property such as total amount of visible purple surface area. Evidence for the processing of numerical (discrete) quantity is evidence for a kind of abstract thought: in carefully controlled experiments, response based on number requires the subject to ignore salient perceptual properties of the stimuli at hand. Many nonhuman species can make discriminations based on discrete quantity, as can very young children and adults even when they are prevented from exact counting (e.g. Barth, Kanwisher, & Spelke, 2003; Barth, La Mont, Lipton, & Spelke, 2005; Cantlon & Brannon, 2006; Hauser, Tsao, Garcia, & Spelke, 2003; Meck & Church, 1983).

This ability appears to arise from what is often referred to as an "analog magnitude representation" system: the discrete numerosity of the set is internally coded by a mental magnitude, which may be pictured as a number line or a vessel filled with liquid. Ratio-dependent discrimination in accord with Weber's Law is a primary signature of this analog magnitude system: discriminability depends on the ratio of the quantities to be compared (e.g. Dehaene, 1997; Gallistel & Gelman, 1992; Gallistel & Gelman, 2000). Of course, we can also make perceptual judgments about non-numerical quantities such as duration, length, area, or volume. Converging evidence suggests that the

E-mail address: hbarth@wesleyan.edu

0010-0277/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2008.09.002

analog magnitude system may be involved in the representation of many continuous quantities as well as number (e.g. Balci & Gallistel, 2006; Walsh, 2003; vanMarle & Wynn, 2006; Brannon, Lutz, & Cordes, 2006; Meck & Church, 1983; for a recent review, see Feigenson (2007)).

The relation between continuous and discrete quantity processing can be difficult to characterize: numerical quantity is correlated with various forms of continuous quantity, and these cannot be independently varied. Attempts to identify the specific quantities that control human and animal behavior under a variety of experimental conditions have led to controversy, particularly with respect to studies of infants and young children, which often involve neither extensive training nor explicit instruction. For example, under some conditions, infants appear sensitive to changes in numerical quantity (Brannon, Abbott, & Lutz, 2004; Lipton & Spelke, 2003; Xu & Spelke, 2000). In other cases they appear oddly insensitive to numerical changes that should be equally discriminable, instead responding to changes in continuous quantity (Clearfield & Mix, 1999; Feigenson, Carey, & Hauser, 2002; Feigenson, Carey, & Spelke, 2002).

These and many other findings have led to a wide variety of proposals describing the kinds of basic, unlearned abilities that might underlie nonverbal quantitative processing. Some researchers suggest that claims about the detection of discrete quantity in very young children are unfounded, as the critical results may be explained in terms of sensitivity to continuous amount. These researchers suggest instead that sensitivity to numerical quantity is relatively late-developing and probably dependent upon language (Mix, Huttenlocher, & Levine, 2002; Newcombe, 2002). Others have appealed to multiple distinct core knowledge capacities in order to explain the wide range of findings with infants, adults, and nonhuman animals (Carey, 2004; Carey & Sarnecka, 2006; Feigenson, Dehaene, & Spelke, 2004; Hauser & Spelke, 2004; Xu, 2003). Still others have suggested that the attentional demands of a particular experimental situation might determine the quantitative dimension that governs behavior in that situation (Cordes & Gelman, 2005; Hurewitz, Gelman, & Schnitzer, 2006). The last suggestion found empirical support in a series of recent behavioral studies in adults. The authors proposed specifically that Stroop interference between analog magnitude representations of continuous and discrete quantity might explain some of the apparent discrepancies in the nonverbal quantity processing literature (Hurewitz et al., 2006). According to the authors' hypothesis, we might automatically extract and represent both discrete and continuous quantities from a scene, and these representations might then compete for control of behavior. The attentional demands of particular study conditions could easily bias the competition in favor of discrete or continuous quantity, producing different results depending on the paradigm and stimuli in question. These researchers argue that adults do appear susceptible to competition between continuous and discrete dimensions of quantity, based on the observation of performance patterns consistent with Stroop interference during quantity judgments (Hurewitz et al., 2006).

These claims of Stroop interference arise from studies in which adult participants were presented with pairs of arrays containing up to seven filled circles, and were asked to make judgments of either total continuous surface area or discrete number (Hurewitz et al., 2006). The irrelevant dimension in each task could provide information helpful to the task (congruent trials), harmful to the task (incongruent trials), or neutral to the task (neutral trials). Often, performance costs for incongruent trials in Stroop paradigms such as these are interpreted to mean that the observer is obligated to process the irrelevant dimension in the task: the automatically-extracted irrelevant information cannot be ignored, and so it interferes with judgments about the relevant dimension. For example, many studies have demonstrated that the physical size of an Arabic numeral interferes with judgments of its numerical magnitude (Besner & Coltheart, 1979) and vice versa (Henik & Tzelgov, 1982). In Hurewitz et al.'s study, participants could perform both the continuous and discrete tasks accurately and rapidly, and in both cases error rate and reaction time were higher for incongruent than congruent trials. The authors concluded that discrete quantity information interfered with continuous quantity judgments and vice versa, that competition between representations was responsible, and that adults seem to rapidly and automatically extract both discrete number and total continuous amount when confronted with a set (Hurewitz et al., 2006).

The present paper considers an alternative explanation for experimental results that appear to indicate Stroop interference in quantity judgments involving sets of objects, focusing on judgments of cumulative area. At least one potentially unwarranted – and consequential – assumption is built into the adult experiments that are said to demonstrate these Stroop interference effects (see Algom, Dekel, & Pansky, 1996, for a related discussion). To understand this assumption, first consider a hypothetical cumulative area judgment task, in which participants are presented with two sets of disks and instructed to choose the set with the larger cumulative area. The researcher manipulates the *ratio* of the cumulative areas such that some comparisons are easy (for example, the ratio of cumulative areas across the two sets might be 1:2) and others are more difficult (for example, the cumulative area ratio might be 7:8). The researcher also manipulates *congruency* by creating two different types of trials. Trials in which the set with the larger cumulative area has a smaller number of disks are classified as "incongruent," while "congruent" trials are those in which the set with the larger cumulative area also has a larger number of disks. Performance is likely to be worse for more difficult comparisons in general (those involving ratios closer to 1:1), but suppose there is also an apparent effect of congruency in addition to the ratio effect: performance is worse for the incongruent trials than for the congruent trials. Such congruent vs. incongruent trial differences have previously been explained in terms of Stroop interference between automatically-extracted representations of cumulative area and number (Hurewitz et al., 2006). The presence of a true congruency effect, dissociable from any effect of ratio, is the critical evidence for Stroop interference.

The hidden assumption in the Stroop interference explanation lies in the idea that performance should be assessed relative to the ratio of *veridical cumulative areas* across the two sets. This assumption is only valid if we make additional assumptions about the mechanisms underlying cumulative area judgments in these tasks: that participants either (a) estimate cumulative area without first arriving at an estimate of individual item size (for example, through some direct perceptual summation of all of the appropriately-colored pixels in the set) or (b) compute cumulative area by first estimating individual item size very accurately, with no systematic bias (for example, through the summation of accurate estimates of individual disk areas, or by performing some computation over accurate estimates of individual disk area and total number of disks). These assumptions are problematic for two reasons. First, we have no compelling reason to assume that direct extraction of cumulative area (with no regard for object boundaries) must be possible. Second, humans are actually rather poor at the estimation of individual two-dimensional areas. Previous psychophysical research on judgments of individual circles' areas shows that we tend to underestimate circles' areas considerably: subjective area increases more slowly than physical area (in other words, the Stevens exponent for the area of a circle is not 1.0; it is often measured at about 0.8, with individual variation, e.g. Chong & Treisman, 2003; Teghtsoonian, 1965).

How could participants' underestimation of individual disk areas affect experimenters' inferences about the cognitive processes underlying these tasks? Patterns of bias in estimates of individual area are relevant here because, if we do judge cumulative area by making use of estimates of individual element areas (or estimates of average element size; Chong & Treisman, 2003; Chong & Treisman, 2005), then performance differences for congruent vs. incongruent trials may not reflect true effects of congruency (which in turn are necessary to provide evidence of Stroop interference). This is because bias in estimates of individual areas should produce systematic differences in performance on congruent and incongruent trials: the cumulative *subjective* area ratios associated with incongruent trials should be more difficult (closer to 1:1) than those associated with congruent trials, even in an experiment designed to equate cumulative *veridical* area ratios across trial types (see Fig. 1 for an example). Performance should suffer on incongruent trials, but not because they are incongruent. Rather this performance cost would be properly attributed to an effect of ratio (discrimination difficulty), and it would not be accurate to interpret the cost as evidence of Stroop interference. The incongruent trial cost observed in previous studies may therefore be misleading: there may be no evidence for the existence of competition and interference between automatically-extracted representations of number and cumulative area. Though the task appears to be a Stroop task in its design and in its effects, the apparent Stroop interference effects found previously could be illusory: there may be no difference in the cognitive processing of congruent and incongruent trials.

The aim of the present study was to distinguish between these two accounts of adults' quantity judgments: the interference hypothesis (Hurewitz et al., 2006) and the alternative "illusory-Stroop" hypothesis described above. The interference hypothesis states that continuous and discrete quantity information are both automatically extracted from visually-presented sets, such that their representations compete for control of participants' behavior, causing interference on incongruent trials. The illusory-Stroop hypothesis states that our well-established tendency to underestimate individual areas can explain performance costs observed for incongruent trials (because incongruent trials are simply associated with more difficult subjective area ratios), and so incongruent trial costs do not necessarily constitute evidence of Stroop interference.

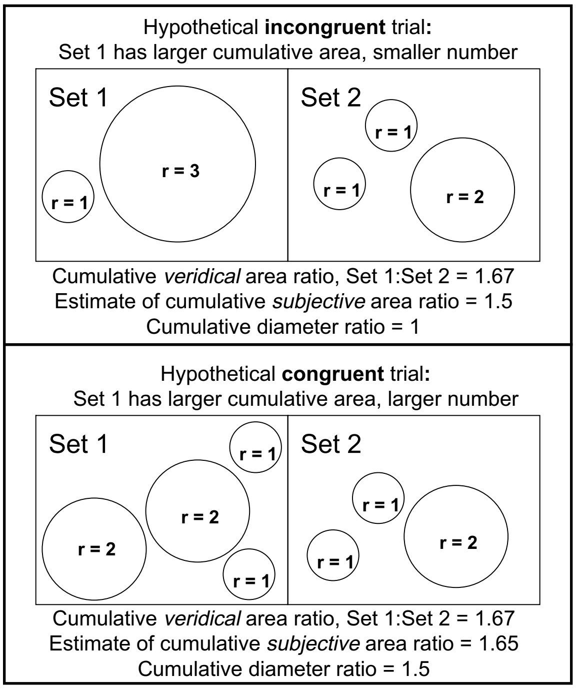
Hypothetical **incongruent** trial:
Set 1 has larger cumulative area, smaller number
Set 1
r = 1
r = 3
Set 2
r = 1
r = 1
r = 2
Cumulative *veridical* area ratio, Set 1:Set 2 = 1.67
Estimate of cumulative *subjective* area ratio = 1.5
Cumulative diameter ratio = 1
Hypothetical **congruent** trial:
Set 1 has larger cumulative area, larger number
Set 1
r = 1
r = 2
r = 2
r = 1
Set 2
r = 1
r = 2
r = 1
Cumulative *veridical* area ratio, Set 1:Set 2 = 1.67
Estimate of cumulative *subjective* area ratio = 1.65
Cumulative diameter ratio = 1.5

Fig. 1. Schematic depiction of hypothetical congruent and incongruent trial types (not actual stimuli). Even though veridical cumulative area ratios are equated across the two trial types, the ratio of subjective areas is much closer to 1:1 for the incongruent trial (where "subjective area" is a power function of veridical disk area with an exponent of 0.8), and the ratio of cumulative diameter is 1:1 for the incongruent trial (where cumulative diameter is equal to the sum of the individual diameters). If participants do not have access to veridical estimates of area, incongruent trials are likely to be associated with more difficult discriminations (with ratios closer to 1:1) than congruent trials, potentially producing incongruent trial performance costs that are not direct results of trial incongruency.

Four experiments were conducted to test the predictions of these competing hypotheses. In the first three experiments, adult participants were shown two sets of filled circles on a computer screen and asked to make a judgment about the set with the larger total continuous amount (cumulative area), under a variety of experimental conditions. In the fourth experiment, participants judged discrete number rather than continuous amount, with the same stimuli presented in the third experiment. In all experiments, eight different models of individual participants' performance were explored in an attempt to establish, first, the content of the quantitative representation controlling behavior in these tasks and, second, the presence or absence of a clear interdimensional Stroop interference effect. The results show that the illusory-Stroop hypothesis can better account for the data: there is no evidence of Stroop interference between competing representations of discrete and continuous quantity in these experiments.

2. Experiment 1: continuous quantity, homogeneous sets, implicit instructions

Adult participants saw two rapidly presented sequential sets of filled circles on a computer screen and were instructed to choose one of the two sets. In this experiment, the instructions to maximize total cumulative area were implicit, and all of the circles in a set were the same size. Participants were given instructions suggesting that total continuous amount (cumulative area) should be maximized ("The circles on the screen represent food; choose which set you would want if you were hungry.") Post-tests were used to discount data from participants whose responses suggested intentional judgments based on criteria other than cumulative area.

2.1. Methods

2.1.1. Participants

Thirty-four adults participated for pay or for credit in the Harvard University Psychology Department Study Pool. All had normal or corrected-to-normal vision. All experimental procedures were approved by an Institutional Review Board and informed consent was obtained from all participants.

2.1.2. Materials

Stimuli were presented on a PowerMac G4 computer with a ViewSonic GS790 color monitor set at 1024 by 768 pixels resolution. For each trial, one array of red disks appeared on the screen and remained for 400 ms, followed by a blank screen for 400 ms and a second array for 400 ms. Disk size was always the same within an array, but disk size and number always varied between arrays. Disks were positioned in a pseudorandom arrangement without touching or overlapping. Array numerosities ranged from 9 to 32, and disk diameters ranged from 10 to 20 pixel-widths such that the largest possible area of a disk was four times the smallest possible area. For half of these trials, the set with the larger number had the larger total cumulative disk area (congruent trials); for the other half, the set with the larger number had the smaller total cumulative area (incongruent trials). Congruent and incongruent trials in these studies are always defined with respect to the congruency of number and physical/veridical cumulative areas. The two trial types were interleaved. There were eight possible set 1:set 2 *numerosity* ratios (four congruent and four incongruent, collapsed across set 2:set 1 and set 1:set 2) and eight ratios describing the total *cumulative areas* of set 1:set 2 (four congruent and four incongruent). Ratios were matched as closely as possible for congruent vs. incongruent trials. Number and cumulative area ratios were necessarily different within each individual trial, but across the entire set of trials, they were closely matched. Congruent trial numerosity ratios (collapsed across set 1:set 2) were approximately 0.38, 0.58, 0.75, and 0.78, and congruent trial cumulative area ratios were approximately 0.38, 0.56, 0.74, and 0.8. Incongruent trial numerosity ratios (collapsed across set 1:set 2) were approximately 0.38, 0.64, 0.67, and 0.78, and incongruent trial cumulative area ratios were approximately 0.38, 0.63, 0.67, and 0.76.

2.1.3. Procedure

Participants were seated approximately 0.75 m from the monitor; viewing distance was not controlled. They were told that the red items represented food, and that they should choose the set that they would want if they were hungry. They received no explicit instructions as to the basis for choosing one group over the other – the purpose was to determine which cues would be the default bases for choice, when the task implicitly suggested that total continuous amount should be maximized. Participants were asked not to consider their choices carefully, but to choose rapidly based on first impressions. After viewing each trial, participants made a two-alternative forced-choice (2AFC) judgment about the presented pair of sets, pressing a left-hand key to choose the first set and a right-hand key to choose the second. After completing six blocks of 96 trials each (576 trials total), participants were questioned about their goals during the tasks and any strategies they thought they had used to accomplish those goals. The task was intended to lead participants to attempt to maximize continuous amount (in this case, cumulative area of all disks), but the vague instructions might have been interpreted very differently by different participants. Post-tests were used to discount all data from participants whose responses suggested intentional judgments based on numerosity, individual item size, or any criteria other than total continuous amount. Twenty out of 34 participants stated unambiguously that they had attempted to maximize continuous amount; these participants often described strategies such as "choosing the one with the most red." Eleven of 34 participants described other goals based on quantities of some kind: some explicitly chose based on the set with the larger number of items, and some described choosing based on some combination of number and size, but did not explicitly mention maximizing total amount. Three additional participants described different non-quantitative strategies. Data from the latter two groups were excluded from the analysis.

2.2. Results and discussion

Both the interference hypothesis and the illusory-Stroop hypothesis predict that participants should respond less accurately on the incongruent trials, and this was the case (consistent with previous data, Hurewitz et al., 2006). These competing hypotheses have different explanations of the source of this performance cost, however, which in turn lead to specific and testable predictions as follows. The interference hypothesis holds that the incongruent trial cost arises from competition between representations of continuous quantity and discrete number: that this cost represents a true effect of congruency, dissociable from any effects of ratio (discrimination difficulty). The interference hypothesis predicts, therefore, that differences between congruent and incongruent trials should endure even when we assess performance with respect to cumulative subjective area (when we take into account the possibility that participants underestimated individual element area). There should be separable effects of ratio and congruency, so performance will not be determined by the difficulty of the discrimination alone: two separate curves should be required to describe participants' performance on congruent vs. incongruent trials.

The illusory-Stroop hypothesis, on the other hand, explicitly states that this interference-based interpretation of the incongruent trial cost results from the problematic assumption that participants make judgments of cumulative area based on veridical, rather than underestimated, estimates of individual area. It also states that when individual area underestimation is taken into account, the differences between congruent and incongruent trials will be attributable to ratio effects. Therefore the illusory-Stroop hypothesis predicts that the apparent congruency effects should disappear when we assess performance with respect to cumulative subjective area, in accord with previous research on the assessment of the areas of two-dimensional shapes. There should be no true effect of congruency dissociable from effects of ratio, so a single curve should be able to describe performance on both congruent and incongruent trial types: performance should determined by the difficulty of the discrimination alone.

In order to test these predictions quantitatively, individual performance was assessed with respect to eight possible models of cumulative area judgment. There were four main categories of models (see Appendix) each with two subtypes ("pooled" and "unpooled" fits – see below). One of the four model categories (AREA) assumes that participants make judgments based on veridical cumulative area (corresponding to assumptions made in previous studies; Hurewitz et al., 2006). Another (POW) assumes that participants tend to underestimate area, such that each circle's subjective area is a power function of its physical area with a Stevens exponent of 0.8, comparable to measurements from previous studies (e.g. Chong & Treisman, 2003; Teghtsoonian, 1965). A third model (DIAM) assumes that participants simply use disk diameter as a rough estimate of area, which does occur under some conditions (e.g. Krider, Raghubir, & Krishna, 2001), in effect maximizing cumulative diameter.[1] A fourth type of model was also considered (NUM), which assumes that participants simply chose based on the total number of disks in each set.

Fig. 2A depicts the group data plotted with respect to these four potential models of performance; however, the group data were highly variable and the analyses to follow focus on individual participants' data. Note that the incongruent and congruent trials are fairly well matched for difficulty with regard to veridical area ratios: in the AREA plot, in which cumulative veridical area ratios are shown on the *x*-axis (leftmost panel of Fig. 2A), the congruent and incongruent trials are located at roughly equal *x*-values. This difficulty-matching across trial types breaks down when performance is assessed with respect to models that account for individual area underestimation, as in the POW and DIAM plots (in which cumulative subjective areas or cumulative diameters, respectively, are shown on the *x*-axis): the incongruent trials now appear shifted toward the center of the *x*-axis relative to the congruent trials. According to the illusory-Stroop hypothesis, this ratio-based difference in trial difficulty is the cause of the incongruent trial cost: the incongruent trial cost reflects an effect of ratio, not a true effect of congruency.

Within each one of the four main model types, two subtypes were also tested. The first was a single-curve model in which data from congruent and incongruent trials were pooled (one curve was fit to each participant's data, regardless of trial type). The second was a two-curve model in which data from congruent and incongruent trials were unpooled (two separate curves were fit to each participant's data – one for congruent and one for incongruent trials). The critical predictions for the results of these analyses are as follows. If interference does indeed produce the performance costs for incongruent trials, as predicted by the interference hypothesis, then the two-curve models of performance should receive more empirical support than the single-curve models. This is because the interference hypothesis predicts that there should be a true effect of congruency, separable from the effect of discrimination ratio, even if alternative models that account for individual area underestimation are considered. If there is no clear evidence of interference, as predicted by the illusory-Stroop hypothesis, however, a single curve should be sufficient to fit data from both trial types, and the one-curve model should receive more support from the data. This is because the illusory-Stroop hypothesis predicts that ratio effects alone – not congruency – will explain the observed performance patterns, if we consider the possibility that single element areas are likely to be underestimated.

For each of the eight models considered (AREA-single curve, AREA-two curve, POW-single curve, etc.), a sigmoidal function was fitted to the data produced by each participant (see Appendix). Likelihood ratios were used to assess the best model of performance for each participant (Glover & Dixon, 2004). The likelihood ratio determines the explanatory power of a particular model. Because the likelihood ratio is based on model fit, it will favor more complex models over simpler ones (generally, more parameters lead to better fits). The models tested here vary in complexity in that the single-curve models are nested within their corresponding two-curve models: the single-curve models are special cases of the two-curve models in which parameters are shared for both congruent and incongruent trial types. In cases like this, it is necessary to correct for model complexity. Akaike's information

[1] These models do not necessarily identify the mechanisms that might create these quantity estimates. For example, a participant whose data support the DIAM model could be (unconsciously) summating the diameters of all of the disks, or performing a computation over the diameter of a single disk and the total number of disks in the set. The model simply says that an estimate equivalent to cumulative diameter is the quantity upon which the choice is based.

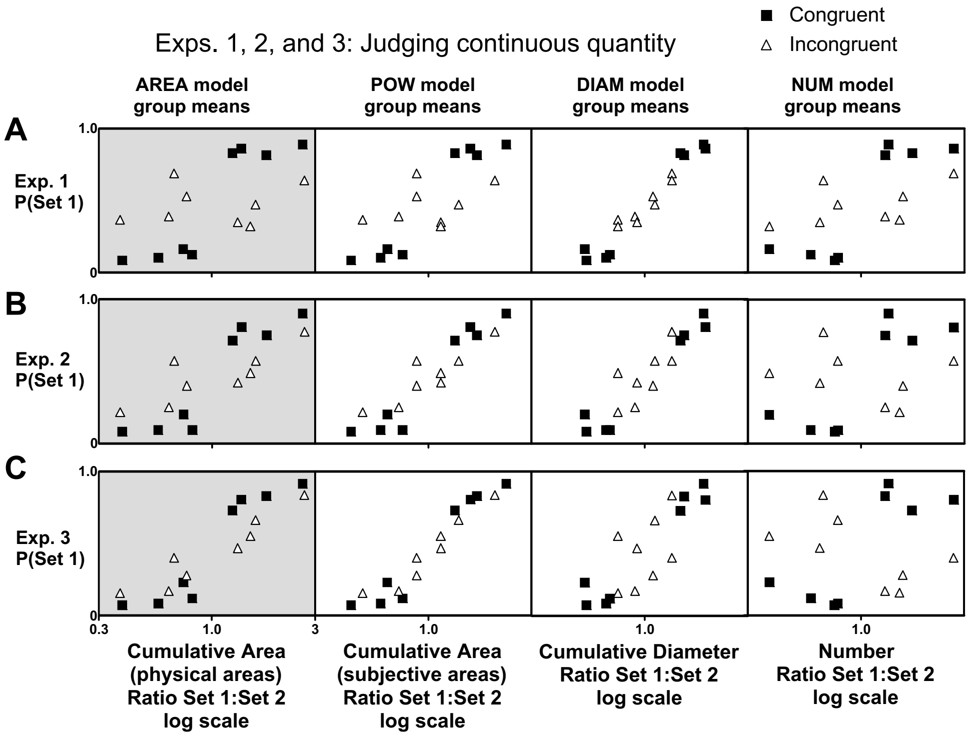

Fig. 2. Group data from the continuous quantity judgment tasks of Experiment 1 (Fig. 2A: implicit instructions, homogeneous sets), Experiment 2 (Fig. 2B: explicit instructions, homogeneous sets), and Experiment 3 (Fig. 2C: explicit instructions, heterogeneous sets). The *x*-axis shows the ratio of the quantities that are compared according to each model. A shaded graph indicates that those data are plotted against the comparison ratios that *should* determine performance on the task, according to the task instructions (cumulative area, in this Figure). The ratios that are more difficult to discriminate are those close to 1, and the easiest ratios are those farther from 1. The *y*-axis shows the proportion of trials in which participants chose the first set. Group means are plotted with respect to the four main types of models tested. The AREA column shows the ratios of the total *Cumulative Veridical Areas* of sets (the quantity that should be maximized in the task). The POW column depicts the data plotted with respect to ratios of total *Cumulative Subjective Areas* of the sets, assuming individual areas are underestimated in accord with previous psychophysical findings. The DIAM column depicts the data plotted with respect to *Cumulative Diameter* ratios. The NUM column depicts the data plotted with respect to ratios of the *Number* of items in the sets. The data produced for all three continuous quantity tasks were highly variable: see Tables 1–3 for the best models of individual participants' performance.

criterion (AICc) was used to identify the best model of performance, corrected for model complexity, for each participant (AICc = AIC corrected for small sample sizes; Glover & Dixon, 2004; Burnham & Anderson, 2002; Hurvich and Tsai; 1989; Akaike, 1973). Differences in AICc scores, which provide a measure of the relative explanatory power of each model, are reported in Table 1 for every model and every participant in Experiment 1. The table identifies the best model for each participant in addition to any other models that were reasonably well supported (Burnham & Anderson, 2002). Because this method will always choose a "best" model even if no model provides a good explanation of the data, fits with R^2 values less than 0.80 were excluded (signified by blank lines in the table).

For every participant, the model that best explained the data was a single-curve model: a single explanation, based solely on the ratio of quantities being compared, could best account for performance patterns on both congruent and incongruent trials. The data appear to support the illusory-Stroop hypothesis rather than the interference hypothesis: apparent effects of congruency disappear when participants' responses are considered with respect to models of performance that account for the underestimation of element area. Therefore, there appears to be no true congruency effect at work: the data do not provide evidence of Stroop interference between competing representations of discrete and continuous quantity. Rather, incongruent trials may produce lower levels of performance because, when the well-documented tendency to underestimate individual disk area is taken into account, the trials categorized as "incongruent" simply require more difficult discriminations.

Does this finding generalize to an explicit cumulative area judgment task? Participants in the first experiment received instructions that did not explicitly require them to choose the set with the larger total cumulative area, which might have led participants to perform in a manner that would not apply to an explicit task. Experiments 2 and 3 tested this possibility.

3. Experiment 2: continuous quantity, homogeneous sets, explicit instructions

Participants in Experiment 2 were presented with the same stimuli used in Experiment 1, but they were asked explicitly to choose the set with the larger continuous amount (cumulative area).

Table 1
Experiment 1, best models for each participant as determined by AICc differences

	$AREA_1$	$AREA_2$	POW_1	POW_2	$DIAM_1$	$DIAM_2$	NUM_1	NUM_2	Best R2
1					**		*****	**	.822
2			*****						.965
3			*****						.955
4									
5							*****		.987
6			*****						.895
7					*****	*			.901
8	*		*****						.884
9					*****				.798
10							*****		.956
11							*****	**	.830
12							*****		.965
13									
14					*****				.902
15					*****				.938
16							*****		.974
17					*****				.901
18					*****			*	.852
19			*****						.906
20			*****						.993

The leftmost column gives participant numbers and the next eight columns represent the eight models tested. Best model of the eight models tested, *****; other models with substantial empirical support relative to the best, **; other models with considerably less support, *; models that are out of the running or AICc difference > 10 (essentially no empirical support; Burnham & Anderson, 2002), left blank. R-squared values are listed for the best supported model.

3.1. Methods

3.1.1. Participants

Eighteen adults participated for pay or for credit in the Harvard University Psychology Department Study Pool. All had normal or corrected-to-normal vision. All experimental procedures were approved by an Institutional Review Board and informed consent was obtained from all participants.

3.1.2. Stimuli and procedure

The stimuli and procedure were as in Experiment 1, except that these participants were explicitly instructed to choose the set with the greater aggregate area. This was described in two ways to make sure participants understood the goal: they were told to maximize the total summed area of all the disks, or in other words, to choose the set with more red pixels. All other aspects of the procedure were the same as in Experiment 1, but no participants' data were excluded.

3.2. Results and discussion

Fig. 2B depicts the group means for Experiment 2 plotted with respect to each of the four models described earlier. Again, there was great across-participant variability, despite the explicit instructions in this task; the analyses focus on the individual data. Tests of individual participants' performance were carried out as in Experiment 1. Table 2 summarizes the results of this analysis for every participant in Experiment 2, listing the best model for each. Again, there is no evidence of Stroop interference between competing representations of discrete and continuous quantity. For every participant, the model that best explained the data was a single-curve model, demonstrating that a single explanation provides the best account of performance on both congruent and incongruent trials. The performance cost for incongruent trials again appears to be an effect of ratio (discrimination difficulty), not a true effect of congruency dissociable from ratio.

The results of Experiment 2, like those of Experiment 1, provide evidence against the idea that conflicting numerical information interfered with these judgments of cumulative continuous quantity. In Experiment 2, however, all of the circles within a set were the same size, as in Experiment 1. It is possible that the assessment of cumulative area across homogeneous sets is a special case, and that the use of heterogeneous sets would produce different results. Experiment 3 tested this possibility.

4. Experiment 3: continuous quantity, heterogeneous sets, explicit instructions

The experiment was conducted again with explicit instructions and one additional change: in this experiment the disk size varied both within and across sets.

4.1. Methods

4.1.1. Participants

Seventeen adults participated for pay or for credit in the Harvard University Psychology Department Study Pool. All had normal or corrected-to-normal vision. All experimental procedures were approved by an Institutional Review Board and informed consent was obtained from all participants.

4.1.2. Stimuli and procedure

Participants were again explicitly instructed to choose the set with the greater aggregate area. The stimuli and procedure were the same as in Experiment 2 except that in Experiment 3, disk size varied within a set as well as

Table 2
Experiment 2, best models for each participant as determined by AICc differences

	$AREA_1$	$AREA_2$	POW_1	POW_2	$DIAM_1$	$DIAM_2$	NUM_1	NUM_2	Best R^2
1					*****				.854
2	*		*****						.895
3			*****						.903
4					*****				.859
5			*****						.945
6					*****				.940
7	*****								.842
8					*****				.921
9					*****				.947
10	*		*****						.893
11	*****								.957
12	*****								.952
13							*****	**	.873
14									
15	*****								.903
16							*****	*	.853
17					*****				.912
18									

The leftmost column gives participant numbers and the next eight columns represent the eight models tested. Best model of the eight models tested, *****; other models with substantial empirical support relative to the best, **; other models with considerably less support, *; models that are out of the running or AICc difference > 10 (essentially no empirical support), left blank. R-squared values are listed for the best supported model.

between sets. The fixed disk sizes from Experiments 1 and 2 now became the *average* sizes for Experiment 3. Disk sizes varied from four pixel-widths less than their previous fixed diameters (in Experiments 1 and 2) to four more than their previous diameters, such that sets that had previously contained the smallest disks, with 10-pixel-width diameters, now contained disks that varied from 6–14. Sets that had previously contained the largest disks, with 20-pixel-width diameters, now contained disks that varied from 16 to 24. This meant that the diameters varied from 60%–140% of the mean for the smallest-disk sets, and from 80%–120% of the mean for the largest-disk sets, in order to ensure that the mean diameters of the sets corresponded well with the mean areas (for example, the sets with mean diameters of 10 pixel-widths also had mean areas close to 25π pixels).

4.2. Results and discussion

Participants chose the set with the larger number of items on 77% of the trials. Fig. 2C depicts the group means for Experiment 3 plotted with respect to each of the four model types. Again, the group data were highly variable (this was the case for all three experiments depicted in Fig. 2, even though participants completed a large number of trials and received explicit instructions for Experiments 2 and 3). Tests of individual participants' performance were carried out as in the previous experiments[2]; Table 3 summarizes the results of this analysis for each participant in Experiment 3, listing the best model for each. Fig. 3 depicts individual data from four participants in Experiment 3 plotted for all four of the models, providing an example of the range of responses produced.

For Experiments 1, 2, and 3, each type of single-curve model provided the best explanation for some of the participants (approximately 11% AREA, 25% DIAM, 33% POW, and 18% NUM). For 11% of the participants, none of the four models tested provided a good explanation of the data. Considering the results summarized in Tables 1–3 taken together, there was not a single participant for whom an unpooled two-curve model received the most empirical support. This means that for every participant, a single curve fit both incongruent trials and congruent trials. Therefore there was no evidence of an effect of trial congruency separable from effects of discrimination ratio, and therefore no support for an interference-based explanation of these data.

Reaction time (RT) data also support the illusory-Stroop interpretation. Both the Stroop interference hypothesis and the illusory-Stroop hypothesis predict that there should be differences in RT between congruent and incongruent trials, but the two hypotheses do provide differing explanations of these RT differences which lead to testable predictions. The illusory-Stroop hypothesis predicts that there should be systematic differences in RT patterns across participants whose data are best fit by the AREA, POW, and DIAM models. Recall that the cumulative diameter discrimination ratios for the incongruent trials are *much* more difficult (closer to 1:1) than the cumulative diameter ratios for the congruent trials (see Fig. 2). This means that a participant whose data are best fit by the DIAM model (who is, in effect, underestimating individual disk areas severely) is likely to exhibit a clear RT difference between congruent and incongruent trials, due to the systematic difference in cumulative diameter ratios across

[2] Power functions describing perceived individual element size have been measured in tests of magnitude estimation, when the size of an individual circle is explicitly estimated. These tasks show that the exponent determining the perceived sizes of individual circles and many other 2D shapes is around 0.8. The same rule also appears to apply for judgments of mean size across heterogeneous sets (that is, the size of an individual circle is estimated according to this rule, and so is the mean size within a set; Chong & Treisman, 2003; Chong & Treisman, 2005). For this reason, the models may reasonably be expected to apply in the same manner for both homogeneous and heterogeneous sets.

Table 3
Experiment 3, best models for each participant as determined by AICc differences

	$AREA_1$	$AREA_2$	POW_1	POW_2	$DIAM_1$	$DIAM_2$	NUM_1	NUM_2	Best R^2
1	*		*****						.985
2			*****						.973
3							*****		.952
4	**		*****						.908
5					*****				.941
6			*****						.941
7									
8	*		*****						.983
9							*****		.970
10			*****						.936
11			*****						.975
12									
13					*****				.933
14	*****		*						.883
15	*****								.946
16			*****		*				.845
17	*		*****						.953

The leftmost column gives participant numbers and the next eight columns represent the eight models tested. Best model of the eight models tested, *****; other models with substantial empirical support relative to the best, **; other models with considerably less support, *; models with AICc difference > 10 (essentially no empirical support) or models that are out of the running entirely, left blank. R-squared values are listed for the best supported model.

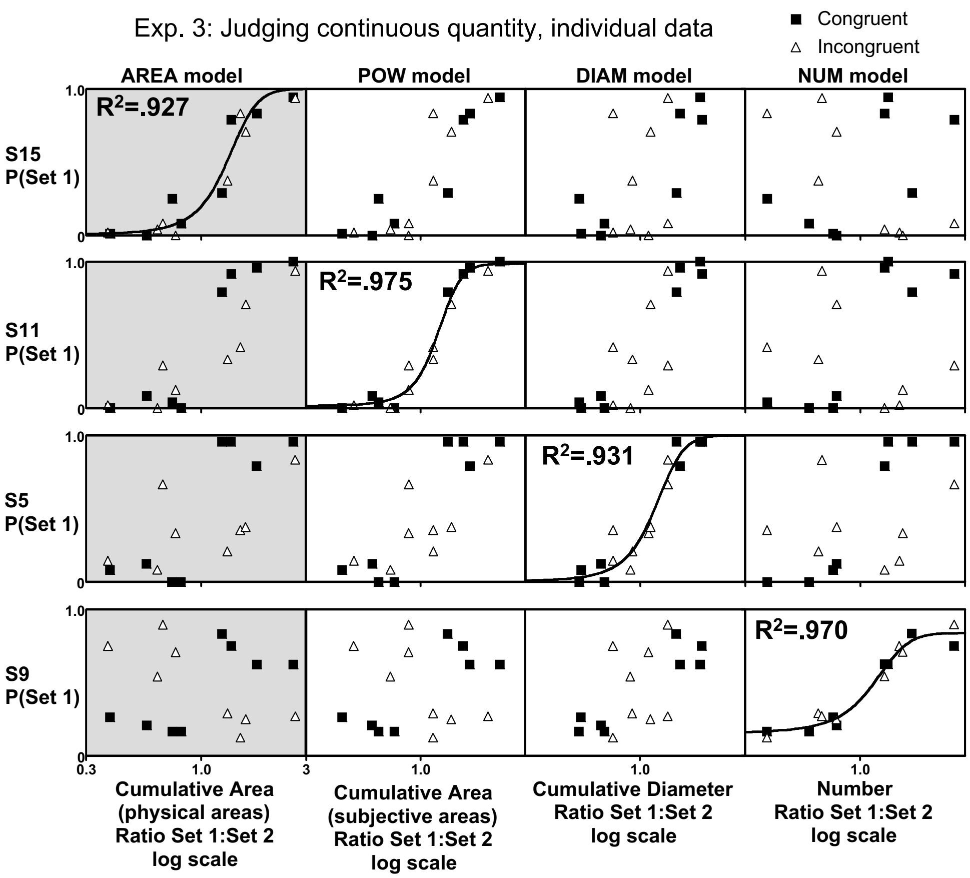

Fig. 3. Examples of models' fits to four individual participants' data from Experiment 3 (judging continuous quantity, with explicit instructions, in heterogeneous sets). Curve fits are shown for the best model for each participant (see Table 3 for best models for every participant).

these two trial types (*not* as a direct result of their congruency status). In contrast, a participant who can estimate individual disk areas quite accurately (one whose data are best fit by the AREA model) may exhibit little or no

RT difference between congruent and incongruent trials. This is because the two trial types were *closely equated* with respect to cumulative veridical area ratios. Participants whose data are best fit by the POW model should fall somewhere in between: these participants apparently did not have access to accurate estimates, but their underestimation of individual disk area was not as severe as that of the DIAM participants. For these participants, the incongruent trials were associated with cumulative subjective area ratios that were *somewhat* more difficult than those associated with the congruent trials. Therefore the illusory-Stroop hypothesis makes the following predictions about RT differences for congruent vs. incongruent trials: AREA participants should show little or no difference, POW participants should show some difference, and DIAM participants should show the greatest difference. This prediction was tested using RT data from the participants in Experiments 1 and 2 whose data had been best fit by these three models (nearly all participants in Experiment 3 were in the POW category, so their data cannot help to differentiate between the two hypotheses). Four participants fell into the AREA category, twelve fell into the POW category, and ten had been classified as DIAM participants. Average RT difference scores (Incongruent trial RT – Congruent trial RT) were 2 ms for the AREA group, 57 ms for the POW group, and 98 ms for the DIAM group, consistent with the predictions of the illusory-Stroop hypothesis. A one-way ANOVA performed on the individual participants' RT difference scores for these three groups confirmed the effect of model category (F(2,23) = 8.086, *p* < .005).

Overall, the first three experiments support the idea that there was no Stroop interference between mental representations of cumulative area and discrete number during these aggregate continuous quantity judgment tasks. The final experiment tested the possibility that a discrete quantity judgment task might produce evidence of Stroop interference effects from the irrelevant (continuous) dimension.

5. Experiment 4: judging number (discrete quantity)

Can adults make accurate judgments of discrete number and ignore conflicting continuous quantity information? This experiment used a procedure identical to that of Experiment 3 with a new set of participants, with one difference: the new task required choosing the set with the larger number of elements.

5.1. Methods

5.1.1. Participants

Seventeen adults participated for pay or for credit in the Harvard University Psychology Department Study Pool. All had normal or corrected-to-normal vision. All experimental procedures were approved by an Institutional Review Board and informed consent was obtained from all participants.

5.1.2. Stimuli and procedure

The stimulus set and procedure were taken from Experiment 3, but participants were explicitly instructed to choose the set with the larger number of elements.

5.2. Results and discussion

Participants chose the set with the larger number of items on 88% of the trials. Group means for all four model types are presented in Fig. 4. Unlike the data from the previous three experiments, these data demonstrate no difference between group performance levels for congruent and incongruent trials. Participants are apparently able to judge number and ignore continuous quantity successfully at every ratio, at least for the stimulus types and comparison difficulties tested here. It is possible that incongruent trial costs would be observed for more difficult numerical

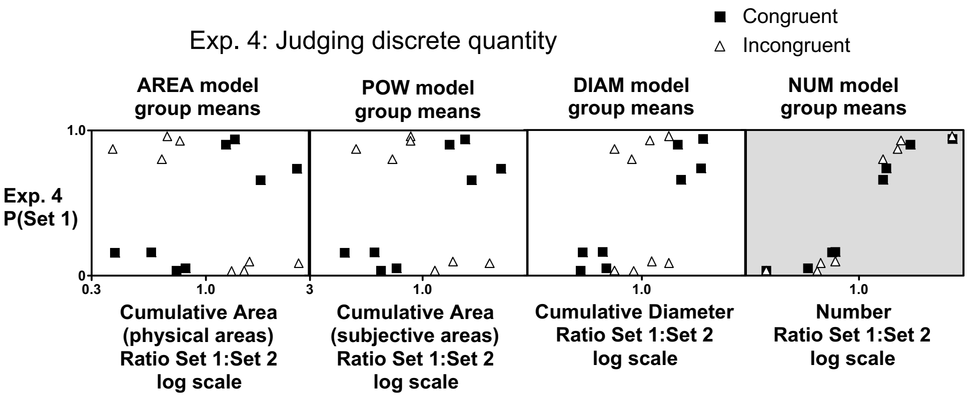

Fig. 4. Group data from the discrete quantity judgment tasks of Experiment 4 (compare to the continuous quantity judgment data in Fig. 2). A shaded graph indicates that those data are plotted against the comparison ratios that *should* determine performance on the task, according to the task instructions (number of items, in this Figure). Group means are plotted with respect to the four main types of models tested. The AREA column shows the ratios of the total *Cumulative Veridical Areas* of sets (the quantity that should be maximized in the task). The POW column depicts the data plotted with respect to ratios of total *Cumulative Subjective Areas* of the sets, assuming individual areas are underestimated in accord with previous psychophysical findings. The DIAM column depicts the data plotted with respect to *Cumulative Diameter* ratios. The NUM column depicts the data plotted with respect to ratios of the *Number* of items in the sets (the quantity that participants should attempt to maximize, according to task instructions). The data produced for the discrete quantity tasks were far less variable than the data produced for the continuous quantity task: see Table 4 for the best models of individual participants' performance.

Table 4
Experiment 4, best models for each participant as determined by AICc differences

	$AREA_1$	$AREA_2$	POW_1	POW_2	$DIAM_1$	$DIAM_2$	NUM_1	NUM_2	Best R^2
1							*	*****	See text
2							*****		.960
3							*	*****	See text
4							*****		.987
5							*****		.968
6							*****		.997
7							*****		.958
8							*****		.976
9							*****		.993
10								*****	See text
11							*****		.932
12							*****		.827
13							*****		.966
14							*****		.972
15							*****		.984
16							*****		.979
17							*****		.933

The leftmost column gives participant numbers and the next eight columns represent the eight models tested. Best model of the eight models tested, *****; other models with substantial empirical support relative to the best, **; other models with considerably less support, *; models that are out of the running or AICc difference >10 (essentially no empirical support), left blank. R-squared values are listed for the best supported model.

ratios. For fourteen of the seventeen participants, performance was best explained by the one-curve NUM model (suggesting that they compared the number of items in the first set to the number of items in the second as they were instructed[3]), and a single curve best explained performance patterns for both the congruent and incongruent trials (see Table 4). The average estimated Weber fraction for these fourteen participants' numerosity judgments was 0.17, comparable to estimates from previous studies. Individuals' Weber fraction estimates varied from 0.07 to 0.28. Fig. 5 shows examples of individual data from four of these fourteen participants.

For three of the seventeen participants, the data were best explained by a two-curve NUM model: there was a difference between the congruent and incongruent trials for these three participants. However, the direction of this difference is opposite to that predicted by the interference hypothesis: if cumulative area information were interfering with number judgments, we would expect to see better discrimination for congruent trials. In fact, these three participants' data show the opposite pattern: the functions fitted to their incongruent trial data have steeper slopes than those fitted to their congruent trials.

The four experiments differ in their findings, but they are consistent with the same general conclusion. Like Experiments 1 through 3, Experiment 4 provides evidence against the idea that total continuous amount and discrete quantity are automatically represented when we look at a set such that their representations compete for behavioral control leading to interference between quantitative dimensions. Experiments 1 through 3 demonstrated performance costs for incongruent trials during cumulative area judgments, but showed that these costs were not due to Stroop interference. Experiment 4 produced no performance costs for incongruent trials during numerical quantity judgments.

6. General discussion

This study found no evidence of Stroop interference between representations of discrete and continuous quantity in adults' comparative judgments with large sets of elements. Instead, performance patterns that are often interpreted as evidence of Stroop interference were shown to reflect an "illusory-Stroop" effect.

In the first three experiments, participants were asked to make comparative judgments of the cumulative areas of large sets of disks under a variety of experimental conditions. Performance costs were observed for the incongruent trials (when numerical quantity information conflicted with cumulative area information), but these costs were *not* best explained in terms of interference from the irrelevant dimension of number. Instead, the illusory-Stroop hypothesis provided a better explanation of the incongruent trial costs observed in these tasks. According to this hypothesis, participants judge cumulative area by combining estimates of individual element sizes such that individual areas are underestimated, in accord with a long history of psychophysical research. The hypothesis predicts that incongruent trials should produce performance costs relative to congruent trials (in which number and cumulative area are not in conflict), because if individual elements' areas are underestimated, trials classified as incongruent will simply require more difficult discriminations (see Fig. 1).

Participants in a fourth experiment were asked to make comparative numerical (discrete) quantity judgments, choosing the numerically larger set from two sets of disks. Here, no performance differences between congruent and incongruent trials were observed. Participants' numerical judgments were apparently not subject to interference from competing representations of cumulative area in this

[3] This experiment does not necessarily tell us about the mechanism used to create this representation of number; the data are consistent with multiple theories of the construction of number representations.

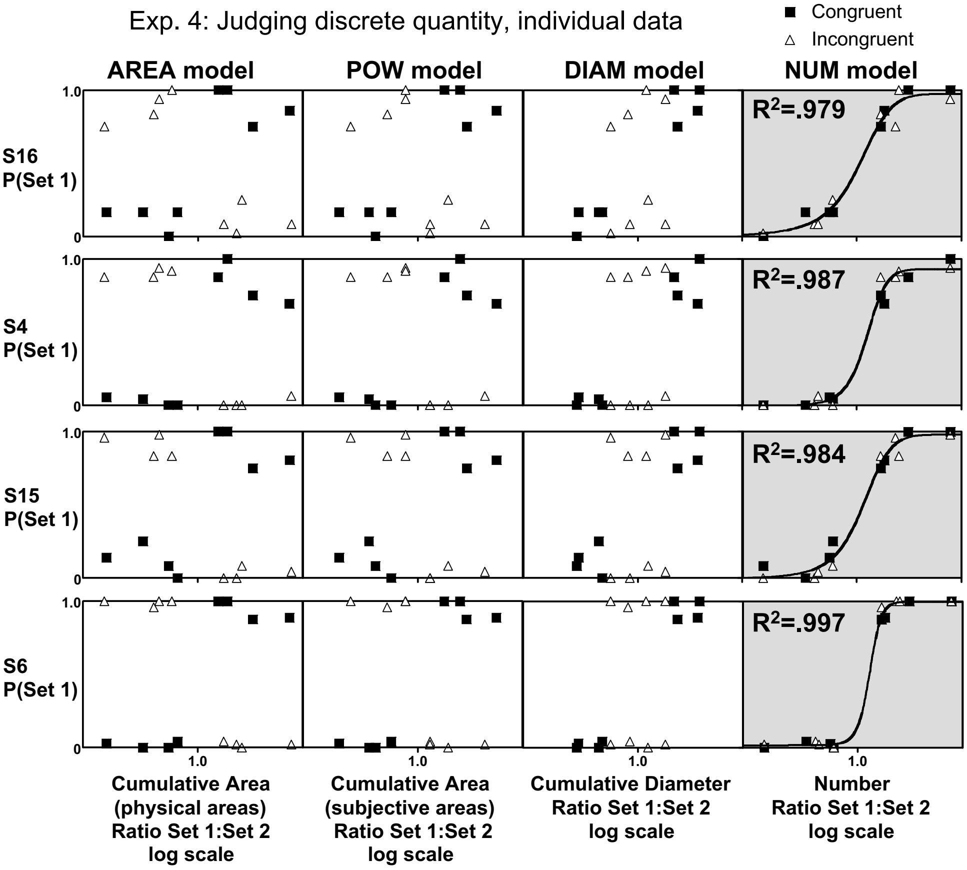

Fig. 5. Examples of models' fits to four individual participants' data from Experiment 4 (judging discrete quantity, explicit instructions, heterogeneous sets). The stimuli were the same as those used in Experiment 3; only the task instructions were different. Curve fits are shown for the best model for each participant (see Table 4).

task. Taken together, the findings from these four experiments show that, at least under the conditions tested here, there is no evidence of Stroop interference between representations of number and cumulative area. Though it is constructed like a Stroop task and produces data that mimic Stroop interference, this task does not function as an interference paradigm.

6.1. Relation to previous studies using similar tasks

Previous findings of performance costs for incongruent trials in similar studies of adults' discrete and continuous quantity judgments have been interpreted as evidence for the automatic extraction and representation of both number and cumulative area, such that these representations compete for behavioral control and produce Stroop interference on incongruent trials (Hurewitz et al., 2006). The present study offers an alternative explanation for incongruent trial performance costs, while providing evidence against Stroop interference effects (at least for the procedures used in this series of studies). What can this tell us about the conclusions drawn from previous studies using similar tasks?

One possibility is that the incongruent trial costs observed in previous studies are also instances of an illusory Stroop effect, and that a similar analysis of the data would cause apparent interference effects to disappear (as in the present study). This may be the case. The current cumulative area judgment tasks produced data compatible with the results obtained from previous studies: the studies differ fundamentally in their interpretation of the incongruent trial cost, but both studies did find an incongruent cost. If similar mechanisms do underlie participants' cumulative area judgments in both types of tasks, even greater illusory Stroop effects could have been produced in the previous study, because the ratios of cumulative areas for congruent vs. incongruent trials were not equated in the previous study as they were in the present study. This could amplify any illusory Stroop effects of the sort observed here, rendering incongruent trial discriminations even *more* difficult relative to congruent trials (again, provided that the same mechanisms are at work).

A second possibility is that Stroop interference effects *do* exist under the conditions of the previous study, though they do not exist for the task conditions reported here. The current experiments used large sets (sets of nine elements or more), and the previous study used smaller sets (sets of seven elements or fewer). It is certainly possible that in the smaller-set range, there are Stroop interference effects between quantity dimensions; different patterns of performance have been reported in quantity processing paradigms that employ small vs. large sets (e.g. Hauser & Spelke, 2004; Mandler & Shebo, 1982; Xu, 2003; but see Balakrishnan & Ashby, 1992; Cordes, Gelman, Gallistel, & Whalen, 2001; Izard, Dehaene-Lambertz, & Dehaene, 2008). There were other differences between this study and the previous study as well: the former used smaller disks and sets were presented sequentially, and the latter used larger disks and sets were presented simultaneously. It is possible that interference does exist under the conditions of the previous studies, and that these stimulus differences are responsible. Further experiments are needed to distinguish between these possibilities.

The present numerical judgment task did not produce a difference in performance between congruent and incongruent trials, so there was no opportunity to explore potential Stroop interference effects from irrelevant continuous quantity information when participants explicitly judged discrete quantity. This finding differed from the results of the previous study, in which incongruent trial performance costs were observed when participants made numerical judgments (Hurewitz et al., 2006). The two tasks may have produced different results because of differences in their stimuli and procedures; for example, the smaller sets used in the previous task or the simultaneous presentation of the sets may be responsible for the incongruent trial costs observed previously. The data reported here cannot speak to the question of whether the incongruent trial costs observed in the Hurewitz et al. task do indeed reflect a Stroop interference effect, as reported by the authors. Further experiments with similar stimuli will be necessary to explore the possibility that related types of illusory Stroop effects might explain incongruent trial costs in numerical judgments.

6.2. Constructing representations of cumulative area

These experiments do not provide explicit tests of the mechanisms underlying the construction of cumulative area representations, but they do suggest that, at least under the conditions tested here, adults do not simply ignore object boundaries and assess cumulative area through some perceptual summation process, perhaps adding up all of the appropriately-colored pixels in the scene. The models that were required to explain performance in these tasks involved the estimation of individual element size, suggesting that participants assessed cumulative area through some computation over individual size estimates.

The models did not specify the specific kinds of computations that might have been employed to combine individual size estimates. For example, a participant whose data best supported the DIAM model could have been summating all of the individual diameters in a set, or s/he could have been assessing the diameter of a single disk (or the mean diameter in the set) and multiplying by the total number of disks. Which type of mechanism is more likely to underlie these judgments? Are summation processes more likely, or is it more likely that participants construct representations of cumulative area by multiplying the mental magnitudes that represent number and individual element size? A growing body of work on statistical summary representations of sets suggests that items in sets may not be processed as individuals by the visual system; rather, statistical information about the set as a whole appears to be rapidly extracted (Ariely, 2001; Chong & Treisman, 2003; Chong & Treisman, 2005; Sussman & Scholl, accepted for publication; but see Myczek & Simons, 2008). It is possible that multiplicative models incorporating estimates of numerical quantity and mean element size are more parsimonious and more compatible with other findings; however, further experiments will be needed to distinguish between these possibilities quantitatively.

Whether representations of cumulative area are the result of a summation process or of a computation over estimates of element size and numerosity, it seems likely that the element size estimates that build these representations are not veridical. At least under the conditions of these experiments, only a very few participants produced data suggesting that they used accurate estimates of individual area to construct representations of cumulative area (these were the few participants whose data supported the AREA, or cumulative veridical area, model of performance – consistent with the idea that for these few participants, the Stevens exponent for disk area was 1.0). Most participants in the cumulative area tasks produced data suggesting they either underestimated individual element area according to a power law (with an exponent of 0.8, as suggested by previous research), or that they simply used diameter as an estimate of disk size.

6.3. Sources of individual difference

When participants were asked to make judgments of cumulative area, they rarely produced data that could be fit well by a model that assumed their choices were truly based on some quantity equivalent to physical cumulative area (the AREA model). Only six individual participants produced data that were fit best by such a model, and there was a great deal of individual difference in participants' data. The analyses used in these studies binned participants into categories determined by the performance model that best fit their data: AREA (corresponding to the AREA model described above), POW (for those whose data were best explained by a model assuming that subjective disk area was a power function of veridical disk area, with an exponent of 0.8 as suggested by previous literature), DIAM (for those whose comparisons were based on some quantity equivalent to cumulative diameter), and NUM (for those who appeared to choose the set with the larger number). In the three experiments involving cumulative area judgments, participants produced data that placed them in all four categories (though nearly all of the participants fell into the first three). Cumulative area judgments clearly produce individual differences in performance, but what is the best explanation of these differences?

It is possible that the categories used in this initial investigation are indeed accurate reflections of the perceptual and cognitive processes underlying these judgments. On this view, AREA participants had access to some computation that allowed them to accurately assess areas (perhaps a learned strategy), POW participants used a different computation, and DIAM participants really did perform a computation involving an estimate of diameter. If so, the differences in individual patterns of performance are due to true differences in the computations employed by these participants.

It is also possible, and perhaps more likely, that all of the AREA, POW, and DIAM participants performed the cumulative area task using the same underlying process, and that the categories chosen for this study do not reflect a true categorical difference in participants' judgments. The AREA model, for example, is equivalent to a model in which subjective individual disk area is a power function of veridical disk area with an exponent of 1. Note that the POW model used in this initial report of the illusory-Stroop phenomenon imposed an exponent of 0.8 (based on previous findings), but studies do report wide individual variation in power functions for two-dimensional area (e.g. Teghtsoonian, 1965). Future experiments may find that the AREA, POW, and DIAM categories should be subsumed under a more general POW model, in which the Stevens exponent is an additional parameter rather than being fixed at 0.8. The strongest test of this possibility would measure the exponents that govern estimates of disk area in each individual participant, allowing researchers to use the measured value of the Stevens exponent to guide the fit of this revised and more general power function model of cumulative area judgment.[4] If this speculative explanation is correct, the individual performance differences observed in the present tasks may simply reflect individual differences in Stevens exponents for the estimation of individual disk areas, rather than resulting from individual differences in the computation of cumulative area.

Nearly all participants in the explicit number judgment task were placed in the NUM category. These participants nearly always produced data that could be fit very well by the NUM model, suggesting that they did base their comparisons on the numbers of disks present in each set (or on some other equivalent quantity). Despite the fact that the number task required discrimination based on an abstract quantitative dimension, participants were clearly more accurate when judging numerical quantity (Experiment 4) than when judging cumulative area for identical stimuli (Experiment 3). This result is consistent with the recent finding that infants can detect smaller changes in numerical quantity than in cumulative area when presented with large sets of elements (Cordes & Brannon, 2008).

6.4. Potential relevance to controversies from the developmental literature

Although it is clearly speculative to make connections between these adult data and patterns of behavior observed in infants, the Stroop interference effects explored here have been explicitly proposed as a potential explanation for a pattern of inconsistent findings from the developmental literature: that infants exhibit sensitivity to discrete quantity under some conditions, while they appear sensitive only to continuous quantity under others (Cordes & Gelman, 2005; Hurewitz et al., 2006). The logic of the argument is as follows: if we form representations of cumulative area and discrete number whenever we look at a set, such that irrelevant quantitative information intrudes even upon adults' explicit judgments, we might reasonably expect infants to exhibit shifting patterns of behavior when confronted with sets under differing experimental conditions (Hurewitz et al., 2006).

The present findings suggest that claims of Stroop interference should be examined carefully before they are favored as an explanation of inconsistent infant behavior, though further research in adults and children will be needed in order to explore the relative explanatory strengths of the interference hypothesis and the illusory-Stroop hypothesis under a broader range of experimental conditions. If the illusory-Stroop hypothesis does turn out to apply to adults' quantity judgments with small sets as well as the large sets tested here, then there will be no compelling evidence for the existence of competition between representations of continuous and discrete quantity in adults. While we cannot assume that the same would hold true for other populations, a lack of evidence for competition and interference in adults would remove some of the motivation for favoring this account in infants. If, alternatively, the interference hypothesis does turn out to provide a good explanation of adults' performance in quantity judgments with small sets, then Stroop interference may be able to provide an explanation of the infant findings. Many of the most puzzling results from studies with infants involve their behavior when confronted with very small sets of objects: sometimes they appear to notice changes in numerical quantity, but under other conditions they ignore seemingly obvious numerical changes, responding only to changes in continuous amount (e.g. Clearfield & Mix, 1999; Feigenson et al., 2002; Feigenson et al., 2002; but see Cordes & Brannon, in press).

7. Conclusion

Recent adult behavioral studies have led to the suggestion that both cumulative area and numerical quantity are extracted automatically when we apprehend a set of elements (perhaps along with many other quantitative dimensions), and that their representations compete for behavioral control, such that quantitative judgments regarding one stimulus dimension are subject to interference from the irrelevant dimension (Hurewitz et al., 2006). The present study found no evidence of interference between number and cumulative area representations in

[4] Studies measuring individual participants' power functions for disk area may also allow for more nuanced analyses of RT data than the present studies could provide (see Exp. 3 Results & Discussion). The illusory-Stroop hypothesis predicts that the magnitude of the incongruent-congruent RT difference observed, in a task designed like this one, should be correlated with the Stevens exponent. Participants with an exponent near 1 should show little RT difference for congruent vs. incongruent trials (provided that cumulative area discrimination ratios are equated across trial types). Participants with smaller exponents should show greater RT differences, because the more extreme the underestimation of individual element area, the more difficult the discriminations will tend to be for incongruent trials relative to congruent trials.

adults, at least for the types of stimuli tested here. Though participants in the present cumulative area judgment tasks did produce the patterns of performance that are often interpreted in terms of Stroop interference (performance costs for "incongruent" trials, in which the task-irrelevant dimension, number, conflicted with the task-relevant dimension, area), these patterns were shown to be due to an illusory Stroop effect. The performance costs were *not* best explained in terms of interference from the irrelevant dimension of number. Instead, the data support the idea that participants judge cumulative area by combining estimates of individual element sizes such that individual areas are underestimated. This model of cumulative area judgment predicts that incongruent trials should produce a performance cost because if individual elements' areas are underestimated, trials classified as incongruent will simply require more difficult discriminations.

These experiments provide evidence against Stroop interference between representations of continuous and discrete quantity in adults' judgments of large sets, under the conditions tested here, and they demonstrate that adults' discriminations of cumulative area are more difficult than discriminations of number. They further provide support for a particular class of models of the construction of representations of cumulative area, in which cumulative area is assessed through a computation that combines inaccurate estimates of individual element area.

Acknowledgments

The author thanks Barbara Juhasz, Andrea Patalano, and Elizabeth Spelke for comments on an earlier version of this manuscript, Elizabeth Spelke for the use of her laboratory facilities for data collection, and three anonymous reviewers for their helpful input. This work was supported by NSF grant REC-0087721 to E. Spelke, a Spencer Foundation Postdoctoral Fellowship to H.B., and a Project Grant from Wesleyan University. Address correspondence to Hilary Barth (hbarth@wesleyan.edu) at the Department of Psychology, Wesleyan University, 207 High St, Middletown CT 06459.

Appendix

A sigmoidal function in the form of a four-parameter logistic equation was fitted to individual participants' data such that upper and lower asymptote parameters, in addition to slope and position parameters, were estimated rather than being fixed in order to avoid bias in slope and position estimates caused by observers' lapses (Wichmann & Hill, 2001). Analyses were also performed with asymptotes fixed at 0 and 1, with the same results for most participants; however, some participants' data could not be fitted well in this manner (e.g. Fig. 3, S9). The equation was Y = LA + (UA − LA)/(1 + 10^((α − CR)*β)), where α is the position parameter, β is the slope parameter, LA is the lower asymptote, UA is the upper asymptote, and CR is the Comparison Ratio (the basis for comparison under each model tested). The AREA model assumes that the basis for comparison (CR) is equivalent to $(\pi r_1^2 N_1)/(\pi r_2^2 N_2)$, where r_1 and r_2 are the fixed disk radii (in Experiments 1 and 2) or average disk radii (in Experiments 3 and 4) in sets 1 and 2, respectively, and N_1 and N_2 are the number of disks in each set. The POW model assumes that the basis for comparison is $((\pi r_1^2)^{0.8}(N_1))/((\pi r_2^2)^{0.8}(N_2))$, where r_1 and r_2 are the fixed disk radii (in Experiments 1 and 2) or average disk radii (in Experiments 3 and 4) in sets 1 and 2, respectively, and N_1 and N_2 are the number of disks in each set. The DIAM model assumes that the basis for comparison is equivalent to $(r_1N_1)/(r_2N_2)$, where r_1 and r_2 are the fixed disk radii (in Experiments 1 and 2) or average disk radii (in Experiments 3 and 4) in sets 1 and 2, respectively, and N_1 and N_2 are the number of disks in each set. The NUM model assumes that the basis for comparison is equivalent to N_1/N_2, where N_1 and N_2 are the number of disks in each set.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Académiai Kiadó.

Algom, D., Dekel, A., & Pansky, A. (1996). The perception of number from the separability of the stimulus: The Stroop effect revisited. *Memory and Cognition, 24*, 557–572.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*, 157–162.

Balakrishnan, J. D., & Ashby, F. G. (1992). Subitizing: Magical numbers or mere superstition. *Psychological Research, 50*, 555–564.

Balci, F., & Gallistel, C. R. (2006). Cross-domain transfer of quantitative discriminations: Is it all a matter of proportion? *Psychonomic Bulletin and Review, 13*, 636–642.

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition, 86*, 201–221.

Barth, H., La Mont, K., Lipton, J., & Spelke, E. (2005). Abstract number and arithmetic in preschool children. *Proceedings of the National Academy of Sciences, 102*, 14116–14121.

Besner, D., & Coltheart, M. (1979). Ideographic and alphabetic processing in skilled reading of English. *Neuropsychologia, 17*, 467–472.

Brannon, E. M. (2006). The representation of numerical magnitude. *Current Opinion in Neurobiology, 16*, 222–229.

Brannon, E. M., Abbott, S., & Lutz, D. L. (2004). Number bias for the discrimination of large visual sets in infancy. *Cognition, 93*, B59–B68.

Brannon, E. M., Lutz, D., & Cordes, S. (2006). The development of area discrimination and its implications for number representation in infancy. *Developmental Science, 9*, F59–F64.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference. A practical information-theoretic approach* (2nd ed). New York: Springer.

Cantlon, J., & Brannon, E. M. (2006). Shared system for ordering small and large numbers in monkeys and humans. *Psychological Science, 17*, 401–406.

Carey, S. (2004). Bootstrapping and the origin of concepts. *Daedalus*, 59–68.

Carey, S., & Sarnecka, B. W. (2006). The development of human conceptual representations. In M. Johnson & Y. Munakata (Eds.), *Processes of change in brain and cognitive development: Attention and performance XXI* (pp. 473–496). New York: Oxford University Press.

Chong, S., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*, 393–404.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*, 891–900.

Clearfield, M. W., & Mix, K. S. (1999). Number vs. contour length in infants' discrimination of small visual sets.. *Psychological Science, 10*, 408–411.

Cordes, S., & Brannon, E. M. (in press). The relative salience of discrete and continuous quantities in infants. *Developmental Science*.

Cordes, S., & Brannon, E. M. (2008). The difficulties of representing continuous extent in infancy: Using number is just easier. *Child Development, 79*, 476–489.

Cordes, S., & Gelman, R. (2005). The young numerical mind: When does it count? In J. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 127–142). New York: Psychology Press.

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychological Bulletin and Review, 8*, 698–707.

Dehaene, S. (1997). *The number sense*. New York: Oxford University Press.

Feigenson, L. (2007). The equality of quantity. *Trends in Cognitive Sciences*. doi:10.1016/j.tics.2007.01.006.

Feigenson, L., Carey, S., & Hauser, M. D. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychological Science, 13*, 150–156.

Feigenson, L., Carey, S., & Spelke, E. S. (2002). Infants' discrimination of number vs. continuous extent. *Cognitive Psychology, 44*, 33–66.

Feigenson, L., Dehaene, S., & Spelke, E. S. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*, 307–314.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44*, 43–74.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences, 4*, 59–65.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin and Review, 11*, 791–806.

Hauser, M. D., & Spelke, E. S. (2004). Evolutionary and developmental foundations of human knowledge: A case study of mathematics. In M. Gazzaniga (Ed.), *The cognitive neurosciences III*. Cambridge: MIT Press.

Hauser, M. D., Tsao, F., Garcia, P., & Spelke, E. S. (2003). Evolutionary foundations of number: Spontaneous representations of numerical magnitudes by cotton-top tamarins. *Proceedings of the Royal Society London B, 270*, 1441–1446.

Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory and Cognition, 10*, 389–395.

Hurewitz, F., Gelman, R., & Schnitzer, B. (2006). Sometimes area counts more than number. *Proceedings of the National Academy of Sciences, 103*, 19599–19604.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297–307.

Izard, V., Dehaene-Lambertz, G., & Dehaene, S. (2008). Distinct cerebral pathways for object identity and number in human infants. *PloS Biology, 6*, 275–285.

Krider, R. E., Raghubir, P., & Krishna, A. (2001). Pizzas: ∏ or square? Psychophysical biases in area comparisons. *Marketing Science, 20*, 405–425.

Lipton, J. S., & Spelke, E. S. (2003). Origins of number sense: Large-number discrimination in human infants. *Psychological Science, 14*, 396–401.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General, 11*, 1–22.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes, 9*, 320–334.

Mix, K. S., Huttenlocher, J., & Levine, S. C. (2002). Multiple cues for quantification in infancy: Is number one of them? *Psychological Bulletin, 128*, 278–294.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception and Psychophysics, 70*, 772–788.

Newcombe, N. (2002). The nativist-empiricist controversy in the context of recent research on spatial and quantitative development. *Psychological Science, 13*, 395–401.

Nieder, A. (2005). Counting on numbers: The neurobiology of numerical competence. *Nature Reviews Neuroscience, 6*, 177–190.

Sussman, R. S., & Scholl, B. J. (accepted for publication). Finding the mean: The flexibility and limitations of visual statistical processing. *Perception and Psychophysics*.

Teghtsoonian, M. (1965). The judgment of size. *American Journal of Psychology, 78*, 392–402.

vanMarle, K., & Wynn, K. (2006). Six-month-old infants use analog magnitudes to represent duration. *Developmental Science, 9*, F41–F49.

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

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception and Psychophysics, 63*, 1293–1313.

Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition, 89*, B15–B25.

Xu, F., & Spelke, E. S. (2000). Large number discrimination by human infants. *Cognition, 74*, B1–B11.