Journal of Numerical Cognition
jnc.psychopen.eu | 2363-8761

PsychOpen
publishing psychology

Research Reports

A Drift Diffusion Model Account of the Semantic Congruity Effect in a Classification Paradigm

Angelo Pirrone*[ab], James A. R. Marshall[b], Tom Stafford[a]

[a] Department of Psychology, The University of Sheffield, Sheffield, United Kingdom. **[b]** Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom.

Abstract

The semantic congruity effect refers to the facilitation of judgements (i) when the direction of the comparison of two items coincides with the relative position of the items along the dimension comparison or (ii) when the relative size of a standard and a target stimulus coincides. For example, people are faster in judging 'which is bigger?' for two large items, than judging 'which is smaller?' for two large items (selection paradigm). Also, people are faster in judging a target stimulus as smaller when compared to a small standard, than when compared to a large standard, and vice versa (classification paradigm). We use the Drift Diffusion Model (DDM) to explain the time course of a semantic congruity effect in a classification paradigm. Formal modelling of semantic congruity allows the time course of the decision process to be described, using an established model of decision making. Moreover, although there have been attempts to explain the semantic congruity effect within evidence accumulation models, two possible accounts for the congruity effect have been proposed but their specific predictions have not been compared directly, using a model that could quantitatively account for both; a shift in the starting point of evidence accumulation or a change in the rate at which evidence is accumulated. With our computational investigation we provide evidence for the latter, while controlling for other possible explanations such as a variation in non-decision time or boundary separation, that have not been taken into account in the explanation of this phenomenon.

Keywords: drift diffusion model, semantic congruity effect, decision making, magnitude comparison

Journal of Numerical Cognition, 2017, Vol. 3(1), 77–96, doi:10.5964/jnc.v3i1.79

Received: 2016-10-26. Accepted: 2017-03-02. Published (VoR): 2017-07-21.

*Corresponding author at: Department of Psychology, The University of Sheffield, Floor D, Cathedral Court, 1 Vicar Lane, Sheffield, S1 2LT, United Kingdom. E-mail: apirrone1@sheffield.ac.uk

When subjects are required to judge two stimuli that differ on a single contrastive polar continuum (e.g., 'big' vs. 'small'), subjects are faster to judge which of the two stimuli is higher on that continuum, when the stimuli are high on that particular dimension, and they are faster to judge which of the two stimuli is lower on that continuum, when the stimuli are low on that particular dimension. Furthermore, when subjects are required to judge whether a target stimulus is bigger or smaller than a standard stimulus, subjects are faster when the relative size of the standard and of the target coincides (see Dehaene, 1989). Dehaene (1989) defined the first paradigm (i.e., chose the bigger/smaller of two stimuli) as a selection paradigm, and the second paradigm (i.e., is the target bigger/smaller than a standard) as a classification paradigm. The result that characterises these two paradigms is referred to as the *semantic congruity* effect. The semantic congruity effect has been replicated in perceptual and symbolic judgements across different domains, including surface area (Moyer & Bayer, 1976), line length (Petrusic, Baranski, & Kennedy, 1998), brightness (Wallis & Audley, 1964), scalar adjectives of quality (Holyoak & Mah, 1982), the distance between two cities (Holyoak & Mah, 1982) and Arabic numerals (Banks, Fujii, & Kayra-Stuart, 1976; Holyoak, 1978).

Many theories have been proposed to account for the semantic congruity effect. These theories vary greatly in the level of description of the phenomenon, with some theories being able to account for semantic congruity effects only in the case in which comparative instructions are presented to the subject (selection paradigm), but not when subjects have to decide whether a target is bigger or smaller than a standard stimulus (classification paradigm). For a detailed and exhaustive review of the models proposed for the explanation of the semantic congruity effect, refer to Petrusic (1992) and Leth-Steensen and Marley (2000); here we present a brief description of some of the theories that have been proposed for the explanation of this phenomenon.

According to the expectancy effect (Banks & Flora, 1977; Marschark & Paivio, 1979), the direction of the comparison (e.g., is the target stimulus bigger than the standard?) prepares the subject for the range of stimuli that will be presented. This results in a facilitation in case of congruency between the comparison and the stimuli. However, even when the comparative is presented together or after the presentation of the stimuli, the semantic congruity effect can still be observed (Holyoak & Mah, 1982), undermining a basic assumption of this model. Alternatively, the semantic coding model (Banks, Clark, & Lucy, 1975; Banks et al., 1976) explains the congruity effect by referring to linguistic codes; however, this struggles with the finding that even non-human primates show a semantic congruity effect when comparing magnitudes (Cantlon & Brannon, 2005).

A further verbal theory, the frequency explanation (Ryalls, Winslow, & Smith, 1998), explains the semantic congruity effect by the fact that each comparative is associated with one unique dimension during learning (i.e., subjects learn to use 'bigger' for high magnitude stimuli, and 'smaller' for low magnitude stimuli); yet, this explanation struggles with the result that the expectancy effect is found also when subjects are taught new comparisons with novel comparatives (Chen, Lu, & Holyoak, 2014). A further class of models are reference point models (Chen et al., 2014; Dehaene, 1989; Holyoak, 1978; Holyoak & Mah, 1982; Marks, 1972), according to which, subjects, when making a magnitude judgement, compare the numerical value of the stimulus with reference values stored in memory. Under this view, the subject is assumed to establish a reference point near one of the extreme values encountered in a given context and this results in a facilitation when the stimulus to discriminate is nearer to the reference point. From this perspective, the use of reference points has been suggested to affect the strength of evidence accumulation (see Chen et al., 2014; Dehaene, 1989); meaning, for example, that when the magnitude of the standard stimulus coincides with the magnitude of the target, this results in higher rates of evidence accumulation, compared to when there is not congruency between the relative sizes of the two stimuli. Other authors have explained the semantic congruity effect adopting random walk models (Birnbaum & Jou, 1990; Link, 1990; Link & Heath, 1975; Poltrock, 1989); these studies explain the semantic congruity effect as arising from a starting point adjustment dictated by the instructions. However, as argued in Leth-Steensen and Marley (2000), in tasks in which subjects are presented with symmetric differences (i.e., the same number of bigger and smaller comparisons are presented), it is not clear why subjects should adjust their starting point of evidence accumulation towards one of the two alternatives in selection paradigms. Finally, some evidence-accumulation models and instructional pathway interference accounts have been proposed (Leth-Steensen & Marley, 2000; Petrusic, 1992; Petrusic, Shaki, & Leth-Steensen, 2008), according to which the semantic congruity is due to a variation in the rate of evidence accumulation in case of congruency/incongruency between the instructions and the relative size of the stimulus pair.

Here, we focus on a computational model of decision making, known as the Drift Diffusion Model (Ratcliff & McKoon, 2008). This computational model has been applied to an impressive variety of tasks, paradigms and domains, including perceptual decision making, value-based decision making and also the description of the integration of sensory signals towards a motion-discrimination decision in monkeys (Gold & Shadlen, 2002; Ratcliff, 1978, 2002; Ratcliff & McKoon, 1988, 2008; Ratcliff & Rouder, 1998; Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Van Zandt, & McKoon, 1999; Shadlen & Newsome, 2001; Thapar, Ratcliff, & McKoon, 2003; Voss, Rothermund, & Voss, 2004).

In the DDM the decision maker integrates difference in evidence supporting two alternatives until a certain positive or negative threshold is crossed, and a decision is made in favour of that alternative.

In its simplest formulation, defined as 'the *reduced* version', the DDM is the continuous case of a random walk process (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006) and it is described by the following equation

$$dx = \mu dt + \theta dW,\ x(0) = 0 \tag{1}$$

where dx is the increment in evidence in a small time window dt, μ denotes the mean increase in evidence per unit time, and θdW denotes Gaussian white noise with mean zero and variance $\theta^2 dt$.

Interestingly, the DDM - in its *reduced* version - implements the Sequential Probability Ratio Test (Wald, 1947; Wald & Wolfowitz, 1948), which is the procedure that gives the shortest decision time given a fixed error rate in a two-alternatives forced-choice task (Bogacz et al., 2006). It is possible to demonstrate (Bogacz et al., 2006) that as discrete samples are taken more frequently and one approaches continuous-time sampling of a variable, the SPRT converges to Equation 1. In this way, the DDM is statistically optimal for stationary distributions of evidence in conditions in which the subject has to manage a speed-accuracy trade-off (Bogacz et al., 2006). Given this feature of the model, the DDM not only represents a descriptive model of decision making, but has been proposed also as a normative model (Basten, Biele, Heekeren, & Fiebach, 2010; Wang, 2013) towards which, under the influence of natural selection, the decision maker may be supposed to have evolved (but see Pirrone, Stafford, & Marshall, 2014).

A further reason for the popularity of the DDM is that, as shown by Bogacz et al. (2006), other prominent models of choice, under specific parametrization implement or approximate the DDM, with the exclusion of race models (Vickers, 1970) - models with one accumulator for each alternative that accumulate evidence but do not inhibit each other.

Although there are numerous variants of the DDM, here we focus in particular on the version of the DDM as formalised in Ratcliff and McKoon (2008), a more refined and psychologically plausible version of the *reduced* DDM.

A DDM process (Figure 1) is determined by seven parameters (Ratcliff & McKoon, 2008; Vandekerckhove & Tuerlinckx, 2007).

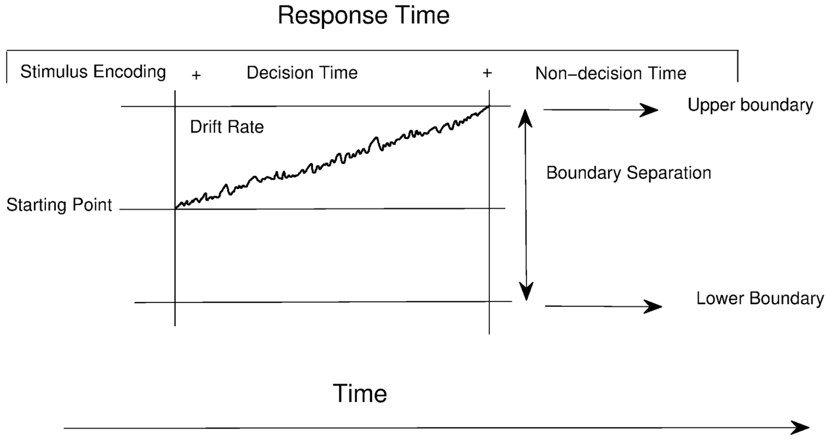

Figure 1. Graphical representation of the DDM.

The first, denoted by *a*, is the boundary separation and it captures the distance between the two thresholds for a decision. When *a* is small the decision is faster but less accurate since, given noisy fluctuations in the accumulation of evidence, it is more likely to end up at the wrong boundary; when *a* is large the decision is slower and more accurate. An interpretation for this parameter is therefore the trade-off between speed and accuracy for a decision. Second, is the starting point of evidence accumulation, denoted by *z*. An interpretation for this parameter is the bias for either response; if *z* is not equidistant from the boundaries but nearer to the one of the two limits, the subject will be 'biased' to make the choice corresponding to the nearer boundary; when the accumulation of evidence starts at *a/2* the process is unbiased. In the case of a biased process, fast reaction times (RTs) towards the nearer boundary and slow RTs towards the opposite boundary are predicted, given that the distance from the decision boundary is small in one case and large in the other. Third is the inter-trial variability of *z*, defined as s_z. Fourth is the drift rate, denoted as *v*, which represents the mean rate at which information is accumulated over time. This parameter can be interpreted as the quality of the stimulus and the amount of information carried by it for the perceiver. Experimental conditions for which the correct decision is 'easy' will have a higher drift rate compared to more difficult conditions. Also, a further interpretation of this parameter is the sensitivity of a subject towards a stimulus. The accumulation of information varies according to the drift rate and to a fifth parameter, the inter-trial variability in drift rate, denoted by *eta*. This parameter can be interpreted as the variability in attention or motivation of the decision maker or, in the case of changing stimuli, it can be thought of as the variability in stimulus quality. The last two parameters of the DDM refer to the non-decisions time, since the decision maker has to encode the stimulus and execute the motor response when making a decision. The non-decision component of a RT is denoted by *ter* and its inter-trial variability is defined as s_t.

It is interesting to note that the DDM can account for the full range of correct and incorrect RTs and for the probability of correct and wrong answers. Additionally, the DDM offers several advantages in terms of the relation between model parameters, experimental design, and wider theoretical interpretation. The main parameters of the DDM have clear interpretations in terms of psychological processing (e.g., the speed-accuracy trade-off is reflected in the separation of the decision thresholds). Model fitting using the DDM tends to reveal single parameters changing their values to track changes across experimental conditions. Inter-related to both of these, the intuitive nature of some aspects of DDM function means that changes to experimental design can often produce clear predictions in terms of DDM parameter change.

Relatively few studies have applied the DDM to questions of numeracy judgement (Park & Starns, 2015; Ratcliff, 2006; Ratcliff, Thompson, & McKoon, 2015). However, these examples show the benefits of a DDM decomposition of data in this field. For example, in Ratcliff et al. (2015), through a series of four numerosity experiments, authors have found that accuracy is largely dependant upon drift rate while RTs are determined by threshold settings. The values of drift rate and boundary separation were correlated across tasks but interestingly, across subjects, these two parameters were not correlated. With four further experiments in which speed and accuracy instructions were manipulated, the authors replicated the results of accuracy-drift correlation, RTs-boundary correlation and the consistency across tasks, however between-subjects differences were maintained even when the internal response criterion of subjects was manipulated. This result shows the benefit of a computational decomposition of data and lays the foundation for the understanding of the contrasting results regarding presence/absence of correlation between RTs and accuracy in numeracy judgements. A second important application of the DDM to numeracy judgements comes from Park and Starns (2015). In Park and Starns (2015), authors were interested in acuity measures of the approximate number system - the cognitive system that allows to estimate numerosity non-linguistically. Traditionally, measures of acuity of the approximate number system only involve accuracy. However, using the DDM, the authors show that measures of acuity only based on accuracy cannot account for speed-accuracy trade-offs confounds that do affect acuity measurements. This means that traditional measures of acuity are likely to be inaccurate, since they are contaminated by speed-accuracy trade-off confounds. Furthermore, the authors found that drift rate is a better predictor of symbolic mathematical ability compared to previously proposed measures.

Comparing directly the semantic congruity effect theories described above is out of the scope of this work, since some of them are not framed within the evidence accumulation framework. Here, we bring the semantic congruity effect within the same framework as many other decision phenomena; we use the Drift Diffusion Model (Ratcliff & McKoon, 2008) and show how it can account for the semantic congruity effect, by fitting it to behavioural data from a magnitude comparison experiment conducted with human subjects. Since the semantic congruity effect manifests in changes in decision time, the use of the DDM, which explicitly considers the time course of decision-making, is natural. In contrast, some of the heuristic proposals outlined above lack such formal description of how decisions evolve over time, or when they specify how the decision evolves, they do so by adopting ad-hoc models that only make predictions for the specific task but cannot be generalised to other tasks or domains. A unifying framework such as the DDM overcome the limitations of task-specific models. Furthermore, with a DDM decomposition we can investigate which decision parameters account for the semantic congruity effect. Together with the explanations proposed (i.e., drift rate or starting point) other parameters that have never been taken into account, such as non-decision time or boundary separation, could play a role in the semantic congruity effect. For example, the non-decision time, which has never been taken into account in the previous literature, could as well contribute to a semantic congruity effect given that the congruency/incongruency between the magnitude of the stimuli (or between the instructions and the relative sizes of the target and standard stimulus) could affect the motor response of the subjects.

Usually, in two-alternative forced choice tasks, parameters such as the starting point of evidence accumulation or the boundary separation are assumed to take time to change and are assumed to be set before the stimulus appears (Bogacz et al., 2006); here, however, we assume that the size of the standard, to which subjects pay attention at first during the trial presentation, is apprehended quickly, and it affects the decision process. In the literature similar mechanisms that affects the early stages of a decision are described; for example, Provost and Heathcote (2015) provided a similar explanation for a mental rotation task, and in their computational investigation they found that participants adjusted their boundary separation on the basis of a property of the stimulus, rotation angle. Also, it should be noted that typically in the kind of tasks in which the DDM is used, subjects evaluate one single stimulus; in this case a change in decision parameters cannot be contingent on the outcome of the decision. However, in our case we have that one feature of the stimulus, the size of the standard stimulus, to which subjects pay attention at first, can affect the subsequent discrimination of the target.

In our experiment, participants had to decide whether a target stimulus was smaller or bigger than a standard array, hence ours is a classification paradigm. A stimulus example is reported in Figure 2. Our experiment presents some differences with semantic congruity tasks in which the direction of the comparison is explicitly given. However, with our task we elicit a semantic congruity effect similarly to what done before by other authors (e.g., Dehaene, 1989; Link, 1990; Mewhort, Smith, & Kohly, 1996).

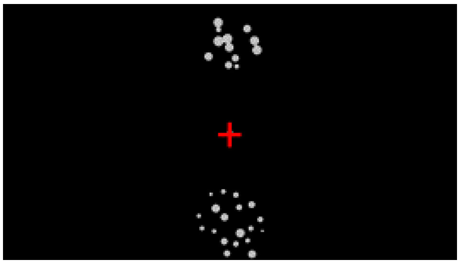

Figure 2. Stimulus example (stimuli are exaggerated for visibility). In each trial subjects had to decide whether the array presented on bottom (target) was smaller or bigger in numerosity than the array presented on top (standard). After their response, subjects were presented with a fixation cross that over the course of 600 ms was varying in size, as a warning signal to maintain fixation at the centre of the screen.

Experiment

Participants

Four right-handed subjects, one male, mean age = 20.5 years (*SD* = 3.2) with normal or corrected-to-normal vision participated voluntarily in the experiment in exchange of credits for course requirements. Each participant was tested in four sixty-minutes sessions on different days. The experiment was approved by the University of Sheffield, Department of Psychology Ethics Sub-Committee, and carried out in accordance with the University and British Psychological Society ethics guidelines and subjects gave their informed consent before performing it.

Materials

The experiments were programmed in Matlab, using the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). We used a modification of an established perceptual decision task (Gertner, Arend, & Henik, 2012; Piazza et al., 2010; Piazza, Fumarola, Chinello, & Melcher, 2011; Revkin, Piazza, Izard, Cohen, & Dehaene, 2008; Revkin, Piazza, Izard, Zamarian, et al., 2008) and a type of 'congruity' task similar to that used by Link (1990) and Dehaene (1989) - similar since also in our case subjects decided whether a target stimulus was bigger or smaller than a standard stimulus, however Link (1990) and Dehaene (1989) used two-digit numbers in their experiments. In our task, participants judged if a cluster of dots presented on the bottom of a laptop screen was 'smaller' or 'bigger' in numerosity than one presented on the top of the screen without counting and responding by button press.

Table 1

Stimuli Values for Each Condition

Condition	N of Dots	Ratio	Magnitude of Standard	Target (compared to standard) is
1	12 vs 5	0.42	small	smaller
2	12 vs 6	0.50	small	smaller
3	12 vs 7	0.58	small	smaller
4	12 vs 8	0.66	small	smaller
5	12 vs 9	0.75	small	smaller
6	12 vs 10	0.83	small	smaller
7	12 vs 11	0.91	small	smaller
8	12 vs 19	0.42	small	bigger
9	12 vs 18	0.50	small	bigger
10	12 vs 17	0.58	small	bigger
11	12 vs 16	0.66	small	bigger
12	12 vs 15	0.75	small	bigger
13	12 vs 14	0.83	small	bigger
14	12 vs 13	0.91	small	bigger
15	24 vs 10	0.42	medium	smaller
16	24 vs 12	0.50	medium	smaller
17	24 vs 14	0.58	medium	smaller
18	24 vs 16	0.66	medium	smaller
19	24 vs 18	0.75	medium	smaller
20	24 vs 20	0.83	medium	smaller
21	24 vs 22	0.91	medium	smaller
22	24 vs 38	0.42	medium	bigger
23	24 vs 36	0.50	medium	bigger
24	24 vs 34	0.58	medium	bigger
25	24 vs 32	0.66	medium	bigger
26	24 vs 30	0.75	medium	bigger
27	24 vs 28	0.83	medium	bigger
28	24 vs 26	0.91	medium	bigger
29	36 vs 15	0.42	big	smaller
30	36 vs 18	0.50	big	smaller
31	36 vs 21	0.58	big	smaller
32	36 vs 24	0.66	big	smaller
33	36 vs 27	0.75	big	smaller
34	36 vs 30	0.83	big	smaller
35	36 vs 33	0.91	big	smaller
36	36 vs 57	0.42	big	bigger
37	36 vs 54	0.50	big	bigger
38	36 vs 51	0.58	big	bigger
39	36 vs 48	0.66	big	bigger
40	36 vs 45	0.75	big	bigger
41	36 vs 41	0.83	big	bigger
42	36 vs 39	0.91	big	bigger

On each trial, one array - the standard - contained a fixed numerosity (12 dots for one third of the trials, 24 dots for one third of the trials, 36 dots for the other third), and the other array - the target - contained a varying numerosity that was smaller or bigger than the fixed numerosity by one of seven possible ratios. The ratio defined the difficulty of the judgement, with ratios closer to 1 being harder. The seven ratios, in order of increasing difficulty, were 0.42, 0.50, 0.58, 0.66, 0.77, 0.83, 0.91. The absolute number of dots in each choice pair and a description of conditions is shown in Table 1.

There were in total 42 conditions; seven increasing ratios (i.e., increasing difficulty) for each of three levels of standard stimulus magnitude (small, medium and big) for each type of response 'smaller' or 'bigger' (i.e., half of the times the target stimulus was bigger/smaller than the standard). For each trial, subjects had to decide whether the target stimulus was smaller or bigger than the standard stimulus by pressing 'left' or 'right' on the keyboard. Conditions were chosen so that for each standard stimulus we would have accuracy levels that range from floor to ceiling on the basis of the results of previous pilot studies.

To avoid participants relying upon continuous quantities associated with numerosity (i.e., dot size and envelope area), in this experiment the dot arrays were generated following the method and the MATLAB code provided by Gebuis and Reynvoet (2012). This method was used to produce four sets of images with all possible combinations of correlation (positive vs. negative) between the two features of the stimuli (envelope area, dot size) and dot number.

Procedure

During the whole experiment, subjects had to put their head on a chin rest at a viewing distance of 57 cm from the screen of a 14-inch laptop monitor (Dell Latitude E5430) with a refresh rate of 60 Hz. Subjects were required to fixate a red cross at the centre of the screen. The two dot arrays were presented simultaneously on the screen at ± 4.25 degrees of visual angle from the fixation cross, and participants were asked to judge if the cluster presented on the bottom of the screen was bigger or smaller than the one presented on top by pressing 'left' or 'right' on a keyboard. Each dot was randomly assigned an item size ranging between 0.08 and 0.59 degrees of visual angle. If subjects answered below 300 ms or above 3000 ms the sentence 'Too fast!' or 'Too slow!' was displayed on the screen. After giving a response, subjects were presented with a fixation cross that over the course of 600 ms was varying in size (i.e., small and then bigger for two times), as a warning signal for subjects to pay attention to the centre of the screen, and after subjects were presented with a new trial. Trials were presented in random order - within each day of the experiment - and participants performed 50 trials per condition after a training phase to familiarize them with the task which involved 1 trial per condition. Subjects participated in 4 different sessions on 4 different days (within a week from the first session), for a total of 200 trials per condition and 8400 trials for the whole experiment.

Results

Behavioural Results

Figure 3 shows the psychometric functions averaged across subjects. This figure shows the imbalance in response probability due to the semantic congruity effect. The probability of answering bigger increases with the standard, although the same ratios across conditions are maintained; when the magnitude of the standard was small the probability of answering 'bigger' was lower compared to when the magnitude of the standard was big.

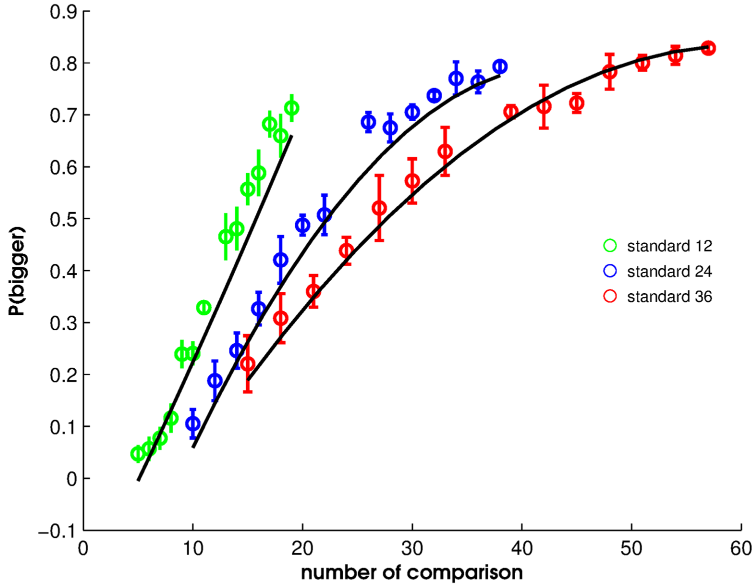

Figure 3. Psychometric functions showing the probability of answering 'bigger' as a function of the numerosity of the target and of the standard stimulus. On the x-axis is reported the number of dots of the target stimulus. Error bars are standard errors of the mean.

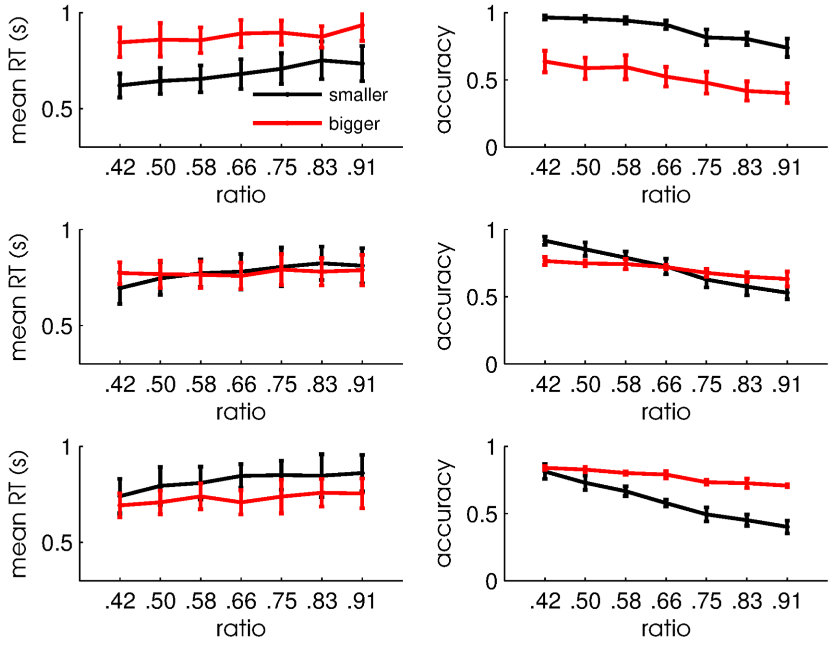

Figure 4. Mean correct RTs, and accuracy levels averaged across subjects for small standard conditions (first row), medium standard conditions (second row) and big standard conditions (third row). Error bars represents standard errors of the mean. The legend shows whether the target stimulus was smaller or bigger than the standard stimulus.

Figure 4 shows mean correct RTs as a function of the experimental condition when data are collapsed across participants and RTs lower than 0.3 s and bigger than 3 s are eliminated (about 0.5% of the data). The second column of plots of Figure 4 shows mean accuracy averaged across participants. The two plots on the top row show mean RTs and accuracy for conditions for which the standard stimulus was 'small' (i.e., it had 12 dots), the two plots on the middle row show mean RTs and accuracy for conditions for which the standard stimulus was 'medium' (i.e., 24 dots) and the two plots on the bottom row show mean RTs and accuracy for conditions for which the standard stimulus was 'big' (i.e., 36 dots).

Figure 3 and Figure 4 clearly show the presence of a semantic congruity effect, given that subjects, for conditions having the same ratio (e.g., 12 and 5 dots vs 12 and 19 dots), have different RTs and especially different accuracy depending on the congruency between size of the standard and of the target stimulus.

We entered correct RTs and accuracy levels in two different mixed-effect regression with ratio, magnitude of standard (abbreviated as 'magnitude') and correct response category (abbreviated as CRC) as dependent variables. In each regression, we included random effects for subject-specific constants and slopes. Regarding correct RTs, the regression showed that magnitude affected RTs, $B = .228$, 95% CI [.137, .318], $t = 4.977$, $p < .001$, with RTs increasing as magnitude increased. In particular, for each increase in magnitude, RTs increased between .137 s and .318 s. This effect suggests that subjects were biased towards answering 'smaller'. CRC affected RTs, $B = .449$, 95% CI [.333, .564], $t = 7.628$, $p < .001$, with RTs being higher when the CRC was ‘bigger’ compared to when it was 'smaller'. Also this effect suggests a bias towards answering smaller. As expected, ratio affected RTs, $B = .462$, 95% CI [.223, .701], $t = 3.794$, $p < .001$, with RTs increasing as ratio (i.e., difficulty) increased. The interaction effect of magnitude and CRC affected RTs, $B = -.155$, 95% CI [-.227, -.082], $t = -4.322$, $p < .001$.

As shown in Figure 4 when magnitude increased and CRC was 'bigger', RTs decreased, while when magnitude increased and CRC was 'smaller', RTs increased. Ratio by CRC affected RTs, $B = -.201$, 95% CI [-.365, -.037], $t = -2.406$, $p = .016$. As expected and as shown in Figure 4, the effect of CRC was larger for difficult discriminations compared to easy discriminations. Regarding accuracy, the binary logistic mixed effect regression showed that magnitude affected accuracy, $B = -2.903$, 95% CI [-4.408, -1.397], *Exp(B)* = .055, $t = -3.986$, $p = .001$, with accuracy decreasing when magnitude increased. CRC affected accuracy, $B = -5.665$, 95% CI [-7.235, -4.096], *Exp(B)* = .003, $t = -7.388$, $p < .001$, with accuracy decreasing when the CRC category was 'bigger', confirming that subjects were generally biased towards answering 'smaller'. Ratio affected accuracy, $B = -8.534$, 95% CI [-10.673, -6.395], *Exp(B)* < .001, $t = -7.912$, $p < .001$, with accuracy decreasing when difficulty increased. The interaction of magnitude and CRC affected accuracy, $B = 1.670$, 95% CI [.213, 3.126], *Exp(B)* = 5.31, $t = 2.394$, $p = .027$. As shown in Figure 3 and Figure 4 when magnitude increased and CRC was bigger accuracy increased, while when magnitude increased and CRC was smaller, accuracy decreased. Ratio by CRC affected accuracy, $B = 3.139$, 95% CI [1.454, 4.824], *Exp(B)* = 23.079, $t = 3.768$, $p = .001$, confirming that the effect of CRC was larger for difficult discriminations compared to easy discriminations.

Model Fitting

To fit the DDM to our data we used the Diffusion Model Analysis Toolbox (DMAT; Vandekerckhove & Tuerlinckx, 2007, 2008) for Matlab (version 2013b). Among the options available, we used as objective function a chi-square function. We decided to represent the RT distributions of responses in terms of six bins, defined by the boundaries of the conventional .1, .3, .5, .7 and .9 quantile bins dividing the correct and error RT distributions (Vandekerckhove & Tuerlinckx, 2007). In DMAT, the observed response frequencies are compared to the expected response frequencies and a chi-square statistic is minimised to find the best fitting parameters.

For each participant the drift could be (i) fixed across conditions, or (ii) free to vary across conditions; the boundary separation could be (i) fixed across conditions, or (ii) free to vary across conditions; the starting point could be (i) fixed across conditions, or (ii) free to vary across conditions; and finally the non-decision time could be (i) fixed across conditions, or (ii) free to vary across conditions. Across-trials variabilities in drift, non-decision time and starting point were kept constant across conditions in order to avoid over-fitting. It should be noted that we also fitted a series of models in which, when a parameter was free to vary, its across-trials variability parameter was also free to vary. However, in this case DMAT warned that the variability parameters were not identified by the data or that their standard error estimates were biased. In theory, we could have used bootstrapping for an estimation of the parameters and their standard errors, but given the number of models to be fitted and the number of iterations required for the bootstrapping, this would have been computationally intensive (i.e., the fitting would have taken days to complete). Furthermore, when across-trials variabilities were fixed across conditions, DMAT did not provide warnings for the best model, so we opted for this option.

All possible combinations of models were fitted to each individual resulting in a total of 16 models per participant. To assess which model best satisfies the trade-off between simplicity and goodness of fit, we used a statistical criterion for model selection, the Bayesian Information Criterion (BIC; Raftery, 1995), calculated as $-2 \cdot loglikelihood(data|model) + k \cdot logN$, where k is the number of free parameters in the model and N the total number of observations. The BIC is a measure of goodness of fit to which a penalty for the introduction of parameters is added. The best model is the model with the lowest BIC value; as proposed in Kass and Raftery (1995), a difference of ten in BIC scores between two models is considered a strong evidence towards the model with the lowest BIC score. For all participants, the model in which only the drift rate was allowed to vary across conditions, was selected by far as the best model, with differences in BIC scores being always greater than 75 if the best model is compared to the second-best model, showing a striking preference for this model.

As it is clear from plotting the drift rate recovered from the fitting for each participant - Figure 5 -, the drift rate was (i) a function of the ratio between the standard and the target stimulus (i.e., higher the ratio, lower the drift) and (ii) a function of the magnitude of the standard (i.e., when the magnitude of the standard increases, the drift shifts towards the boundary for the response 'bigger'). In Figure 5, when drift values are positive, it means that they drifted towards the threshold for the response 'bigger', while when drift values are negative, it means that the process was directed towards the boundary for the response 'smaller'. Figure 5 shows also that in general participants were more biased towards answering 'smaller' given that the slope for this alternative is generally steeper than the slope for the opposite response; this is in line with the behavioural analyses showing a main effect of CRC on accuracy and RTs.

A linear regression on drift rates showed a main effect of CRC, B = .570, 95% CI [.408, .731], t = 6.961, p < .001, with drifts being higher when the CRC was smaller compared to when it was bigger. Magnitude affected drift, B = .065, 95% CI [.012, .118], t = 2.435, p = .016; as the magnitude increased, the slope of the drift rate increased suggesting an additive influence of the magnitude of the standard on the drift rates. As expected ratio affected drift rates, B = .319, 95% *CI* [.152, .486], t = 3.767, p < .001, with drift rates being higher when the comparison was simpler. The interaction effect of CRC and ratio resulted significant, *B* = -.534, 95% CI [-.770, -.298], *t* = -4.466, *p* < .001. In line with the behavioural results, this shows that the effect of CRC was stronger for difficult conditions.

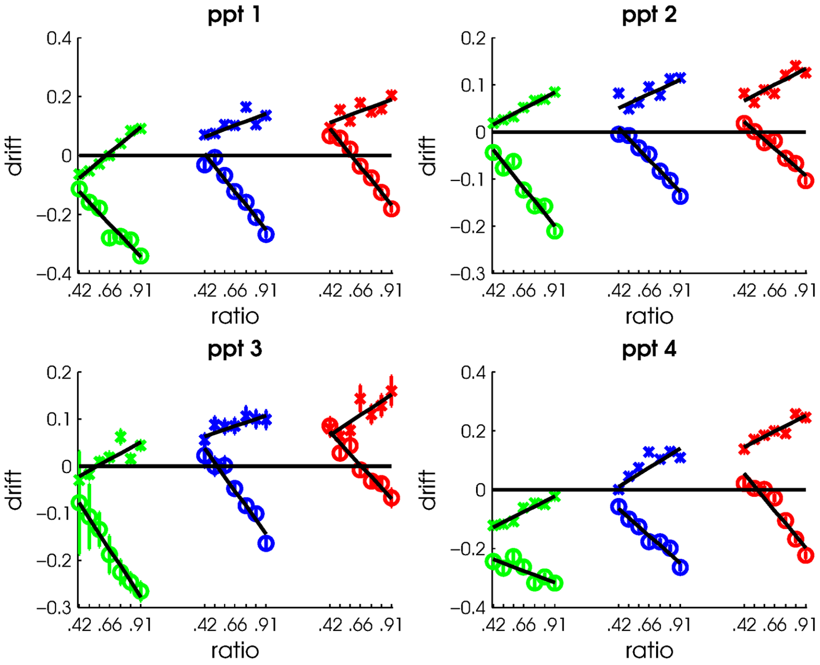

Figure 5. Graphical representation of drift rate for each of the four participants, abbreviated 'ppt'. The horizontal line for the drift represents the level at which the drift is 0. The green symbols are for conditions for which the standard was smaller (12 dots), the blue symbols are for conditions for which the standard was medium (24 dots), the red symbols are for conditions for which the standard was big (36 dots). 'o' for conditions for which the CRC was smaller, 'x' for conditions for which the CRC was bigger. All ratios are reported, even though on the x label are labelled only the two extreme ratio conditions (.42 and .91) and the .66 ratio conditions. Error bars are standard errors of parameters' estimates recovered by DMAT.

Interestingly, in our plots the numerosity is represented on a linear scale (best linear fit reported in Figure 5) of the ratio of the two numerosities to compare and this is not in line with the results of Park and Starns (2015) - in their study drift rates followed the logarithm of the ratio of the two numbers to compare.

The remaining parameters and their standard errors estimated by DMAT, for each participant, are shown in Table 2.

Table 2

Estimates and Standard Errors of a, ter, eta, z, s_z and s_t for Each Participant (Abbreviated 'ppt').

Participant ID	**Parameter estimate and SE**	**a**	**ter**	***eta***	**z**	**s_z**	**s_t**
ppt1	estimate	.148	.299	.160	.077	<.001	.132
	SE	.002	.002	.011	.001	.019	.012
ppt2	estimate	.185	.266	.081	.087	<.001	.004
	SE	.002	.003	.005	.001	.009	.039
ppt3	estimate	.144	.507	.120	.075	<.001	.408
	SE	.006	.012	.044	.004	.012	.024
ppt4	estimate	.148	.272	.173	.064	<.001	.100
	SE	.001	.002	.006	.001	.008	.007

Fits of the model to the data are represented by quantile probability plots, Figure 6. Quantile probability plots are a powerful way of showing the goodness of fit; on the x-axis it is shown the probability of a correct and of a wrong response for the model and the data, while on the y-axis are shown the quantile-RTs that divide the distributions of correct and wrong response, both for the model and the data. Here, we show the conventional .1, .3, .5, .7 and .9 quantiles that divide the RT distributions, for correct and error responses (Ratcliff & McKoon, 2008).

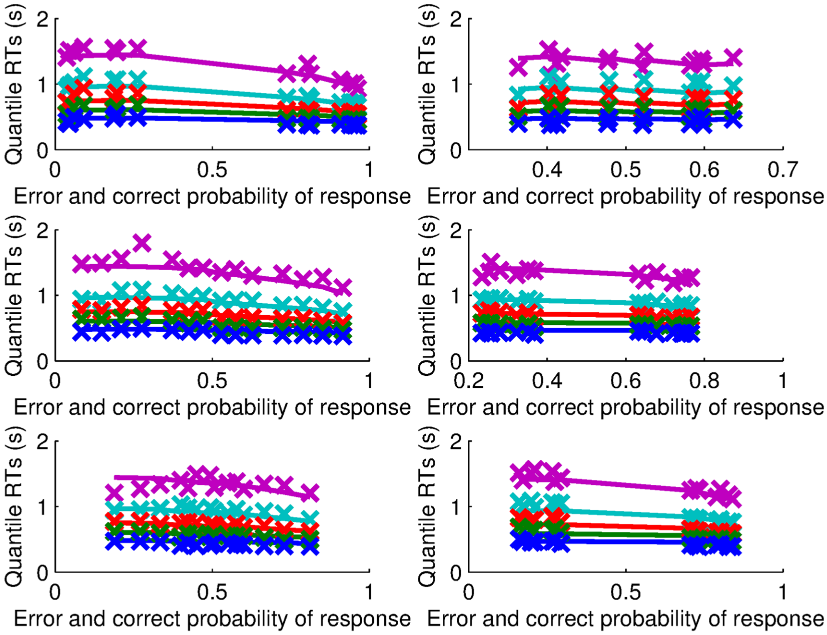

Figure 6. Quantile probability plots showing predictions of the model (recovered from the parameters averaged across individuals and represented by the lines) and the data (averaged across individuals and represented by 'x'). The different colours reflect the different quantiles (.1, .3, .5, .7, .9) of the RT distribution for each condition.

In Figure 6, we compare the predictions of the model based on the parameters averaged across individuals (represented by the lines), and the observed data pooled across individuals (represented by 'x'). Figure 6 has 6 plots; the two plots on top show conditions for which the standard was small, the plots on the middle show conditions for which the standard was medium and the plots on the bottom show conditions for which the standard was big. The plots on the left of Figure 6 show conditions for which the correct response category was 'smaller', while the plots on the right show conditions for which the correct response category was 'bigger'. Note that, as the behavioural analyses show, for conditions with a high ratio (i.e., high difficulty), the overall performance of subjects dropped below chance in some cases. As a consequence, for these conditions, the probability of a correct choice lays on the left of the graph, and the probability of an incorrect choice is on the right side of the graph, mostly near to chance level. In general, for conditions with highly discriminable stimuli (i.e., conditions with low ratio) little weight should be accorded to the quantiles for error responses since these are mainly influenced by a very limited and potentially unreliable number of measurements given that subjects made very few errors in these extreme conditions.

The quantile probability plots show that the model obtained from our fitting can capture the averaged data well, especially considering that the data are averaged across four experimental sessions with clear repercussions on the within-subject variability, and considering the high number of conditions present in this study.

Discussion

As we described earlier, several theories have been proposed for the explanation of the semantic congruity effect (Banks et al., 1975; Banks & Flora, 1977; Banks et al., 1976; Holyoak & Mah, 1982; Marschark & Paivio, 1979; Ryalls et al., 1998). Here, we have adopted a computational framework, the drift diffusion model (DDM) that is psychologically plausible, mathematically rigorous and that has been shown to fit data in various psychological tasks (Ratcliff, 1978, 2002; Ratcliff & McKoon, 1988; Ratcliff & Rouder, 1998; Ratcliff et al., 2004, 1999; Thapar et al., 2003; Voss et al., 2004). Our results show that the DDM (Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff et al., 1999) can account for the data in an experiment in which we have elicited a semantic congruity effect.

We found that the changes in decision time and accuracy associated with our manipulation, can be best explained by a change in the drift rate. The drift rate is associated with the discriminability of the experimental condition, as it is commonly assumed in the DDM, but it is also affected by the magnitude of the standard stimulus. This effect seems to suggest that subjects were first assessing the numerosity of the standard and the magnitude of the standard biased them towards one of the two response categories. In particular, when the standard was small subjects were biased in answering 'smaller'; vice versa, when the standard was big subjects were biased in answering 'bigger'. Specifically, in this study subjects may have learnt to use the two extreme standard magnitudes as reference points for the values 'small' and 'big', since over the four experimental session the numerosity of the standard only consisted of three possible values. This strategy would result in the pattern observed in the data with subjects being faster and more accurate in judging which of the two stimuli is bigger/smaller when there was congruency between the magnitude of the standard and of the target. Interestingly, for the medium magnitude standard, being equidistant from the small and the big standard magnitude, the semantic congruity effect cancels out. The decision process in this study can be described as a two-stage DDM; first subjects had to asses the size of the standard stimulus. Afterwards, subjects had to assess whether the target stimulus was bigger or smaller than the standard and in case of congruency between the size of the standard and the relative size of the target, the response was faster and more accurate (i.e., drift rates were higher).

The main result of this study is in line with reference point models (see Chen et al., 2014; Dehaene, 1989) and it is relevant for theories in which the congruency between magnitude of the stimulus and direction of the comparison affects the strength of the evidence signal (Leth-Steensen & Marley, 2000; Petrusic et al., 2008). However, a key point of the models proposed by Leth-Steensen and Marley (2000) and by Petrusic et al. (2008) is that the semantic congruity effect arises when there is congruency between the comparison instruction and the relative size of the stimuli, while in our case, the semantic congruity is driven by the magnitude of the standard stimulus. Further theoretical work - in which such theories are framed within a DDM framework - and experimental work - in which the direction of the comparison is explicitly given - is needed to test Leth-Steensen and Marley (2000) and Petrusic et al. (2008) explanations, given that the experimental paradigm presented here and the conceptual explanation that we provided vary greatly from their conceptualisation of a similar phenomenon (i.e., semantic congruity effect in selection paradigm). For these theories, it has been proposed (Leth-Steensen, Petrusic, & Shaki, 2014) that it is the relative size of the stimulus pair that 'primes' the corresponding congruent form of the instruction, resulting in a facilitation in case of congruency. Our result is, in theory, also in line with an account in which the relative size of the stimulus pair 'primes' the corresponding congruent response category, resulting in a facilitation in case of congruency. However, in our case, it is not clear why an assessment of the overall size of the stimulus pair is necessary since it is not explicitly required. Nevertheless, it should be noted that an effect of the overall magnitude of the stimulus pair is in line with recent findings showing that in two-alternatives forced choice tasks, both absolute and relative evidence are integrated by participants (Starns, Chen, & Staub, 2017). If this is the case, subjects may not be able to ignore the absolute size of the two stimuli and also this could in theory account for our results. However, the result that semantic congruity effects arise even when the standard stimulus and the target stimulus are presented sequentially (Dehaene, 1989; Link, 1990), seems to undermine the role of the size of the stimulus pair in the explanation of semantic congruity effect for classification paradigms; given that subjects are presented with a target *after* the presentation of a standard, assessing the overall magnitude of the stimuli means assessing the magnitude of the target itself, and this is a clearly problematic assumption (Leth-Steensen & Marley, 2000) - as once the magnitude of the target is assessed, the response can be executed without any need to bias the decision. This argument leads us to conclude that even though semantic congruity effects in classification and selection paradigms can be due in both cases to an increase in the rate at which evidence is accumulated, the decision process faced by subjects for these two tasks varies greatly, hence it is reasonable to expect that different conceptual explanations are needed for the two tasks.

Here, we invalidate theories which interpret the semantic congruity effect as a modification in starting point of evidence accumulation (Birnbaum & Jou, 1990; Link, 1990; Link & Heath, 1975; Poltrock, 1989). Furthermore, the other principal theories that have been proposed for the explanation of the semantic congruity effect - the expectancy effect (Banks & Flora, 1977; Marschark & Paivio, 1979), the semantic coding model (Banks et al., 1975, 1976) and the frequency explanation (Ryalls et al., 1998) - seem to be already falsified by the contrastive results presented in the introduction. Furthermore, the expectancy theory and the semantic coding model do not apply in our study, given that they are dependent on the direction of the comparative instruction that is not used in the current task.

The choice of previous authors to not consider other decision mechanisms (e.g., boundary separation or non-decision time) in accounting for semantic congruity effects, is questionable. Here, we show directly - with the model selection procedure - that neglected mechanisms, such as boundary separation or non-decision time variations, do not play a role in the semantic congruity effect.

Our application of the DDM further highlights the heuristic power of the DDM, and shows that different phenomena that have been previously explained by descriptive and or task-specific theories can be accounted for by sequential sampling models of evidence accumulation and decision making, when the focus is shifted to the computational level of analysis. Our formal account of this phenomenon is parsimonious, as it uses a unifying model of choice rather than proposing an ad-hoc model for the explanation of the phenomenon, and rigorous, as we account for the full distributions of correct and error responses, by taking into consideration all the cognitive processes that underlie a decision.

Supplementary Materials

The data and the script used for the fitting. doi:10.17605/OSF.IO/QQGYU

Funding

Angelo Pirrone was supported by the University Studentship Network in Neuroeconomics. The funding source had no other role other than financial support.

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

We thank Stephen Link, Andrew Heathcote, Jeffrey Starns and two anonymous reviewers for their insightful comments and suggestions.

General Note

This work appeared in Angelo Pirrone’s PhD dissertation.

References

Banks, W. P., Clark, H. H., & Lucy, P. (1975). The locus of the semantic congruity effect in comparative judgments. *Journal of Experimental Psychology: Human Perception and Performance, 1*(1), 35-47. doi:10.1037/0096-1523.1.1.35

Banks, W. P., & Flora, J. (1977). Semantic and perceptual processes in symbolic comparisons. *Journal of Experimental Psychology: Human Perception and Performance, 3*(2), 278-290. doi:10.1037/0096-1523.3.2.278

Banks, W. P., Fujii, M., & Kayra-Stuart, F. (1976). Semantic congruity effects in comparative judgments of magnitudes of digits. *Journal of Experimental Psychology: Human Perception and Performance, 2*(3), 435-447. doi:10.1037/0096-1523.2.3.435

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences of the United States of America, 107*(50), 21767-21772. doi:10.1073/pnas.0908104107

Birnbaum, M. H., & Jou, J.-W. (1990). A theory of comparative response times and “difference” judgments. *Cognitive Psychology, 22*(2), 184-210. doi:10.1016/0010-0285(90)90015-V

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700-765. doi:10.1037/0033-295X.113.4.700

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436. doi:10.1163/156856897X00357

Cantlon, J. F., & Brannon, E. M. (2005). Semantic congruity affects numerical judgments similarly in monkeys and humans. *Proceedings of the National Academy of Sciences of the United States of America, 102*(45), 16507-16511. doi:10.1073/pnas.0506463102

Chen, D., Lu, H., & Holyoak, K. J. (2014). The discovery and comparison of symbolic magnitudes. *Cognitive Psychology, 71*, 27-54. doi:10.1016/j.cogpsych.2014.01.002

Dehaene, S. (1989). The psychophysics of numerical comparison: A reexamination of apparently incompatible data. *Perception & Psychophysics, 45*(6), 557-566. doi:10.3758/BF03208063

Gebuis, T., & Reynvoet, B. (2012). The role of visual information in numerosity estimation. *PLOS ONE, 7*(5), Article e37426. doi:10.1371/journal.pone.0037426

Gertner, L., Arend, I., & Henik, A. (2012). Effects of non-symbolic numerical information suggest the existence of magnitude-space synesthesia. *Cognitive Processing, 13*(S1), S179-S183. doi:10.1007/s10339-012-0449-9

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron, 36*(2), 299-308. doi:10.1016/S0896-6273(02)00971-6

Holyoak, K. J. (1978). Comparative judgments with numerical reference points. *Cognitive Psychology, 10*(2), 203-243. doi:10.1016/0010-0285(78)90014-2

Holyoak, K. J., & Mah, W. A. (1982). Cognitive reference points in judgments of symbolic magnitude. *Cognitive Psychology, 14*(3), 328-352. doi:10.1016/0010-0285(82)90013-5

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795. doi:10.1080/01621459.1995.10476572

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What’s new in psychtoolbox-3. *Perception, 36*(14), 1-16.

Leth-Steensen, C., & Marley, A. (2000). A model of response time effects in symbolic comparison. *Psychological Review, 107*(1), 62-100. doi:10.1037/0033-295X.107.1.162

Leth-Steensen, C., Petrusic, W. M., & Shaki, S. (2014). Enhancing semantic congruity effects with category-contingent comparative judgments. *Frontiers in Psychology, 5*, Article 1199. doi:10.3389/fpsyg.2014.01199

Link, S. W. (1990). Modeling imageless thought: The relative judgment theory of numerical comparisons. *Journal of Mathematical Psychology, 34*(1), 2-41. doi:10.1016/0022-2496(90)90010-7

Link, S. W., & Heath, R. (1975). A sequential theory of psychological discrimination. *Psychometrika, 40*(1), 77-105. doi:10.1007/BF02291481

Marks, D. F. (1972). Relative judgment: A phenomenon and a theory. *Perception & Psychophysics, 11*(2), 156-160. doi:10.3758/BF03210364

Marschark, M., & Paivio, A. (1979). Semantic congruity and lexical marking in symbolic comparisons: An expectancy hypothesis. *Memory & Cognition, 7*(3), 175-184. doi:10.3758/BF03197536

Mewhort, D. J. K., Smith, D. G., & Kohly, R. (1996). Interference in memory produces numerical distance effects. *Journal of Mathematical Psychology, 40*, 349.

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology, 8*(2), 228-246. doi:10.1016/0010-0285(76)90025-6

Park, J., & Starns, J. J. (2015). The approximate number system acuity redefined: A diffusion model approach. *Frontiers in Psychology, 6*, Article 1955. doi:10.3389/fpsyg.2015.01955

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437-442. doi:10.1163/156856897X00366

Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. *Journal of Experimental Psychology: Human Perception and Performance, 18*(4), 962-986. doi:10.1037/0096-1523.18.4.962

Petrusic, W. M., Baranski, J. V., & Kennedy, R. (1998). Similarity comparisons with remembered and perceived magnitudes: Memory psychophysics and fundamental measurement. *Memory & Cognition, 26*(5), 1041-1055. doi:10.3758/BF03201182

Petrusic, W. M., Shaki, S., & Leth-Steensen, C. (2008). Remembered instructions with symbolic and perceptual comparisons. *Perception & Psychophysics, 70*(2), 179-189. doi:10.3758/PP.70.2.179

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., . . . Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition, 116*(1), 33-41. doi:10.1016/j.cognition.2010.03.012

Piazza, M., Fumarola, A., Chinello, A., & Melcher, D. (2011). Subitizing reflects visuo-spatial object individuation capacity. *Cognition, 121*(1), 147-153. doi:10.1016/j.cognition.2011.05.007

Pirrone, A., Stafford, T., & Marshall, J. A. R. (2014). When natural selection should optimize speed-accuracy trade-offs. *Frontiers in Neuroscience, 8*, Article 73. doi:10.3389/fnins.2014.00073

Poltrock, S. E. (1989). A random walk model of digit comparison. *Journal of Mathematical Psychology, 33*(2), 131-162. doi:10.1016/0022-2496(89)90027-8

Provost, A., & Heathcote, A. (2015). Titrating decision processes in the mental rotation task. *Psychological Review, 122*(4), 735-754. doi:10.1037/a0039706

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-163. doi:10.2307/271063

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59-108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review, 9*(2), 278-291. doi:10.3758/BF03196283

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology, 53*(3), 195-237. doi:10.1016/j.cogpsych.2005.10.002

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review, 95*(3), 385-408. doi:10.1037/0033-295X.95.3.385

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873-922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*(5), 347-356. doi:10.1111/1467-9280.00067

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19*(2), 278-289. doi:10.1037/0882-7974.19.2.278

Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition, 137*, 115-136. doi:10.1016/j.cognition.2014.12.004

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*(2), 261-300. doi:10.1037/0033-295X.106.2.261

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science, 19*(6), 607-614. doi:10.1111/j.1467-9280.2008.02130.x

Revkin, S. K., Piazza, M., Izard, V., Zamarian, L., Karner, E., & Delazer, M. (2008). Verbal numerosity estimation deficit in the context of spared semantic representation of numbers: A neuropsychological study of a patient with frontal lesions. *Neuropsychologia, 46*(10), 2463-2475. doi:10.1016/j.neuropsychologia.2008.04.011

Ryalls, B. O., Winslow, E., & Smith, L. B. (1998). A semantic congruity effect in children’s acquisition of high and low. *Journal of Memory and Language, 39*(4), 543-557. doi:10.1006/jmla.1998.2594

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area lip) of the rhesus monkey. *Journal of Neurophysiology, 86*(4), 1916-1936.

Starns, J. J., Chen, T., & Staub, A. (2017). Eye movements in forced-choice recognition: Absolute judgments can preclude relative judgments. *Journal of Memory and Language, 93*, 55-66. doi:10.1016/j.jml.2016.09.001

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging, 18*(3), 415-429. doi:10.1037/0882-7974.18.3.415

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*(6), 1011-1026. doi:10.3758/BF03193087

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40*(1), 61-72. doi:10.3758/BRM.40.1.61

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13*(1), 37-58. doi:10.1080/00140137008931117

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206-1220. doi:10.3758/BF03196893

Wald, A. (1947). *Sequential analysis*. New York, NY, USA: Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics, 19*(3), 326-339. doi:10.1214/aoms/1177730197

Wallis, C. P., & Audley, R. J. (1964). Response instructions and the speed of relative judgements. *British Journal of Psychology, 55*(2), 121-132. doi:10.1111/j.2044-8295.1964.tb02712.x

Wang, X.-J. (2013). Neuronal circuit computation of choice. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics: Decision making and the brain* (2nd ed.) London, United Kingdom: Elsevier Academic Press.