Cognition xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Cognition

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

The difficulties of executing simple algorithms: Why brains make mistakes computers don't

Gary Lupyan [*]

University of Wisconsin, Madison, United States

ARTICLE INFO

Article history:
Received 31 December 2012
Revised 9 August 2013
Accepted 12 August 2013
Available online xxxx

Keywords:
Concepts
Categorization
Human algorithms
Prototypes
Distributed representations
Inference

ABSTRACT

It is shown that educated adults routinely make errors in placing stimuli into familiar, well-defined categories such as TRIANGLE and ODD NUMBER. Scalene triangles are often rejected as instances of triangles and 798 is categorized by some as an odd number. These patterns are observed both in timed and untimed tasks, hold for people who can fully express the necessary and sufficient conditions for category membership, and for individuals with varying levels of education. A sizeable minority of people believe that 400 is more even than 798 and that an equilateral triangle is the most "trianglest" of triangles. Such beliefs predict how people instantiate other categories with necessary and sufficient conditions, e.g., GRANDMOTHER. I argue that the distributed and graded nature of mental representations means that human algorithms, unlike conventional computer algorithms, only approximate rule-based classification and never fully abstract from the specifics of the input. This input-sensitivity is critical to obtaining the kind of cognitive flexibility at which humans excel, but comes at the cost of generally poor abilities to perform context-free computations. If human algorithms cannot be trusted to produce unfuzzy representations of odd numbers, triangles, and grandmothers, the idea that they can be trusted to do the heavy lifting of moment-to-moment cognition that is inherent in the metaphor of mind as digital computer still common in cognitive science, needs to be seriously reconsidered.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

In October, 2012 *Slate* magazine reported on a court case concerning a disputed election of a juvenile judge in Hamilton County, OH (Hasen, 2012). At issue were split-precinct polling places that required poll workers to hand out the appropriate ballots based on a rule such as whether the voter's address was even or odd. A poll worker testified to sending a voter with the address "798" to vote in the precinct for voters with odd-numbered addresses. Court testimony reveals that when asked whether the house number 798 was even or odd, the poll worker responded: "Odd." (*Tracie Hunter v. Hamilton County Board of Elections*, 2012). The remaining testimony follows:

> Q. . .So on Election Day, if somebody came in with an address 798 and you had two ranges to choose from, you would choose the odd for them?
> A. Yes.
> Q. Okay. And is that how you did it for all the ballots that you looked up on Election Day?
> A. To determine if they were even – yes.
> Q. To determine if they were even or odd, you looked at the first digit of the address?
> A. No. I looked at the whole address.
> Q. And [if] there were more odds than even numbers, it would be an odd address?
> A. Yes.

Although we can all agree with Hasen's conclusion that "no one should lose the right to vote because a poll worker can't tell an odd from an even number," it is worth considering whether such mistakes reveal something deeper

[*] Address: 1202W. Johnson St., University of Wisconsin, Madison, Madison, WI 53706, United States. Tel.: +1 917 843 4868.
E-mail address: lupyan@wisc.edu

0010-0277/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2013.08.015

about human cognition than an individual's confusion about the definition of numerical parity. In a series of experiments, I show that such classification errors are endemic, even when individuals' explicit definitions for determining category membership appear entirely correct. I argue that the reason people err in classifying items into categories with clear boundaries and known membership criteria is that human categorization algorithms are inherently sensitive to the particulars of the input. Thus, although the proposition *N* IS EVEN is *either* true or false, the mental representations—the *psychological* concept of parity[1]—may display the kind of graded, probabilistic structure that is characteristic of other concepts with fuzzier boundaries.

The question of how concepts are represented by the mind is at the very core of cognitive science (Fodor, 2001; Murphy, 2002; Prinz, 2004). The past 50 years has seen classical theories of concepts stressing necessary and sufficient conditions give way to theories stressing vagueness and context-dependence (Barsalou, 1987; Hampton, 2006; Lakoff, 1990; Medin & Smith, 1984; Prinz, 2004; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Rosch, 1973). In large part, these theories were created to account for the ease with which people adapt their knowledge to novel contexts (e.g., Clark, 1983; Fauconnier & Turner, 2003). Much of the evidence used to support these probabilistic and prototype-based theories of concepts came from studies in which one measures how people identify category members under various circumstances. In a now classic paper, Armstrong, Gleitman, and Gleitman (1983) challenged the idea that such tests can tell us much about the nature of conceptual representations by showing that both categories like FRUIT and ODD NUMBER showed graded structure as revealed by typicality ratings and longer classification times of "atypical" members (cf. Larochelle, Richard, & Soulières, 2000; cf. Sandberg, Sebastian, & Kiran, 2012). Armstrong et al. (1983) argued that because it is inconceivable that someone who knows the definition of numerical parity would truly believe that some numbers are odder than others, the finding that categorizing members from such categories as ODD NUMBERS—palpably different, according to the authors from fuzzier categories like PET and FRUIT—meant that the results from rapid classifications tasks must reflect functioning of peripheral identification procedures rather than tapping into "core" conceptual content (see Geeraerts, 1989 for discussion; and Gleitman, Armstrong, & Connolly, 2012 for a restatement of this position). Thus, although the difference in the time to classify an apple and a coconut as fruits may stem from a more "central" position of apple within the feature-space of FRUIT, the finding that it takes longer to identify 18 than 22 as EVEN cannot, according to the standard view, reflect such a difference. I present a series of studies showing that people, in fact, represent some numbers as odder than others, some triangles as more triangular, and argue that these effects stem from a failure to fully abstract from the details of the input making human algorithms qualitatively different from context-free computer algorithms that have inspired classical cognitive science.

Understanding the computations that underlie classification is relevant not only for understanding explicit categorization, but also informs theories of cognition more broadly. For example, many language parsers require words to be classified into abstract categories on which further computations are performed (Chomsky, 1995; cf. Anderson, 2006; Sleator & Temperley, 1995). Such assumption have led some to argue that e.g., infants' sensitivity to the similarity structure of the syllable sequences ABA and CDC arises from algebraic computations that treat syllables as context-independent variables (Marcus, 1999; cf. Seidenberg, 1999). On some theories, such symbolic manipulation is not limited to any special domain, but characterizes the entirety of mental processes (e.g., Gallistel & King, 2009). Given the relative simplicity of e.g., the algorithm for computing numerical parity, any symbolic device worth its salt should be able to abstract from the "surface" properties of the input in computing parity. The 13 experiments below test this basic hypothesis. For convenience, a summary of the basic manipulations and results is listed in Table 2.

2. Experiment 1. Speeded parity judgments

In the first experiment participants completed a standard classification task requiring judgments of numerical parity. Of interest was whether people who could all articulate the correct definition of parity would nevertheless make errors in classifying numbers having opposite-parity digits, such as 798.

2.1. Participants and procedure

Ten undergraduate students participated for credit. Each trial began with a fixation cross (0.9–1.1 s) followed by a 1–4 digit numeral displayed for 1.0 s or until response. The numerals appeared in a random position within an invisible horizontally-oriented rectangle (~15° × 5°). Each digit subtended ~0.6° × 1° of visual angle. On half of the trials, the numerals were shown obliquely (±45° or ±60°). This oblique presentation helped to measure the contribution of perceptual-selection errors, as described below. Each participant completed 16 practice trials during which incorrect answers or timeouts were indicated by buzzes, followed by 243 experimental trials (Table 1) with timeout

Table 1
Distribution of trials in Exp. 1.

Number of digits	Number of opposite-parity digits			
	Zero	One	Two	Three
One	27[a]	–	–	–
Two	24	24	–	
Three	24	24	24	–
Four	24	24	24	24

[a] Zero was omitted; (1–9) × 3 repetitions.

[1] It is necessary to distinguish between concepts in the philosophical sense, concerned with the actual state of the world, and concepts in the psychological sense, concerned with mental content—how people actually represent the world. It is this psychological definition that is used here (see also Hampton, 2012).

Please cite this article in press as: Lupyan, G. The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition* (2013), http://dx.doi.org/10.1016/j.cognition.2013.08.015

Table 2
A summary of tasks, manipulations, and core findings.

Experiment	*n*	Task description	Basic manipulation	Basic finding
1	10[a]	Speeded parity judgments	Parity judgments of 1–4 digit numerals with a 1s response deadline	Participants are slower/less accurate at judging the parity of numbers having more opposite-parity digits (#OPD), e.g., 798
2A	193	Unspeeded parity judgments	No response deadline	Replication of 1: participants make more errors on numerals with higher #OPD
2B	100	Unspeeded parity judgments	Numerals are spelled out	Replication of 2A
3A	25[a]	Speeded parity judgments + flanker congruity task	Correlating performance between speeded parity-judgments and flanker congruity task	Replication of 1: Effect of digit-length is on RTs is predicted by flanker performance, but effect of #OPD is not
3B	108	Unspeeded parity judgments + unspeeded flanker congruity task	Correlating performance between unspeeded versions of parity and flanker tasks	Replication of 2A: No relationship with flanker performance
3C	98	Unspeeded parity judgments	Individual digits are now color-coded as in Exp. 3B. People are asked to explain strategy	Replication of 2A. People with correct definitions still show effect of #OPD. Those who explicitly mentioned looking at the last digit did not
4	75	Unspeeded parity judgments + typicality questions	Predicting effect of #OPD from beliefs about gradedness of parity	Participants who believe e.g., that 400 is a better even number than 798 are more adversely affected by #OPD
5A	145[b]	Classifying shapes as triangles and judging typicality of triangles	Classification of triangles as a function of their typicality	Non-canonical triangles are less likely to be classified as triangles
5B	80	Classifying shapes as triangles	Like 5B but choices now include non-triangles	Replication of 5A
5C	85	Classifying shapes as triangles	Like 5B but participants have to respond twice and justify their choices	Replication of 5A, 5B. Individual responses are internally consistent
5D	65	Making inferences of a basic geometric property	Like 5A, but participants make inferences instead of over classification	Atypical triangles are more likely to be judged as not having angles that add up to 180°
6A	83[c]	The "Eligible Contestants" task	Only grandmothers can enter the contest	Grandmothers who are older and have more grandchildren are judged as *more* likely to win a contest for which all grandmothers are *equally* likely to win
6B	50	The "Eligible Contestants" task	Anchoring effect confound check	Results inconsistent with anchoring-based explanation of Exp. 6B

[a] Exps. 1 and 3A were run in a lab with college undergraduates. The remaining experiments were run on Amazon Mechanical Turk.
[b] Two separate groups participated: 96 completed the triangle classification task and 49 the typicality-rating task.
[c] Two separate groups participated: 50 completed the "Eligible contestants" task and a separate group of 33 were asked to rate the contestants on how typical of a grandmother each was.

buzzes only. Participants responded by using their left and right index fingers with hand-to-parity counterbalanced between participants.

2.2. Results and discussion

The dependent measures were accuracy and correct reaction times (RTs). Timeouts (1.8%) were coded as errors; excluding timeouts entirely did not change any of the analyses. Overall accuracy and RTs were 90.2% and 600 ms, respectively. A comparison of odd and even trials showed that parity did not affect RTs, $F < 1$, but odd numbers were classified somewhat more accurately ($M_{\text{odds}} = .92$, $M_{\text{evens}} = .88$), $F(1,9) = 8.94$, $p = .004$.

The critical analyses involved examining responses as a function of properties of the digit: number of digits, and the number of *opposite-parity digits* (henceforth *#OPD*), ranging from 0 to 3. As shown in Fig. 1 (top), RTs increased substantially when the numeral contained more digits (even controlling for numerical magnitude), $F(1,9) = 7.15$, $p = .008$, and, critically, with increasing *#OPD*, $F(1,9) = 51.08$, $p < .0005$; judging the parity of 3-OPD numerals took about 50 ms longer than judging parity of 0-OPD numerals.

Accuracy was not predicted by number of digits, $F < 1$, but, as clearly visible in Fig. 1 (bottom), was very reliably predicted by *#OPD*, $F(1,9) = 19.49$, $p < .0005$ (GLMs). A logistic regression including parity and *#OPD* as response predictors showed that for each opposite parity digit contained in the numeral, participants were 1.24 times more likely to classify it incorrectly, $z = -2.85$, $p = .004$.

Displayed orientation of the numerals did not affect accuracy, but reliably predicted RTs: $F(1,9) = 19.97$, $p < .0005$ with slower RTs for non-canonically oriented numbers (linear model including also the predictors above). The predictive power of opposite-parity digits did not reliably interact with the orientation of the displayed number, $F(1,9) = 2.00$, $p = .16$; the trend was for a stronger effect for canonically oriented than oblique numbers (the relevance of this finding will become important in the discussion below).

The results show that the time and accuracy with which people classify numbers as odd or even are strongly predicted by surface properties of the input. All participants "knew" what made a number even or odd. Nevertheless, their errors were not random, as might be expected if they simply pressed the wrong key or misapplied the "rule" for

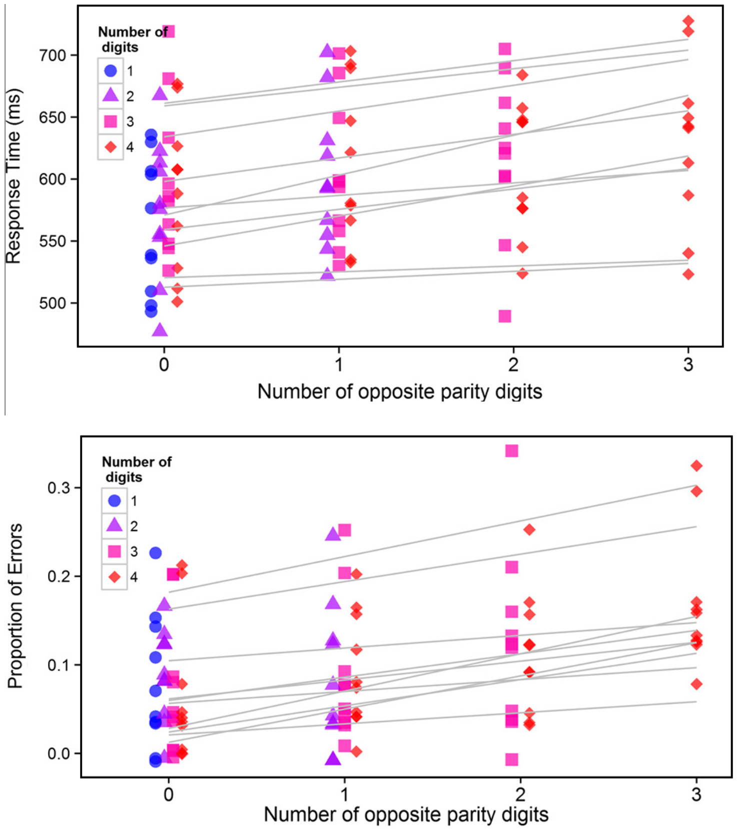


Fig. 1. Results of Experiment 1, showing RTs (top) and error-rates (bottom) as a function of the number of opposite-parity digits (*x*-axis) and number of digits (color). Points represent subject means and lines represent linear fits to performance of individual subjects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

some numerals. People who "knew" perfectly well that numbers ending on 8 are even numbers, were nevertheless likely to make errors on numbers like 798 (indeed, 798 was classified as odd 17.5% of the time).

3. Experiment 2A. Removing time pressure

Although participants in Experiment 1 had ample time to respond to each digit (timeouts occurred on <2% of the trials), the experiment imposed some time pressure, by design. The goal of Experiment 2 was to determine whether people are also more likely to misjudge the parity of numbers such as 798 when there is no time pressure.

3.1. Participants and procedure

Two hundred eight participants from the US and India were recruited from the online service *Amazon Mechanical Turk*. In this and subsequent studies using Mechanical Turk, it is important to validate the quality of the answers by distinguishing individuals who genuinely tried to do the task from those who attempted to minimize effort by e.g., not reading the instructions or clicking randomly. I eliminated any participants who failed to respond to one of the questions ($n = 2$), those who appeared to respond randomly (overall accuracy <60%), or those who responded identically to the two parity questions, $n = 13$, leaving a final sample of 193.

Each participant was shown the same 18 three-digit numbers and asked to click a checkbox next to all the numbers that were odd. They were then shown the same 18 numbers again and asked to click on all the ones that were even (parity order counterbalanced). The original number list was generated randomly with the constraint that the numbers have 0–2 opposite parity digits. Following the two parity questions, participants self-reported their educational level, age, and gender. To measure explicit knowledge regarding the rules governing odd/even membership, participants were asked to describe the difference between odd and even numbers. Three research assistants independently coded the responses for correctness/completeness (1–5 scale) and for classified definition type: definition (e.g., even numbers can be divided by 2 without a remainder), examples (e.g., numbers ending on 0, 2, 4, 6, or 8 are even), both, or neither. Agreement was high (Cronbach alphas >.9); a final score/definition type was obtained through subsequent consultation between the raters.

3.2. Results and discussion

The majority of participants had perfect performance, but almost 30% made errors, either failing to include a number in the correct category or including a number in the incorrect category. The basic error patterns are shown in Figs. 2 and 3. To determine the source of the errors, I ran

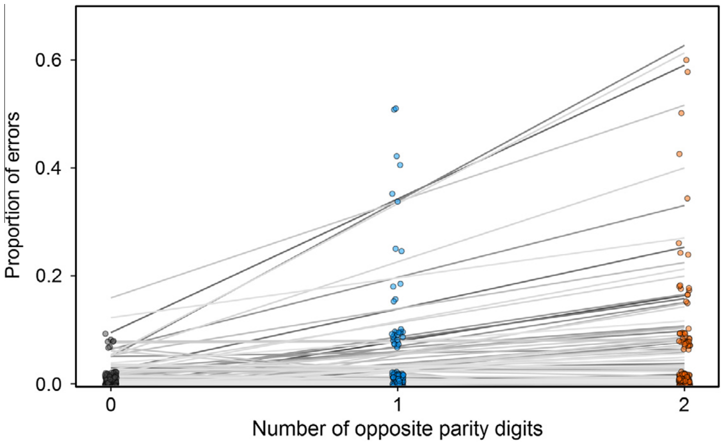


Fig. 2. Results of Exp. 2A. Points represent (jittered) means; lines represent linear fits to performance of individual subjects.

a logistic regression predicting responses from the actual parity of the digit, *#OPD*, participants' educational level and the correctness and type of their definition. The likelihood of choosing a numeral was, of course, predicted by the actual parity of the number and the parity being sought by the prompt, $z = 17.628$, $p \ll .0001$ but it was additionally predicted by *#OPD*, $z = -4.034$, $p \ll .0001$. Neither education, nor the correctness and type of definition predicted performance, nor did these factors interact with the observed effect of *#OPD*, $z < 1.5$. There were no differences between American and Indian participants in accuracy, nor was there a difference in the effect of opposite-parity digits on the performance of the two groups, $z \sim 0$, despite the Indian group reporting considerably higher educational achievement, $F(1,189) = 84.44$, $p < .0001$ (51% of the American sample reported having a college degree compared to 91% among the Indian group).

Although the likelihood of making an error for a 0-OPD number like 400 or 119 was very low ($M = 0.3\%$), the likelihood increased almost 10-fold for numbers like 213, $M = 2.8\%$, and then increased 1.25 times again for numbers like 798: $M = 3.5\%$, $\chi^2(2) = 80.97$, $p < .0005$. If subjects with perfect accuracy are excluded, the error rates rise to 1.1%, 9.9%, and 12.4%, respectively, $\chi^2(2) = 84.64$, $p \ll .0005$.[2]

Exp. 2A shows that participants make systematic errors in classifying the parity of digits even without time-pressure, failing to apply uniformly the definition they are fully able to articulate.[3] A primary function of a concept is to allow for the discrimination of category members and nonmembers. Although it may be possible to brush off differences in reaction times as being somehow peripheral of "true" categorization (Armstrong et al., 1983; Gleitman et al., 2012), it is difficult to dismiss the present findings with the same argument.

4. Experiment 2B. Parity judgments of spelled-out numerals

One possibility is that the errors in Exp. 2A stemmed from participants occasionally responding to the wrong digit (a possibility further tested in Exps. 3A–3C). If true, then better performance might be expected if the numbers are spelled out because, with the exception of numbers divisible by 10 and the 'teens', the English number system spells out the last digit—the only digit critical for making a parity judgment—in its entirety, marking it off in a highly salient way. Thus, if the error patterns in Exp. 2A stemmed from digit selection errors, they should be considerably reduced when numbers are spelled out.

4.1. Participants and procedure

One hundred sixteen participants were recruited. Excluding participants using the criteria listed in Exp. 2A left a final sample of 100. The procedure was identical to Experiment 2A except that the numbers were now spelled out, i.e., "Two-hundred and five" instead of 205.

4.2. Results and discussion

Results were remarkably similar to Exp. 2A with no reliable differences or interactions. Thirty-one percent of participants made errors and these errors were made disproportionately for numerals with opposite-parity digits, $z = 5.106$, $p \ll .0005$ (logistic regression including the same predictors, as in Exp. 2A). As before, education and country of origin were not predictive of performance and did not enter into reliable interactions, $z < 1$.

Because the exact same numerals were used in Experiments 2A and 2B, it is possible to compare the error rates for the individual numerals (e.g., 350 vs. Three-hundred fifty, 798 vs. Seven-hundred ninety-eight, etc.). Although

[2] Perfect accuracy of some participants combined with systematic errors of others produce highly non-normal distributions. The effects reported here replicate under a variety of tests: Logistic regression using raw responses, Chi-squared analysis of numbers of errors, and logistic model comparisons using maximum likelihood (Bates & Maechler, 2012). In fact, reliable predictors remain reliable even after adding Gaussian noise to the subject means until the distributions are normalized.

[3] At the end of the questionnaire participants were also asked whether they thought that some odd numbers were "more odd than others" (Armstrong et al., 1983); 26% of our participants concurred. These individuals had significantly lower accuracy ($p = .009$), but the performance of both concurrers and deniers was similarly affected by the number of opposite-parity digits, cf. Exp. 4.

Please cite this article in press as: Lupyan, G. The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition* (2013), http://dx.doi.org/10.1016/j.cognition.2013.08.015

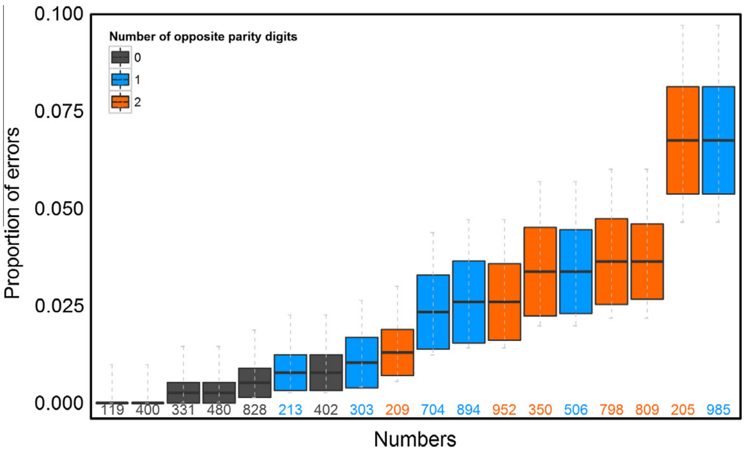


Fig. 3. Results of Exp. 2A. Proportion or errors for each numeral is shown, ordered by accuracy. The box contains the mean ±1 SE. The error bars include the 95% binomial CI. Colors represent the number of opposite-parity digits for the numeral. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

#OPD predicted errors to similar extent in both Experiments 2A and 2B (coefficients −.52 vs. −.48), the error rates for the individual numerals in the two studies were only moderately correlated, $r = .48$. Consider the number 350. When displayed in Arabic numerals, the error rate was 3.3%. When it was spelled out, the errors increased to 11.5%, $F(1,291) = 11.57$, $p = .001$. One speculation is that the psychological oddness of "Three-hundred fifty" stems from the use of 50 as a midpoint (e.g., half of a century, half of a dollar). When written using Arabic notation, the 0 at the end may help override the apparent oddness of the numeric representation activated by the word "fifty." The role of notation has been discussed at length in the past, e.g., in McCloskey's Triple Code model (Dehaene, Bossini, & Giraux, 1993 for discussion). The present results are consistent with a model in which numerical concepts, as activated by different cues (e.g., "three-hundred fifty" vs. "350") are, *however subtly*, different. Such a distinction is not sensible if concepts have fixed cores (Armstrong et al., 1983; Machery, 2009), but entirely consistent with a view in which conceptual representations are viewed as the in-context activation patterns with no dividing lines between core and periphery (Casasanto & Lupyan, in press; Elman, 2009; Lupyan & Thompson-Schill, 2012).

Experiment 2B replicates and extends the finding of Exp. 2A: when judging numerical parity of spelled out numerals, people are overwhelmingly more likely to make errors for numbers with opposite parity digits.

5. Experiments 3A–3C. Are parity-judgment errors due to failures of perceptual analysis?

One explanation for the error patterns observed in Exps. 1–2 is that participants represent some numbers as being truly odder than others. On this account, contrary to the arguments of Armstrong et al. (1983), and Gleitman et al. (2012), membership in formally defined categories—as represented by the human brain—is always a matter of degree. Giving participants additional time allows the representation to settle into a more binary state, but even at the limit, many participants continue to represent some odd numbers as odder than others and these differences manifest as differences in explicit classification tasks. However, an alternative that needs to be ruled out is that people's errors arise from failures of perceptual analysis. Perhaps the parity-classification algorithm is actually free of systematic errors, but the input to it is sometimes faulty. If one assumes that people judge parity by first analytically decomposing the numeral and then applying a context-free algorithm to the last digit then a failure in the decomposition routine would lead to an increase in incorrect responses with increasing *#OPD*, e.g., a number like 798 yields 2 opportunities for feeding the wrong-parity digit into the algorithm whereas inputting any digit of 400 would lead to the same answer. This possibility is systematically tested by Experiments 3A–3C.

In Experiment 3A, participants completed a timed parity-classification task, as in Exp. 1, followed by a flanker congruity task that required perceptual analysis and decomposition similar to the parity-judgment task (Eriksen, 1995; Miller, 1991), but did not require placing stimuli into formal categories. By correlating various aspects of participants' performance on the parity-classification task with their performance on the flanker task, it is possible to test the degree to which perceptual selection failures predict the effect of #OPD.

As shown by Exps. 2A–2B, #OPD continue to affect performance on untimed parity-classification tasks. Exp. 3B seeks to replicate this, and to correlate it with an *untimed* version of a flanker task. If the errors in judging the parity of 798 stem from feeding the wrong digit into the parity-classification algorithm, then such errors should correlate with errors on an untimed flanker task requiring participants to report the color of the last digit while ignoring colors of other digits.

Experiment 3C provides a control to Exp. 3B by using the very same color-coded numerals as Exp. 3B, but now asking people to report parity rather than color. This experiment also asks people for their classifications strategies to determine if the strategy-type predicted performance.

Please cite this article in press as: Lupyan, G. The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition* (2013), http://dx.doi.org/10.1016/j.cognition.2013.08.015

5.1. Experiment 3A: Participants and procedure

Twenty-five undergraduate students participated for credit. Each person completed a flanker congruity task followed by the parity-classification task described in Exp. 1. Each trial began with a fixation cross (700–900 ms) followed by a target and flanker display containing a total of 7 shapes. The targets, always centrally located, were small triangles that faced left (◄) or right (►). On valid trials, the display contained the target flanked by 3 triangles on each side (7 triangles in total). On any one trial, the flankers all faced the same direction. On invalid trials, the target faced opposite the flankers. On neutral trials, the flankers faced upward (▲). Because 'up' is not a possible response, upward facing flankers are predicted to interfere minimally with responding to the direction of the target. The trials were evenly split into the three trial-types (valid, invalid, and neutral). The trials were also evenly split between three delay conditions: simultaneous presentation (standard flanker display), and flanker-first presentations in which the target was presented 150 ms or 500 ms after the flankers. A longer flanker-to-target delay provides additional time in which to selectively attend to the central location in which the target will appear while also inhibiting the representations of the irrelevant flankers and thus should lead to smaller conflict scores. This factor was introduced to explore hypotheses not central to the present paper and will not be discussed in detail. Each participant completed 270 trials.

5.2. Results and discussion of Exp. 3A

5.2.1. Flanker congruity effect

The RT analysis excluded errors trials (2.4%) and trials with RTs were over 1200 ms (0.2%). Overall RTs were 472 ms and accuracy was 97.3%. Not surprisingly, trial-type (neutral, valid, invalid) was a reliable predictor of both RTs, $F(2,48) = 54.39$, $p < .0001$ and accuracy, $F(2,48) = 19.49$, $p < .0001$. People showed an overall RT cost on invalid trials, $M_{\text{invalid}} = 504$ ms; $M_{\text{neutral}} = 464$ ms, $t(24) = 8.13$, $p < .00005$, and an overall advantage for valid relative to neutral trials, $M_{\text{valid}} = 450$ ms; $t(24) = 3.27$, $p = .003$. Accuracy analyses revealed a reliable cost on invalid trials but no advantage for valid trials due to ceiling effects. Due to restricted range of the accuracy analysis, I will focus on RTs, as is conventional for the flanker congruity task.

A linear mixed-effect model analysis that included delay as a covariate showed that the invalidity-cost significantly decreased with longer flanker-to-target onset delays, $t = -6.10$, $p < .0005$, and the validity advantage increased, $t = 5.68$, $p < .0005$ (indeed, the validity advantage was only present for the trials on which the flankers appeared before the target).

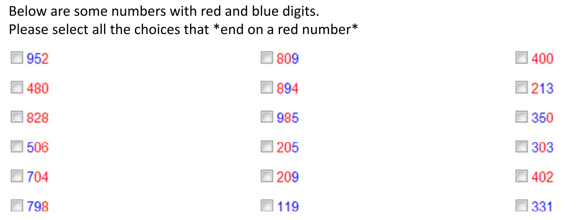
5.2.2. Parity judgments

Overall RTs were 592 ms. and accuracy was 91.5%. This experiment replicated the effects described in Exp. 1, as revealed by a linear-mixed-effect model analysis when predicting RTs and mixed-effect logistic analysis when predicting errors (computed on dichotomous 1/0 values). Controlling for number of digits, increases in #OPD reliably predicted higher RTs, $t = 6.19$, $p < .00005$ and lower accuracy, $z = -3.48$, $p = .0005$. Controlling for #OPD, classifying numerals with more digits took longer, $t = 3.21$, $p = .001$, and was marginally associated with higher accuracy, $z = 1.78$, $p = .07$, though there was no evidence of a by-subject speed-accuracy tradeoff, p's > .2. A direct comparison of the effects of #OPD and digit-number of the present experiment to Experiment 1 showed that the effects of #OPD on RTs was somewhat larger in Exp. 1 than in the present experiment as revealed by a marginally significant experiment-by-#OPD interaction, $t = 1.95$, $p = .05$. A model comparison between a base model predicting RTs from #OPD and digit-number only, to a model including the experiment (Exp. 1 vs. Exp. 3A) and the #OPD-by-experiment interaction showed that the more complex model was not significantly better, $\chi^2(2) = 3.91$, $p = .15$. In short, the present study successfully replicated Exp. 1.

5.2.3. Relationship between the flanker congruity effect and parity-classification

If the effect of #OPD on parity-classification stems solely from selection failures, then controlling for measures of flanker performance that are theoretically linked to such selection difficulties—namely, the difference between neutral and incongruent trials—should account for much of the observed difficulties people have with classifying the parity of numbers like 798 (i.e., the predictive power of #OPD). If, on the other hand, the detrimental effect of increasing #OPD is not purely perceptual, then the predictive power of #OPDs should remain once flanker performance is partialed out. If this were the sole prediction, then support for the hypothesis that effect of #OPD is not due to perceptual-selection failures would require arguing from a null result. However, recall, that in addition to people performing more poorly as a function of #OPD, people were also slower to classify numbers having more digits. That is, controlling for #OPD, people were slower on 4-digit than 3-digit numbers, etc. This effect of number-length may indeed be perceptual in origin insofar as it takes more perceptual processing to isolate the last digit in a 4-digit number than in a 2-digit number. If true, and assuming that the flanker congruity effect measures, among other things, people's ability to perceptually isolate the target in the presence of irrelevant distractors, then controlling for flanker performance should eliminate the effect of digit-number on parity-classification performance.

To test these hypotheses, I compared a series linear-mixed effect models in which parity RTs are predicted from a progressively more complex set of variables. The base model included number of digits, #OPD, and flanker performance on the neutral trials. Both number of digits ($t = 3.08$), and #OPD ($t = 5.96$) continued to predict performance. Performance on the neutral trials did not, $t < 1$. The invalidity-cost and validity-advantage were then added to the model. These variables were not significant predictors of performance, t's < 1, did not improve the overall fit, $\chi^2(2) < 1$, and did not decrease the predictive power of either number-of-digits or #OPD. I next added an interaction between number of digits and the flanker-invalidity cost. Adding this interaction marginally improved the overall fit, $\chi^2(1) = 3.46$, $p = .06$, but importantly, it entirely



Below are some numbers with red and blue digits.
Please select all the choices that *end on a red number*

□ 952
□ 480
□ 828
□ 506
□ 704
□ 798
□ 809
□ 894
□ 985
□ 205
□ 209
□ 119
□ 400
□ 213
□ 350
□ 303
□ 402
□ 331

Fig. 4. Sample trials of the color-flanker task of Exp. 3B. Exp. 3C used identical displays but asked participants to judge parity as in Exp. 2A.

eliminated the predictive power of number-of-digits on parity-judgment RTs (the *t*-value declined from 3.07 to 0.15). Repeating this comparison with the #OPD-by-invalidity-cost interaction showed that the inclusion of the interaction did not improve the model fit, $\chi^2(1) = 1.58$, $p > .2$. The predictive power of #OPD was reduced somewhat; the *t*-value declined from 5.96 to 2.25, but remained a significant predictor, $p = .02$. Adding additional interaction terms involving the validity-advantage did not further improve model fits.

To summarize: Numbers having more digits take longer to be classified as odd or even and this effect is (marginally) increased for people who show a greater flanker invalidity cost. Including the invalidity-cost by digit-number interaction in the model completely eliminated the number-of-digits effect in parity classification. In contrast, the interaction between #OPD and the flanker-invalidity cost was not reliable and the main effect of #OPD remained a significant predictor controlling for flanker performance. This pattern of results suggests that while longer response times for classifying parity of numerals with more digits is linked to perceptual selection performance, the same cannot be said for poorer performance on numbers with opposite-parity digits, like 798.

5.3. Experiment 3B. Participants and procedure

For Exp. 3B, 115 new participants from Mechanical Turk were recruited. Seven were excluded based on the exclusion criteria described in Exp. 2A. The procedure of Exp. 3B was identical to Exp. 2A except that in addition to the original parity-judgment task, participants also completed an untimed color-flanker task requiring perceptual classification and selective attention, but not requiring placement of stimuli into formally-defined categories. This task used the original numbers used in Exp. 2A with each digit colored red if even and blue if odd (Fig. 4). Instead of judging their parity, participants were asked to select all the numbers that ended on a red-colored digit and then, on a separate screen, to select all the numbers ending on a blue-colored digit (order randomized).

5.4. Results and discussion of Exp. 3B

Most participants demonstrated perfect performance: 85% on the color flanker task and 84% on the parity-judgment task. There were no overall differences between the overall error rates on the two tasks, $F < 1$ (a logistic-regression analysis and an analysis excluding all participants with perfect performance also failed to find any differences in overall accuracy between the two conditions). If classification errors of numbers like 798 stemmed from focusing on the wrong (i.e., non-final) digit, then one would expect that the competition between blue and red colors would produce a similar error pattern, i.e., if participants are more likely to say "odd" for 798 because they mistakenly report the parity of the 7 or 9, then when asked to report the color of the final digit (red), they should on occasion report the wrong color (blue) due to selecting the wrong digit.

Although #OPD predicted performance in the parity-judgment task, $z = -2.35$, $p = 0.02$ (replicating the finding from the previous studies), #OPD did *not* predict performance in the color-flanker task, $z = .57$, $p = .57$. This difference in predictiveness of #OPD for the two questions resulted in a reliable interaction between task-type and number-of-opposite parity digits/colors, as shown by a comparison of logistic regression models: $\chi^2(1) = 9.03$, $p = .003$ (i.e., adding the interaction factor significantly improved the model fit). Moreover, although the item-wise error-rates in the present parity-judgment task were significantly correlated with those in Exp. 2A, $r = .59$, $p = .01$, item-wise error rates for strictly perceptual judgments of the color of the final digit did not correlate with the parity-judgment task in Exp. 2A, $r = .16$, $p > .5$, nor with the parity-judgment task performed by the very same participants, $r = -.15$, $p > .5$.[4] Examining subject-wise correlations also failed to find a correlation between the error rates on the two tasks, $r = -.07$, $p > .4$. This independence of errors in the color-flanker and parity-judgment task is unexpected if both error types stem from the same source.

Exp. 3B replicated the pattern of results observed in Exps. 2A–2B while testing the hypothesis that such parity-classification errors observed in Exps. 1–2 stemmed from failures in perceptual selection.[5] The results showed that performance on an untimed "color flanker" task was not dependent on the number of distracting items (different-color digits) demonstrating that in this paradigm, selectively focusing on the last number while ignoring distracting numerals does not pose a challenge. Unlike the errors made in the parity-judgment task, which are systematic, the errors on the color-flanker task appear to be random.

6. Experiment 3C. Comparison of simple selection and formal classification

It is possible that reporting the color of the last digit, as required by Exp. 3B, calls on perceptual selection of a different sort than that required to isolate the last digit in

[4] Using Spearman correlations produced an identical pattern of results.

[5] It is worth noting that insofar as perceptual selection errors are more likely for obliquely oriented numerals in Exp. 1, the perceptual selection account predicts that the effect of #OPD should be greater when numbers are obliquely oriented. Although obliquely oriented numerals indeed yielded longer RTs, the effect of #OPD on performance was non-reliably *smaller* in magnitude for oblique numerals, not larger, both in Exp. 1 and Exp. 3A.

the service of parity classification. Exp. 3C was isomorphic to Exp. 3B: participants saw the very same displays, but instead of being asked to select all the numerals that ended on e.g., a red number, people were asked to select all the odd and all the even numbers, as before. Note that the two experiments required making all the same choices. In Exp. 3C task, as in Exp. 3B, people could simply click on all the numbers that ended with a digit of a particular color. Finding a continued effect of #OPD would provide further evidence that despite being perfectly able to attend to the last digit in the presence of distractors, for many people the allure of similarity-based classification is too strong to shut off.

6.1. Participants and procedure

One-hundred four new participants from Mechanical Turk were recruited. Six were excluded based on the exclusion criteria described in Exp. 2A. The procedure was identical to Exp. 2A except the digits of the numbers were color-coded as shown in Fig. 4. Two people who both know what it means to be odd or even may nevertheless use different strategies for performing the classification. Following the parity-judgment task participants were asked if they relied on the colors of the digits to respond and also asked, "How do you tell if a number is odd or even?" The answers were independently coded by 3 raters on correctness, completeness, and strategy. Strategies were coded as mentioning something about the last digit, mentioning divisibility by 2 with/without remainder, or both. The three raters showed fairly high reliability (Cronbach α for correctness was .89 and for completeness was .87). Disagreements were settled through consultation among the raters.

6.2. Results and discussion of Exp. 3C

Most participants (79%) had perfect performance. As in the previous studies, #OPD was a significant predictor, $z = -6.12$, $p \ll .0001$ (as before, education level did not predict error rates, $z < 1$). A direct comparison between the present study and Exp. 2A failed to find a difference in overall accuracy, $z < 1$, nor an experiment-by-#OPD interaction, $z = -.04$, $p > .9$. However, the interaction between experiment (3B, 3C) and effect of #OPD was highly reliable, $z = -3.59$, $p < .001$. That is, the #OPD coefficient was reliably different between Exps. 3B and 3C. When people's goal was to report the color of the last digit, conflicting colors did not interfere. When the goal was to report the parity of the number, opposite parity (and opposite color) digits did interfere.

At the group level, Exp. 3C replicates once again the basic effect of #OPD. I next examined performance as a function of people's responses to the question of how to distinguish odd and even numbers. Participants with lower correctness had lower overall performance, $z = 2.59$, $p = .009$ (logistic regression), although this relationship was driven entirely by 3 individuals who had clearly incorrect definitions (and overall accuracy of only 74%). When they were excluded, correctness of stated strategy did not predict performance, $z < 1$. The completeness measure also did not predict performance, $z < 1$. Neither correctness nor completeness interacted with #OPD, $z < 1$. For the 52 participants whose strategies were judged perfectly correct and complete, #OPD continued to predict performance, $z = -3.73$, $p < .001$. Participants appeared to largely ignore the color of the digits. Most (87%) responded that they did not use the color of the digits to respond. The response to this question did not predict overall performance nor effect of #OPD, z's < 1.

Next, I examined performance as a function of people's expressed strategy. Of 98 participants, 44 mentioned the last digit (e.g., "the number ending with 2,4,6,8,0 are even numbers, others are odd"), 38 mentioned division ("Even numbers can be divided evenly into groups of two, Odd numbers cannot be divided evenly"), 8 mentioned both, and the remaining 8 provided strategies with insufficient details to be coded (e.g., "odd numbers are in this order 1, 3, 5, 7, 9, etc., even numbers are in this order 2, 4, 6, 8, 10, etc.", "the first number(1) is odd, the second number(2) is even and the numbers alternate from there").

Performance of the 82 people whose stated strategies mentioned either last digits or division was analyzed with logistic regression. Strategy was not a reliable predictor, $z < 1$ and #OPD continued to predict performance, $z < -4.10$, $p < .0001$. Although the interaction between strategy and #OPD was not reliable, $z = 1.43$, $p = .15$, analyzing performance by strategy showed that #OPD predicted performance for people who mentioned using division, $z < -4.22$, $p < .0001$, but not those who mentioned using the last digit, $z = -.30$, $p > .6$. Interestingly, people who mentioned division tended to have descriptions that were more formally correct, $F(1,80) = 4.96$, $p = .029$ and complete, $F(1,80) = 4.53$, $p = .036$, than people who mentioned relying on the last digit. Although it is difficult to draw strong conclusions from this last finding, it is interesting that people who were *more* formally correct in describing how to tell if a number is even or odd were more prone to mis-classify numbers having opposite-parity digits, while people who relied on a convenient heuristic (if the last digit is 1, 3, 5, etc.) were actually less likely to make a mistake.

6.3. Summary of Experiments 3A–3C

Combined, the results of Exps. 3A–3C are difficult to square with the claim that the only reason people misclassify numbers like 798 is that they accidentally focus their attention on the wrong digit. In Exp. 3A, performance on a task designed (in part) to measure perceptual selection predicted the effect of numeral length (e.g., 2 vs. 4 digit numbers) on parity RTs—an effect that is clearly related to the speed with which one can select the last digit—but did not at all account for the slowing of RTs for numbers with opposite-parity digits. Exp. 3B showed that flanker-type incongruity effects go away when time is not a factor. Exp. 3C showed that in the same untimed conditions, with all the same displays, signatures of graded representations of parity do *not* go away.

7. Experiment 4. Performance quirks or meaningful heuristics?

> "Surely those students [the participants of Armstrong et al., 1983] would not have made it into the prestigious University of Pennsylvania, if they really thought that numbers could be more or less odd" (Pinker, 2000, p. 275).

The studies I have presented thus far show that membership in formal categories shows signatures of being graded, even when no time pressure is imposed. Do the higher error rates in classifying 798 mean that people actually think that 400 is more even than 798? Here, I ask participants this very question and then correlate people's responses with their performance on the parity classification task.

7.1. Participants and procedure

Seventy-eight participants were recruited from Amazon Mechanical Turk. Excluding participants using the same criteria as used in Exp. 2A left a final sample of 75. The procedure was identical to Experiment 2A except participants were asked two multiple choice questions to assess their belief about the gradual membership in odd and even numbers:

(a) *True or false: some odd numbers are more odd than others? (True, False).*
(b) *Is 400 more even than 798? (Yes, No, Unsure).*

The questions were presented either before or after the odd/even classification task. If participants answered True or Yes, they were asked to elaborate and provide examples of odd numbers that were odder than others (a) and to explain what made 400 a more even number than 798 (b). The goal of including these questions was threefold: First, I sought to assess, albeit very coarsely, the degree to which participants thought of parity as existing on a continuum. Second, asking about the relative even-ness of two specific number offers an additional test of the hypothesis that people's mis-classifications are due to inattention or perceptual failures (see below for further discussion). Third, I sought to check whether participants who thought e.g., that 400 was more even than 798 were more affected by #OPD overall.

7.2. Results and discussion

The main results were very similar to Exps. 2A–2B. The majority (73%) had perfect performance, but *#OPD* once again reliable predictor of the likelihood of choosing a numeral, controlling for the prompt (odd, even) and the actual parity of the number, $z = -3.72$, $p = .0002$. Self-reported education was again not a reliable predictor of performance, $z < 1$, $p > .7$.

Of the 75 participants, 23 (31%) thought that some odd numbers were odder than others and 21 (28%) reported that 400 was more even than 798; 5 (7%) reported that they were not sure. These 5 participants were removed from the analyses that follow. I next examined the relationship in the responses to the two questions. Of the 21 people who thought that 400 was more even than 798, 13 (62%) thought that some odd numbers were odder than others. Of the 49 people who thought that 400 was *not* more even than 798, 40 (82%) responded 'False' to the proposition that some odd numbers were odder than others. Altogether, 71% of people responded consistently to the two questions, Fisher's exact test, $p < .0005$.

I next examined people's performance on the parity classification task as a function of how they responded to questions (a) and (b). People who thought that some odd numbers were odder than others had lower overall accuracy in judging parity (95% vs. 99%), $z = 1.98$, $p < .05$. People who thought that 400 was more even than 798 had numerically lower accuracy on the parity task, but this difference was not statistically reliable, $z = 1.53$, $p = .13$ (given the correlation between the answers to the two questions, they were not expected to be simultaneously predictive of performance).

Were people who claimed that 400 was more even than 798 more likely to classify 798 as odd or fail to include is among the evens? Of the 49 respondents who thought that 400 and 798 were equally even, only 1 person (2%) made an error mis-classifying 798. Of the 23 respondents who thought 400 was more even than 798, 4 (17%) made misclassification errors, a reliable difference by Fisher's exact test, $p = .035$.

Finally, I examined whether #OPD interacted with people's responses to the 2 typicality questions. This analysis answered the question of whether the performance of people who e.g., thought that 400 was more even than 798 was more affected by #OPD. To test this hypothesis I compared a series of progressively more complex mixed-effects models. A comparison of the base model that included #OPD to a model that included the interaction terms between #OPD and responses to the two questions above showed that the second model accounted was a better predictor, $\chi^2(2) = 7.96$, $p = .02$. Further examination showed that this effect was driven primarily by the interaction between #OPD and the 400-vs-798 question (question *b* above). The accuracy of participants who thought that 400 was more even than 798 was more adversely affected by #OPD than the accuracy of participants who thought that 400 and 798 were equally even, $z = 2.53$, $p = .01$. Put another way: For participants who thought 400 was more even than 798, the difference in classification accuracy between #OPD = 0 and #OPD = 2 was 9.5%. For those who thought that 400 and 798 were equally even, the difference was only 2%, $F(1,69) = 5.64$, $p = .02$.

Recall that in addition to providing discrete responses, participants who indicated that they thought that some odd numbers were odder than others and/or that 400 was a better even number than 798 were asked to explain their answer. Here are twelve sample responses:

(1) "400 is easily divisible by 2".
(2) "It's easier to identify [400] as even than 798.
(3) "By odd numbers being more odd than others, I mean that some odd numbers are harder to identify quickly when compared to other, less odd numbers."

(4) Some numbers like 7 and 3 only have two factors, themselves and 1. Some others like 15 have more than just that.
(5) A three digit number with all odd numbers like 137 seems more odd to me.
(6) "because 4 0 0 is even"
(7) It is an even hundreds – no odd tens like in 798.
(8) I think that a number ending in 5 is less odd than numbers ending in 1, 3, 7, or 9. 389 is more odd than 455.
(9) It seems to be more even because of the double zero, whereas 98 just doesn't seem a even.
(10) "557 is a right example for odd numbers that are more odd than others."
(11) Two 0s and an even number is better than two odd numbers and one even number.
(12) Both are even, as in divisible by 2. But 400 is more even because it is a clean 00 at the end. More even because it's divisible by 10.

As these responses make clear, participants' thinking about parity is informed by a number of heuristics. One of these appears to be ease of categorization (i.e., fluency). Exemplars that are easier to divide by 2 appear to many people as more even. Although it may seem bizarre to claim that some even numbers are more even than others, consider one use of parity in the real world: dividing a group into two teams. As long as the total number is even, the two teams will have equal numbers. Yet does it not seem that two teams of 10 are a cleaner, better, *more even* split than 2 teams of 9? If it does, you just might agree with the statement that 20 is a better even number than 18. One might protest that this is a perversion of the *real*, concept of parity, but that is precisely the point—parity as represented by people, at least some people, does not fully abstract from real world considerations.[6]

Experiment 4 replicates and extends the findings of the studies above. It is shown that, contrary to Armstrong et al. (1983) and Pinker's (2000), a significant number of participants in fact endorse and justify statements regarding relative oddness and even-ness of numbers, and, importantly, these responses were predictive of performance on the parity classification task. People who agreed with statements about the gradedness of parity were more affected by #OPD when making parity judgments. These effects are themselves graded in nature. Someone who says that 400 is a better even number than 798 does not (in all likelihood) have the "wrong" definition of parity. If they were simply misinformed, their performance would be far worse than what was actually observed. Rather, as was true of the participants in the previous studies, their error patterns were affected by "surface" properties to a greater extent, but were still graded and probabilistic.

8. Experiment 5A: When is a triangle not a triangle?

The results of Exps. 1–4 show that participants make systematic errors when judging numerical parity. As shown by Exps. 3A–3C these errors cannot be easily attributed to failures of perceptual selection. An even stronger case for the claim that classification into formal categories is inherently input-dependent can be made by showing that participants make systematic classification errors even when correct classification does not require decomposing the input into parts like last digit/other digits. In Exps. 5A–5D participants are asked to classify 2-dimensional shapes as triangles. These experiments ask whether triangles that are generally classified as atypical/non-canonical are actually less likely to be classified as triangles altogether. These results cannot be attributed to differences in decomposition/perceptual selection because whatever decomposition is required for classifying a shape as a triangle, is required equally for all types of triangles—isosceles, scalene, etc. If classification errors are due to perceptual selection, then they should occur equally often for all triangle types. In contrast, if triangle classification is invariably affected by triangle typicality, then some triangles—the more typical or canonical ones—ought to be classified as triangles at higher rates than those that are less typical/canonical.

8.1. Participants, materials and procedure

One-hundred four participants were recruited from Mechanical Turk; 8 were excluded based on criteria analogous to Exp. 2A. Participants were shown 10 triangles comprising 4 sub-types (equilateral, isosceles, right, and scalene; Fig. 5) and asked to "select all triangles" by clicking checkboxes next to each shape. A separate group of 50 participants (1 excluded) were asked to rank the typicality of each triangle on a 1–4 scale (very atypical to very typical). All participants saw the same shapes shown on a single webpage in random order. There was no time pressure.

8.2. Results and discussion

Although 77% of participants performed perfectly in this untimed test, errors were strongly modulated by triangle sub-type, $F(3,285) = 13.11$, $p < .0005$ (Fig. 5A and B). A model comparison showed that adding the sub-type predictor substantially improved the fit, $\chi^2(3) = 102.8$, $p \ll .00001$. Logistic regression analyses revealed that equilateral triangles were classified more accurately than isosceles, followed by right, and scalene. All differences in error rates except between right and scalene triangles were highly reliable, p's < .001 (see also Ward, 2004 for an account of similar errors in interviews with pre-service teachers). Education did not predict accuracy or interact with the effect of sub-type.

Participants overwhelmingly rated some triangles as more typical than others, $F(9,432) = 37.60$, $p \ll .0001$: equilateral > isosceles > right > scalene (all contrasts reliable). Canonically oriented triangles were significantly more typical than obliquely oriented triangles (with the

[6] In addition to the parity of a number, one can also define the 2-adic order also referred to as valuation of even numbers. Numbers which can be divided by 2 only once (e.g., 798) are called singly-even, and numbers that can be divided by 2 multiple times (e.g., 400) are called doubly-even. Valuation indeed predicted error rates in Exp. 1 and 3A. However, when controlling for #OPD, the predictive power of valuation disappeared, p's > .5 suggesting that it is not simply the case that participants "definition" of parity explicitly includes valuation.

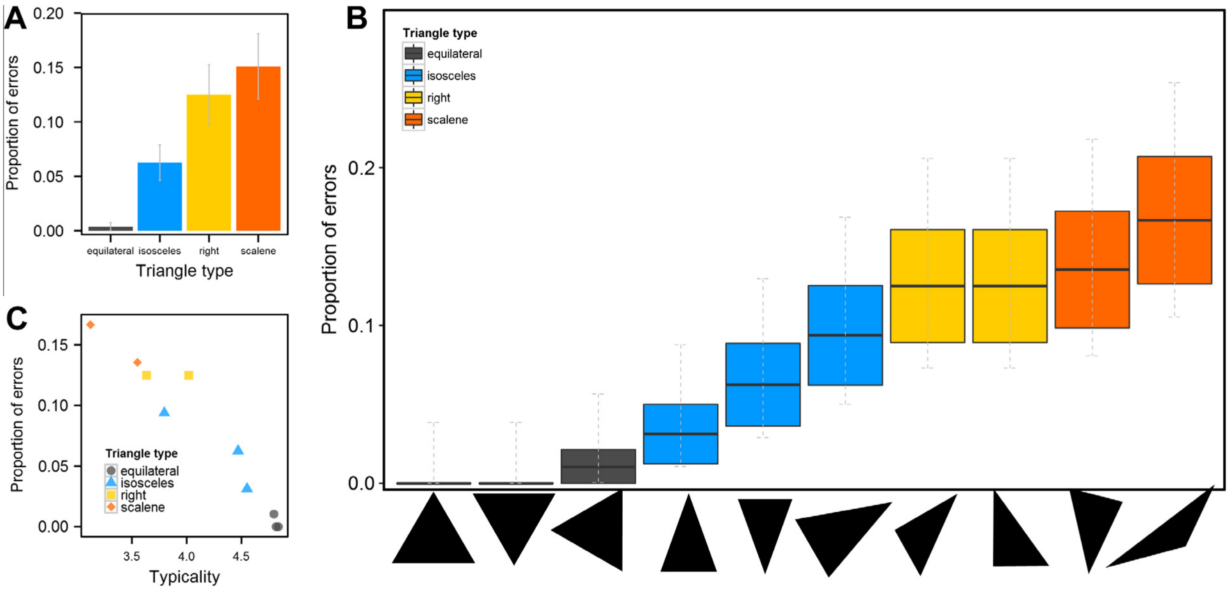


Fig. 5. Results of Exp. 5A. Panel A shows mean errors for each triangle type (±1 SE). Panel B shows the proportion of failures to select each given triangle, with the shapes ordered by mean accuracy. Colors indicate type of triangle. Each box contains the mean and ±1 SE. Errors bars contain the 95% binomial CI. Panel C shows the relationship between error rates and typicality ratings, provided by a separate group of participants. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

exception of "upside-down" triangles: triangles 2 and 5 depicted in Fig. 5 were as typical as upright triangles). As shown in Fig. 5C, typicality accounted for almost the entirety of the error pattern, $r = -.97$, $p < .0005$. The triangles that participants in one group judged to be atypical were the same ones that participants in another group (occasionally) failed to select as "triangles."

These results provide additional evidence of the inherent input sensitivity of the human algorithms that compute membership in formal categories. Unlike judging numerical parity which possibly requires analytic decomposition of the number into its digits, no such analytic decomposition is necessary for classifying shapes as triangles. Any errors stemming from perceptual mis-analysis would presumably impact all triangles equally, but participants were ~2.5 times less likely to classify a scalene triangle as a triangle compared to an isosceles one.

Because all the shapes were triangles, perfect performance required participants to select every option, giving this task flavor similar to a multiple choice exam on which the correct answer is 'c' for suspiciously many questions in a row. It might therefore be objected that the participants' failure to choose all the triangles was pragmatic in nature. This concern is addressed by the next study. To foreshadow the results, the error patterns are unchanged by including non-triangles among the choices.

9. Experiment 5B: Are classification errors in Exp. 5A due to odd question pragmatics?

Experiment 5A showed that many participants systematically failed to include non-canonical triangles as triangles. Correct performance required participants to choose *all* the options. A possibility remains that some individuals understood the question to mean "choose all the *typical* triangles." Although such a (mis) interpretation should not be logically possible on the classical account, it is nevertheless an alternative to my claim that some participants think that, e.g., equilateral triangles are "true" triangles, while scalene triangles are not due to their deviation from a "canonical" triangle. To rule out the possibility that the results of Exp. 5A were due to question pragmatics, I conducted a replication with several non-triangles included among the choices.

9.1. Participants, materials and procedure

Eighty-three participants were recruited from Mechanical Turk; 3 were excluded, as per criteria of Exp. 2A. As in Exp. 5A, participants were asked to click on all the shapes that were triangles. The one differences was that that two additional shapes were now included: a square positioned at 45° off horizontal and an obliquely-presented rectangle.

9.2. Results and discussion

None of the participants selected the square or rectangle and about 78% classified all the remaining shapes as triangles. Once again, which triangles were omitted was strongly dependent on the type of triangle (equilateral, isosceles, right, scalene), $F(3,237) = 8.30$, $p < .0005$. A model comparison showed that adding the sub-type predictor substantially improved the fit $\chi^2(3) = 63.82$, $p \ll .00001$. Education was not a reliable predictor of overall performance, $z < 1$ and did not reliably interact with the classification profile, $p > .15$; however people who reported being only high school graduates were marginally less likely to select right triangles, $z = 1.84$, $p = .07$ and scalene triangles, $z = -1.88$, $p = .06$.

There were no differences in overall accuracy between Experiments 5A and 5B, $F < 1$ and the experiment × triangle sub-type interaction was not significant, $F < 1$. As in Exp. 5A, the likelihood of classifying a particular triangles as a triangle was almost entirely predicted by the typicality measure collected in Exp. 5A, $r = -.94$, $p < .0005$.

This study tested the possibility that the reason that participants failed to select all the triangles in Exp. 5A was due to question pragmatics. One may wonder how fragile a putative rule-based categorization process must be, to be so easily confounded by such pragmatic considerations, but it is nevertheless conceivable that being faced with a question to which the correct answer is to choose all the options leads participants re-interpret the original question. The present study shows that including non-triangles among the choices with a new group of participants led to virtually identical results. Indeed, the correlation between error-rates for the 10 triangles between Exps. 5A and 5B was .96, $p < .0005$. The similarity of the results between Exps. 5A and 5B (see also 5C) speak against the possibility that the results of Exp. 5A were in some way artifacts of question pragmatics.

10. Experiment 5C. Are individuals consistent in their classification performance?

It is admittedly surprising that a sizeable minority of educated adults fail to classify non-canonical triangles as triangles. This experiment tests the stability of these classification decisions by asking people to classify the triangles twice and to justify their choices. If people's classification decisions vary widely from one minute to the next, or if the justifications include such explanations as "the lines did not look straight" or "I thought I did choose these," then there is reason to doubt the validity of the results.

10.1. Participants, materials and procedure

Eighty-seven participants via Amazon Mechanical Turk completed the task; 2 were excluded for selecting one of the rectangle shapes. The procedure was identical to Exp. 5B except that immediately after making their selections, anyone who did not choose all the triangles was shown the omitted triangles and asked "Why did you not select this shape/these shapes?" All participants were also asked to select the triangles a second time with the only difference between the two questions being the order in which the shapes were arranged on the screen.

10.2. Results and discussion

Overall, 85% of participants correctly selected all the triangles. Once again, which triangles were omitted was strongly dependent on the type of triangle (equilateral, isosceles, right, scalene), $F(3,252) = 9.76$, $p < .0005$. A model comparison showed that adding the sub-type predictor substantially improved the fit $\chi^2(3) = 122.08$, $p \ll .0001$. Education was not a reliable predictor of overall performance and did not reliably interact with the classification profile, $z < 1$.

A comparison of errors for the first and second question showed that performance was higher overall the second time they were asked to classify the triangles, ($M = 93\%$ vs. $M = 95\%$). This difference was reliable as shown by comparison of logistic models, $\chi^2(1) = 7.77$, $p = .005$. But importantly, the categorization profile was nearly identical for the two questions. The by-item correlation between the two questions was .97 and by-subject correlation was .84. The responses were also once again strongly predicted by typicality. Correlations between typicality and errors for the 10 different triangles $-.92$ for the first classification question and again $-.92$ for the second question. As suggested by these extremely high correlations, question number (first or second) did not reliably interact with triangle sub-type, $z < 1$.

Finally, it is useful to examine the types of responses people gave when asked to explain why they did not choose certain triangles. Here are some representative justifications:

(1) They aren't true triangles.
(2) Because it is not triangle shaped.
(3) Because a triangle has three equal sides.
(4) Two sides are not equal.
(5) These are not triangles.

Only one person's explanation referred to any sort of (in)attentional factor. This person wrote "I meant to. I guess my mouse malfunctioned" and proceeded to select all the triangles the second time around. A more formal analysis of the free responses would require a much larger number of participants and is beyond the scope of the present work. What the responses make clear, however, is that failures to select all the triangles are not simple oversights and reflect a consistent tendency of (some) participants to mis-classify non-canonical triangles at a higher rate. More generally, these results offer further support for the claim that mental representations of formal categories have a graded structure. If a critic wishes to argue that these results *also* indicate failures of an *identification* procedure as distinct from a putative core, they would need to provide a coherent account of: (1) why, in an untimed categorization task, repeated twice, this core representation is not being "triggered" by all triangles[7] and (2) what is added by positing a conceptual core.

[7] One may wonder whether participants who persist in failing to select non-canonical triangles simply have a different "non-standard" definition of triangle, one that categorically excludes e.g., scalene triangles. It is unclear how this account can be convincingly supported. When asked to provide a definition of a triangle, virtually all participants respond with a version of "a three-sided shape/polygon/figure" without mentioning additional properties. All the reported effects remained significant after excluding participants whose justifications suggested a non-standard definition. At the same time, it is clear that the definition of "triangle" is itself context-sensitive. If it were not, the phrase "upside-down triangle" would be meaningless, yet it is not. Such context-dependence makes judging whether someone has the "correct" definition a less than meaningful enterprise.

Please cite this article in press as: Lupyan, G. The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition* (2013), http://dx.doi.org/10.1016/j.cognition.2013.08.015

11. Experiment 5D. The consequences of graded representations for making inferences

So far, I have shown that participants make systematic errors in tasks requiring explicit classification of items into formally-defined categories. However, categorization is not an end in itself. A major function of categorization is to promote inference (e.g., Markman & Ross, 2003). If we know that property *P* is true of all members of category *C*, then classifying an item as a member of *C*, allows us to infer that it has *P*. Insofar as participants fail to classify certain "non-canonical" triangles as triangles, they may similarly fail to extend properties that are mathematically true of all triangles to these non-canonical triangles. Exp. 6 tested this hypothesis.

11.1. Participants, materials and procedure

Sixty-five participants were recruited from Mechanical Turk. The procedure was almost identical to Experiment 5A, but instead of being asked to overtly classify the triangles, people performed a category inference task. The instructions informed them that "All triangles have angles that add up to 180 degrees" and asked to "select the shapes below that you think have angles adding up to 180 degrees." The shapes were presented to participants in the identical way as in Exp. 5A.

11.2. Results and discussion

Seventy-five percent of participants correctly selected all the triangles as having angles that added to 180°. Errors were strongly modulated by triangle type, $F(3,192) = 6.48$, $p < .0005$. A model comparison showed that adding the sub-type predictor substantially improved the fit $\chi^2(3) = 79.92$, $p \ll .00001$.

Once again, failure to choose individual triangles was strongly predicted by their typicality, $r = -.88$, $p = .001$ (Fig. 6). The one notable difference from Exps. 5A–5B was that the two right triangles were now selected the least often—surprising given their common use in textbooks to illustrate geometric and trigonometric principles. Participants were additionally asked to explain why they failed to select certain shapes. Here are some typical responses from participants who omitted at least one shape:

1. "Well they didn't seem to be quite perfect, I guess".
2. "Because they did not have even sides".
3. "They didn't look like they added to 180 degrees".
4. "They didn't look like they were 180 degrees".
5. "Cause one of them looked like it could have been more than 180 degrees".

12. Experiment 6A. Oddness, triangleness and grandmotherhood: individual differences and further explorations

> "Prototypical grandmothers are women with gray hair, they have wrinkled skin, they wear glasses, and so on. Yet we all know that there are people who fail to exhibit these characteristics who are grandmothers, and that there are people who do exhibit these characteristics who are not. Mrs. Doubtfire (the Robin Williams character) may look like a grandmother, but Tina Turner really is a grandmother (Margolis & Laurence, 2008, p. 196, chap. 8).

The key claim being made across the 11 studies presented so far is that despite people's explicit knowledge of the rules of membership in formal categories like ODD and TRIANGLE, all participants showed signatures of similarity-based processing in speeded classification and a sizeable minority continued to show them in unspeeded classification. This basic finding might be taken to mean that people's concepts come in two forms: one form has fuzzy boundaries and results in prototype effects, while the other has sharp boundaries with all-or-none membership. Exactly this argument has been made on multiple occasions for the concept GRANDMOTHER. As Margolis and Laurence describe, there are "prototypical" grandmothers who share certain characteristics, and then there are the *true* grandmothers, who, by virtue of being women with at least one grandchild, are all *equally* valid instances of the GRANDMOTHER category despite possibly lacking the "peripheral" grandmother features such as gray hair. For example, Pinker writes, "Family resemblance categories are real, but so are classical categories; they live side by side in people's minds, as two ways of construing the world" (Pinker, 2000, p. 275) and "People are not slaves to similarity. We can be told [that] Tina Turner is a grandmother, overriding our statistical experience of what [grandmothers] tend to look like. This suggests an ability to summarize an entire category by a mental variable or symbol, whose meaning comes from the rules it enters into" (Pinker, 1999, p. 41).

The alternative position advanced here is that while it is true that people have the *capacity* for performing classification based on abstract rules (otherwise how could I even write about the difference between rule-based and similarity-based grandmotherhood?), it is wrong to assume that core aspects of human cognition are based on such rule-based computations (as in e.g., Fodor, 1983, 2001; Gallistel & King, 2009; Marcus, 1999; Pinker, 2000). If the algorithms humans deploy in the service of symbolic computation cannot be trusted to produce unfuzzy representations of odd numbers, triangles, and perhaps even grandmothers, can they really be trusted to do the heavy lifting of moment-to-moment cognition if such cognition relies on symbolic computation?

In experiment 6A, I ask two specific questions: First, do people's representations of the highly familiar concept GRANDMOTHER remain graded even in a context that demands decidedly ungraded representations? Second, do people who demonstrate more conceptual gradedness in the domain of grandmothers also more likely to show gradedness in other domains?

12.1. Participants, materials and procedure

Eighty-three participants were recruited from Mechanical Turk, all from the United States. Fifty participated in

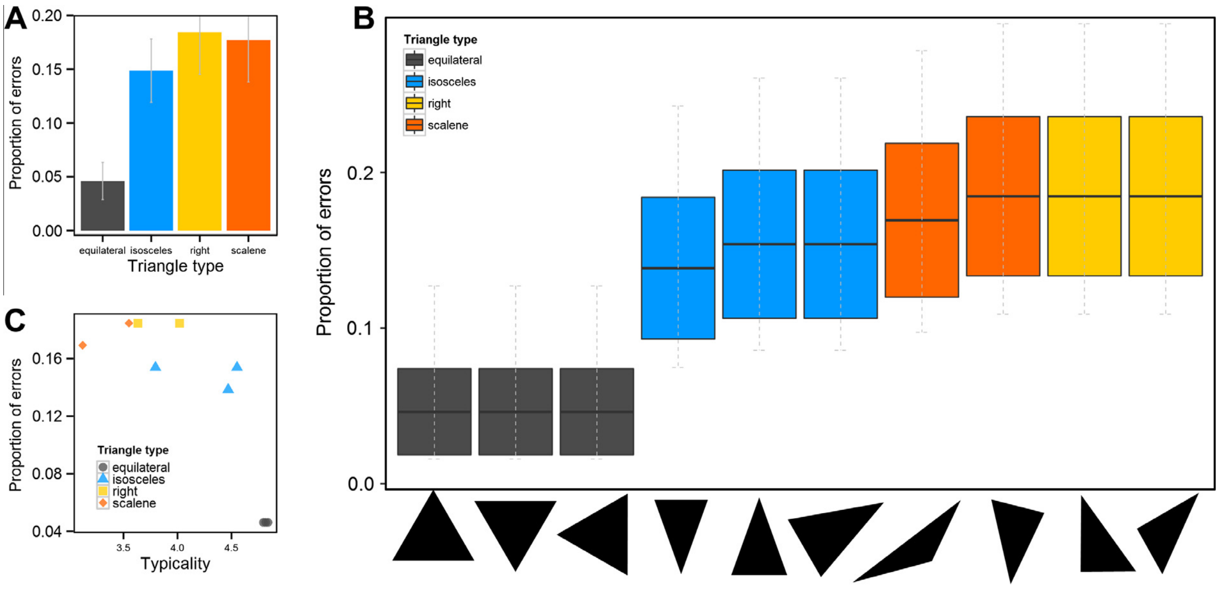


Fig. 6. Results of Exp. 5D. Panel A shows mean errors for each triangle type (±1 SE). Panel B shows the proportion of failures to select each given triangle, with the shapes ordered by mean accuracy. Colors indicate type of triangle. Each box contains the mean and ±1 SE. Errors bars contain the 95% binomial CI. Panel C shows the relationship between error rates and the typicality ratings collected as part of Exp. 5A. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the main experiment. Each participant was asked three questions:

12.1.1. Graded even numbers

> "Is 400 a better even number than 798?" (Yes, No, Unsure)

12.1.2. Graded triangles

> "Are some triangles better than others?" (Yes, No)

Anyone who answered 'yes' was asked a follow-up question asking to explain their answer.

12.1.3. Graded grandmothers

Participants were asked to read the following scenario and respond to the best of their ability.

> "A magazine is running a contest in which they award $100 gift certificates to eligible contestants. To be eligible, you have to be a grandmother. That is the only criterion. The decisions are made completely randomly. All eligible contestants have a 25% chance of winning. Please indicate the likelihood of each person winning the contest, from 0% to 100%"

This prompt was followed by the 10 choices shown below presented in random order. Next to each question was a text box that accepted a number from 0 to 100, with 0 as the default.

- A 24 year old man with no kids.
- A 68 year old man with three adult children and 6 grandchildren.
- A 27 year old woman with no kids.
- A 43 year old woman with two children, aged 11 and 10.
- A 39 year old woman whose daughter just had a baby.
- A 41 year old woman with three grandkids.
- A 59 year old woman with one daughter who recently gave birth to twins.
- A 66 year old woman with 6 kids, but no grandkids.
- A 64 year old woman with 3 sons and 2 granddaughters.
- A 68 year old woman with 2 grandsons and 4 granddaughters.

Following these questions, people were asked to provide demographic information including the highest level of education they completed. A separate group of 33 people were shown the descriptions of people above and asked to rate each on a 1–5 scale from "Not a grandmother" to "Very typical grandmother"

12.2. Results of Exp. 6A

12.2.1. Graded even numbers

Seven people responded 'yes', 38 'no', and 5 were not sure. Of the 45 people who were confident enough in their answer to answer 'yes' or 'no', 16% thought that 400 was a better even number than 798; 95% CI = 6–29%. As in Exp. 4, people who answered 'yes' to this question gave answers that referenced the roundness of 400 and its greater number of factors. Here are three characteristic examples:

1. "While 798 is technically even, 400 is also a perfect square and has more ways that it can be divided".
2. "It ends in double zeros".
3. "Because 400 is comprised of all even numbers while 798 contains two odd numbers".

Table 3
Main results of Exp. 6A.

Contestant	Age	Gender	Num. grand-children	Mean response	95% CI
A 24 year old man with no kids	24	M	0	0.00	–
A 27 year old woman with no kids	27	F	0	0.00	–
A 43 year old woman with two children, aged 11 and 10	43	F	0	0.90	0–1.96%
A 66 year old woman with 6 kids, but no grandkids	66	F	0	1.00	0–2.21%
A 68 year old man with three adult children and 6 grandchildren	68	M	6	6.70	4.90–8.50%
A 39 year old woman whose daughter just had a baby	39	F	1	22.80	21.72–23.88%
A 59 year old woman with one daughter who recently gave birth to twins	59	F	2	24.66	23.68–25.64%
A 41 year old woman with three grandkids	41	F	3	25.20	24.20–26.20%
A 64 year old woman with 3 sons and 2 grand-daughters	64	F	2	26.90	25.90–37.90%
A 68 year old woman with 2 grandsons and 4 grand-daughters	68	F	6	27.00	25.90–28.10%

12.2.2. Graded triangles

A similar percentage of people responded saying that some triangles were more typical than others ($n = 7$), 14%; 95% CI = 6–27%. When asked to explain their choice, participants tended to mention ratios of sides. Here are two examples:

1. "An equilateral triangle, in my mind, is more "triangular" than an extremely obtuse triangle".
2. "Those perfect sexy equilateral triangles are the most trianglest".

12.2.3. Graded grandmothers

Each mentioned contestant was coded on three variables: gender, age, and the number of explicitly mentioned grandchildren (Table 3). For one contestant, the number of grandchildren was deliberately omitted. People were expected to infer that the 43 year old woman with 2 kids aged 10 and 11 was not a grandmother (although, logically she could have had an unmentioned grandchild). The summary statistics for each choice sorted by mean value are presented in Table 3.

The responses were analyzed using linear mixed effects models. The first analysis tested whether the gender of the "contestant" and their number of grandchildren predicted their likelihood of winning. As evident from Table 3, the answer is yes. Every subject made a sharp distinction between men and women, and grandparents and non-grandparents, $t > 10$, $p \ll .0001$. This result serves as a check that participants read and understood the question.[8] All subsequent analyses include only the 5 genuine grandmothers.

First, I tested the hypothesis that contestant age and the *number* of grandchildren were predictive of how likely people thought each contestant was of winning, *despite the instructions stating that all eligible contestants had an equal chance*. To test this hypothesis, I compared a series of linear mixed effect models predicting likelihood of winning from the number of grandchildren and age (Table 3). Adding number of grandchildren significantly improved model fit compared to an intercept-only model, $\chi^2(4) = 9.18$, $p = .002$. Adding age further improved the fit, $\chi^2(5) = 5.46$, $p = .02$. These effects are visualized in Fig. 7.[9]

Although most people's responses were dichotomous (0 or 25), some people input other values. Overall, 29 responses (5.8%) were above 0, but less than 25 ($M = 18.45\%$) and 16 responses (3.2%) were above 25% ($M = 60.8\%$). Of the 50 participants, 10 had at least one response with a value other than 0 or 25. One may protest that anyone who input a number other than 0 or 25 did not understand the basic premise of the question that eligible contestants had a 25% likelihood of winning. Perhaps the graded nature of the grandmother concept revealed by the analysis above is an artifact, being driven exclusively by these participants who grossly misunderstood the question. Perhaps they thought grandmothers with more grandchildren could enter the contest numerous times.

A new indicator variable *nonDiscreteness* was set to 1 for anyone who had at least one response other than 0 or 25. A linear mixed effect model analysis showed a reliable positive interaction between *nonDiscreteness* and contestant-age, $t = 2.47$, $p = .01$ showing that responses of people who were non-discrete in their responses were more affected by contestant age. However, the inclusion of age and number of grandchildren reliably improved model fits *both* for the 40 discrete responders (those who responded with only 0 and 25), $\chi^2(2) = 7.13$, $p = .03$, and for the 10 non-discrete responders, $\chi^2(2) = 8.15$, $p = .02$.

Recall that a separate group of participants viewed the descriptions of the 10 people in Table 3 and were asked to rate how typical each was of a grandmother ("Not a grandmother" was one of the options). Unsurprisingly, people distinguished between grandmothers and non-grandmothers as shown by a reliable improvement in fit when an is-a-grandmother dichotomous predictor is used to predict typicality, $\chi^2(1) = 1112.92$, $p = .02$, $p \ll .0001$. Also not especially surprising is that the model fit is further improved when age, $\chi^2(1) = 394.78$, $p \ll .0001$, and number of grandchildren, $\chi^2(1) = 12.79$, $p = .0003$, are added.

[8] Though note that some people (11/50)—identified the 68 year-old grandfather of 6 as eligible with likelihoods from 25% ($n = 9$), 30% ($n = 1$), and 80% ($n = 1$). The effects of age and number of grandchildren were independently significant both for people who made this (presumably inattentional) error and those who did not.

[9] The description of some eligible contestants explicitly used the word grandchild/grandson/granddaughter and these were (inadvertantly) correlated with contestant age. An additional model using age, number of grandchildren, and an indicator variable coding for explicit use of one of these words showed that explicit use of the grandchild/grandson/granddaughter was indeed correlated with higher likelihoods, $t = 2.22$, $p = .02$, but age continued to be a additional reliable predictor, $t = 2.32$, $p = .02$. In this analysis as in the ones above only eligible contestants, i.e., the grandmothers, were included.

Please cite this article in press as: Lupyan, G. The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition* (2013), http://dx.doi.org/10.1016/j.cognition.2013.08.015

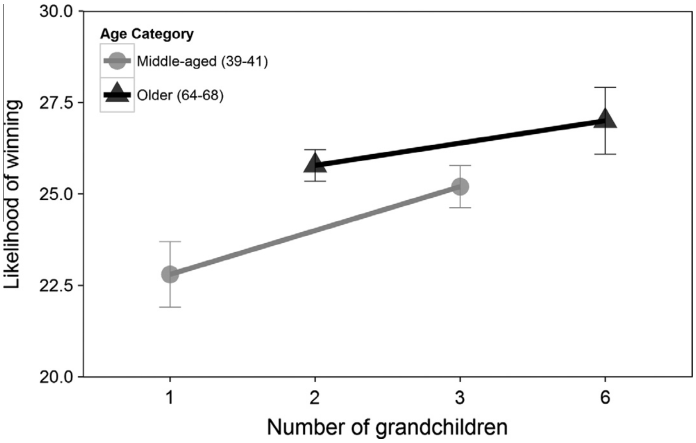


Fig. 7. The mean rating of winning by grandmothers of different ages and having different number of grandchildren. Error bars indicate the within-subject 95% CI of the mean.

Importantly, controlling for grandmother status, typicality predicted the likelihood of winning, $\chi^2(1) = 16.35$, $p < .0001$. More typical grandmothers were judged to be more likely to win.

We see here that even in a context that calls for categorical responses, typicality continues to play a role, just as it did in classifying polygons as triangles. Indeed, *within* the grandmother category, the correlation between mean likelihood of winning and mean typicality was .81.

12.2.4. Relationships between responses

An important question that this study allows us to ask is whether the responses of the 50 participants to these 3 questions were related in any way. One possibility is that people's responses reflect content knowledge only. If true, we might expect the responses to the even number question to be unrelated to responses to the triangle question, and also unrelated to people's concept of grandmothers. A second possibility is that people who tend to discretize formally defined concepts like parity are more likely to discretize other formally defined concepts. Of the 7 people who responded 'yes' to the graded even number question, 3 responded 'yes' to the graded triangle question (43%). Of the 38 people who responded 'no' to the graded even number question, only 2 (5%) responded 'yes' to the graded triangle question. Fisher's exact test showed that these proportions were reliably different, $p = .02$, odds ratio = .08. A linear mixed-effects logistic regression showed that a model predicting the response to the graded even number question from education and the response to the triangle question was reliably better than a model that included education only, $\chi^2(1) = 6.88$, $p = .009$.[10]

In the final analysis, the responses to the even-number and triangle questions were used to predict the responses to the grandmother question. Insofar as all these questions reflect the degree of gradedness of concepts despite clear necessary and sufficient conditions, people who believe that 400 is a better even number than 798 and/or that some triangles are better than others may also represent grandmothers in a more graded manner and thus be more affected by putatively irrelevant information like age and number of grandchildren.

The base model predicted the likelihood of winning (for grandmothers only) from participant's education, contestant age, number of grandchildren, as well as the response to the graded even number question and response to the graded-triangle question. This model was compared to a model that included interactions between the answers to the graded even number and graded triangle question and the variables coding contestant age and number of grandchildren. The latter model was a better fit to the data, $\chi^2(6) = 24.91$, $p < .001$. In particular, participants who thought that 400 was a better even number than 798 were more influenced by the age of the contestant than participants who thought 400 and 798 were equally even, $t = 3.81$, $p = .0001$. The interaction between the triangle-question and age/number of grandchildren variables was not independently reliable.

13. Experiment 6B. Controlling for anchoring effects

In Exp. 6A, the grandmothers judged to be more likely to win were ones that were older and had more grandchildren. It is conceivable that these results reflect an anchoring effect caused by seeing higher numbers as part of the contestant description (Daniel Casasanto, pers. comm.). Anchoring effects of this sort are well-known. For example, Tversky and Kahneman (1974) showed that random numbers (0–100) generated by a "wheel of fortune" apparatus in the subject's presence reliably affected those subjects' estimates of the percentage of African countries in the United Nations: the median responses were 25 and 45 for the groups that received starting random values of 10 and 65, respectively (see also Ariely, Loewenstein, & Prelec, 2003). Although such effects operate in uncertain domains and people *should* be a whole lot more certain of whether a woman with grandchildren is really a

[10] Education was not a reliable predictor in any of the analyses and its inclusion does not change any of the results.

grandmother than how many African countries are in the UN, perhaps the results of Exp. 6A are an artifact of anchoring nonetheless.

13.1. Participants, materials and procedure

Fifty-three people from Amazon Mechanical Turk were recruited. Three participants was eliminated for indicating in the comments that they did not understand the question. The procedure and materials were identical to Exp. 6A except that the criterion for the eligibility in the contest was altered: Rather than only grandmothers being eligible, participants were told that only women under 60 were eligible. As in Exp. 6A, this criterion yielded 5 eligible and 5 ineligible contestants. If anchoring effects rather than the eligibility instructions are responsible for the effects observed in Exp. 6A of graded likelihood within eligible contestants, then participants should continue to rate older eligible female contestants as more likely to win than younger eligible ones. If, on the other hand, the eligibility criterion is responsible for the finding in Exp. 6A, then age might now be negatively correlated with perceived eligibility. At the very least we should find a significant age-by-experiment interaction when examining the likelihood of winning within the eligible group.

13.2. Results of Exp. 6B

Just as in Exp. 6A, participants drew a sharp distinction between eligible and ineligible contestants. The mean likelihood of winning for ineligible individuals was 1.12% and the mean likelihood of winning for eligible entrants was 25.04%, $F(1,49) > 1000$. Given that eligibility was now inversely correlated with contestant age, it was, unsurprisingly, a negative predictor of the likelihood of winning, $b = -.44$, $t = -14.2$, $p \ll .0001$. The critical question, however, was whether anchoring effects led people to rate some eligible contestants as more likely to win than other eligible contestants.

Because the critical tests concerned effects for the eligible contestants only, the analyses that follow include just the eligible contestants: the 5 grandmothers in Exp. 6A and the 5 females under 60 in Exp. 6B. The first analysis tested whether contestant age was differentially predictive of likelihood of winning in Exps. 6A and 6B. A mixed-effects model analysis showed that this interaction was highly reliable, $t = 3.44$, $p = .0006$. Recall that age was positively correlated with winning likelihood in Exp. 6A, $b = .11$. This was not the case for Exp. 6B, $b = -.03$. The interaction between number-of-grandchildren and experiment was likewise reliable, $t = 3.06$, $p = .002$. While number-of-grandchildren predicted greater likelihood of winning in Exp. 6A, $b = .62$, this was no longer true for Exp. 6B, $b = -.33$, $p = .11$. The negative coefficients indicate that rated likelihood was winning was *negatively* associated with age and number of grandchildren. So, for example, the 59 year old female was now rated as numerically *less* likely to win than the 24 year old female ($M_{59\ \text{year old}} = 24.5\%$; $M_{24\ \text{year-old}} = 25.3\%$), although not significantly so. In sum, the results of Exp. 6B make it unlikely that the findings of Exp. 6A stem from anchoring effects.[11]

It may be objected that the present account predicts that the "females under 60" instruction should have led people to rate female contestants just under the age cut-off as being less likely to win (because they are *almost* ineligible) than younger females who are further from the boundary. Although there was a trend in this direction, it was not reliable. "Women under 60" is a purely ad hoc criterion and larger effects may be obtained for categories with more salient examples of good membership. For example, being a teenager, even when formally defined based on an age cut-off, has rich semantic connotations that go beyond age. A precocious 12-year old may be classified as a teenager even when it is made clear that teenager status is to be granted only to 13–19-year olds.

13.3. Discussion of Experiments 6A and 6B

Whatever confusion people may have about what makes a number even or a polygon a triangle, everyone knows what makes someone a grandmother. Yet, when placed in a context that required simply to distinguish between grandmothers and non-grandmothers, a sizeable minority once again relied on an apparently graded representation according to which having more grandchildren makes one a "better" grandmother and thus apparently more likely to win a contest for which grandmothers are eligible. The results of Exp. 6B are inconsistent with the possibility that the findings of Exp. 6A were due to an anchoring effect.

Let us pause to consider why it should even make sense to talk about typical and atypical grandmothers. An obvious answer is that the properties that make someone a *typical* grandmother are correlated with *actually being* a grandmother. Moreover, under normal circumstances perceptual cues such as gray hair, wrinkles, and baking muffins, are much more readily available than the formal definition of "has at least one grandchild" such that even when placed in contexts where such "peripheral" information should not be relevant, people continue to rely on it (i.e., the representations used in the task encode these details).

To reiterate: My claim is not that people are *incapable* of following rules or of forming highly abstract representations. Given enough time, most of the subjects tested here appear to do just that. However, the overall pattern of results strongly suggests that this process of abstraction is far from trivial. Many people continue to rely on similarity-based representations that integrate probabilistic cues suggesting them to be a kind of default state (exactly as

[11] In another version of this experiment in which the instructions stated that all participants were equally likely to win, no contestants were rated as being significantly more likely to win than others, $\chi^2(9) = 8.42$, $p = .49$ (model comparison with/without contestant as a predictor), and neither age nor number of grandchildren predicted the likelihood of winning, a result that also speaks against anchoring as an explanation for the results of Exp. 6A.

conceived in connectionist models, e.g., Bybee & McClelland, 2005; Elman, 2009; McClelland & Rumelhart, 1981).

What of the claim that people possess *two* GRANDMOTHER concepts (and presumably two ODD concepts, two TRIANGLE concepts, etc.)—one graded and perceptually-based and one conforming to classical definitions? Such an interpretation meets with two serious objections. First, what are the conditions necessary to trigger the rule-based representation? After all, participants in Exp. 4 and 6A were not asked whether 400 *looks* more even than 798. They were asked whether it *is* more even than 798 (and additionally asked to justify their answer). Participants were explicitly told that all grandmothers had an equal chance of winning, yet they allowed typicality to override this constraint. If it is true that "our own kinship system gives us a crisp version of 'grandmother': the mother of a parent, muffins be damned" (Pinker, 2009), then such crisp concepts should be found here, and yet here they are "muddied up" by typicality. If Exp. 6A and the earlier explicit classification studies are insufficient to trigger the formal concept, then what is?

The second objection is that any theory that posits a distinct concept for each situation risks a combinatorial explosion. Would we need a new GRANDMOTHER concept to explain each distinct pattern of results likely to be obtained if people were asked whether a woman who did not know she had a grandchild (but actually did) was a grandmother? Or asked at which point a male-to-female transgender individual with a grandchild became a grandmother? Or asked whether women who gave up a child for adoption or lost custody of the child (under various circumstances) could become grandmothers? The very ability to consider these scenarios, each of which requires a reconsideration of what properties are relevant and irrelevant requires flexibly instantiating subtly different conceptual representations and *not* simply following a rule.

14. General discussion

People routinely made errors in classifying items into well-defined categories such as ODD, EVEN, and TRIANGLE and made inferences consistent with their classification decisions and inconsistent with application of trivially simple rules. The response patterns were well predicted by stimulus characteristics. In the case of parity, participants were much more likely to mis-identify the parity of a number if it had opposite-parity digits. Such input dependence was unrelated to the formal correctness of participants' definitions of numerical parity. Similar kind of input-dependence was observed in classifying triangles, e.g. people failed to select scalene triangles as triangles more than 15% of the time and failed to uniformly apply a basic geometric property true of all triangles, to all triangles. The observed response patterns cannot be explained as simple errors of perceptual selection or as simple performance quirks. Many participants insisted—in ways that were quite reasonable—that 400 is *more even* than 798 and that some triangles are more triangular, e.g., by appealing to the relative difficulty of dividing 798 by 2 compared to dividing 400 by 2, and by commenting on the "perfection" of equilateral triangles relative to scalene triangles.[12] The tendency to *not* use hard-and-fast rules predicted people's performance in Exp. 6A which required them to make inferences about the likelihood of winning a hypothetical contest. Despite being clearly instructed that only grandmothers are eligible and that all eligible entrants have an equal chance of winning, people nevertheless rated the more typical grandmothers—ones who were older and had more grandchildren—as being more likely to win.

Computational devices that rely on discrete symbols—digital computers being the paradigmatic example—do not make such mistakes. Storing numbers in discrete registers allows algorithms to operate over them in a context-free way. The algorithm for computing parity is remarkably simple. My laptop takes .038 μs to compute the parity of 2 and .038 μs to compute the parity of 918239748712387482 7. People's performance is not simply slower and more error prone, but qualitatively different, displaying inherent sensitivity to aspects of the input that are formally irrelevant to the operation being performed.

Although such input dependence might be expected in the performance of children just learning the formal rules of membership (Berch, Foley, Hill, & Ryan, 1999; Clements, Swaminathan, Hannibal, & Sarama, 1999; Tsamir, Tirosh, & Levenson, 2008), it is generally assumed to disappear by late childhood. The present results show this not to be the case (see also Ward, 2004 for another demonstration of educated adults failing to classify non-canonical triangles as triangles). Even with no externally imposed time-pressure, adult classification of formally-defined categories continues to show a predictable pattern of errors. These results pose a challenge to theories positing that human concepts of well-defined categories have discrete representations. As recently stated by Gleitman et al., "No person who knows and states that all odd numbers are equally odd should rate some of them more odd than any others, even by a smidgen" (2012). I show here that not just typicality ratings or latencies, but identification itself is sensitive to allegedly irrelevant features of the input.

It is tempting to ascribe the results of the present studies to inattention or perceptual mis-identification. On this account, classification errors in judging parity should be correlated with errors on tests designed to measure selective attention (e.g., flanker congruity task). The bulk of the evidence does not support this account. For example, as shown in Exp. 3A, perceptual selection as measured by a flanker task correlates with the effect of digit-length on parity RTs, but not the effect of the number of opposite-parity digits. It is also unclear how this account can explain the observed findings in the domain of triangles (Exps. 5A–5D) and grandmothers (Exp. 6A).

There is an important difference between the present results and such classic findings as Moyer and Landauer's (1967) demonstration that e.g., people make more errors judging the truth value of $4 < 5$ compared to $1 < 5$. Such

[12] The relative status of triangles is aptly captured by Abbott in *Flatland*: "The birth of a True Equilateral Triangle from Isosceles parents is the subject of rejoicing in our country for many furlongs around. After a strict examination conducted by the Sanitary and Social Board, the infant, if certified as Regular, is with solemn ceremonial admitted into the class of Equilaterals" (Abbott, 1884/2008).

findings show that human computations underlying numerical inequality judgments are analog, reflecting the actual difference between the two numbers. But although the desired response is discrete the *degree* to which one number is larger than another is, in fact, continuous. In contrast, the parity of a number, the triangleness of a triangle, or whether someone is a grandmother is not—or at least should not be—graded. And yet, as shown by the present results, it is. The argument that such graded performance reflects the operation of a peripheral identification system (Armstrong et al., 1983; Gleitman et al., 2012) cannot, it would seem, account for the presently observed systematic failures in identification and inference. Given that a critical function of concepts is discriminating category members from nonmembers (e.g., Prinz, 2004), to shift this burden to peripheral identification procedures would be to deny concepts their very *raison d'etre*.

Not only were people's errors systematic, but individuals' classification performance was correlated in a meaningful way with their explicit endorsements of statements about gradedness of formal categories. Participants who thought that some triangles were better than others were more likely to mis-classify the parity of numbers like 798 (Exp. 4). Participants who thought that 400 was a more even number than 798 were likely to think that more typical grandmothers had a better chance of winning the contest described in Exp. 6A. Do such correlations simply reflect the greater mathematical savvy of some people than others? On first blush, this question is somewhat circular: if we assume that part of mathematical savvy is to view all even numbers as equally even then mathematical savvy predicts mathematical savvy. A more interesting possibility is that there is a dimension of individual variability related to forming discrete concepts and that mathematical savvy arises from the ability to discretize and abstract over task-irrelevant information (although it is possible that real mathematical savvy might actually be hindered by such abstraction). The question of why some people appear to discretize membership in formal categories more than others deserves a *much* more rigorous investigation than is provided here. The individual differences described here bear a strong similarity to those described by Wasserman and colleagues (Wasserman & Castro, 2012 for review) in which about 30% of participants overlook seemingly trivial symbolic rules (if all same → A, else B) in favor of graded category membership based on similarity.

Mathematically, all even numbers are equally even and all odd numbers are equally odd. I have argued that this is not the case for the psychological concept of oddness. But what does it mean for 400 to be more even than 798? Is it that 400 *looks* more even? At issue, I think, is not perceptual similarity, but overall *representational similarity*. People mistake 798 for an odd number not because it *looks* like an odd number. Rather, the reason 798 *looks* odd (at least odder than 400), is that the representation of 798 is closer to that of other odd numbers than the representation of 400. 798 is *almost* odd.[13] On this view, the human algorithm of parity judgments is best conceived as a transformation of a representational state-space in which numbers "reside." This transformation *is* categorization (Lupyan, Mirman, Hamilton, & Thompson-Schill, 2012) resulting in odd numbers being on one side of a decision boundary and even numbers on the other. However, the transformation is partial, and the representational space retains some of its analog structure, also giving rise to numerical distance effects (Moyer & Landauer, 1967) and spatial–numerical interactions (Dehaene et al., 1993).

In rejecting as absurd the idea that people's representations of integers or plane-geometry categories are graded, Armstrong et al. (1983) ask: how could one ever compute with such a graded representation? (p. 284). The answer is: not very well! Or at least far worse than what can be achieved by an algorithm that fully abstracts irrelevant details from the input (i.e., MOD 2). Why aren't people better at tasks like this? One answer is that such context-free computations are simply not needed for ordinary cognition. I expect that individuals who mis-identify the parity of 798 or those who reject scalene triangles as triangles are unremarkable in their everyday behavior and are unlikely to have the sort of problems in everyday cognition that would be expected if running such algorithms constituted the bulk of cognition. Recall that a substantial portion of the sample have college and graduate degrees, and these individuals too make the same errors! On the present account, such errors reflect a failure to fully disengage similarity-based computations on which the rest of cognition depends. This reliance on graded context- and task-dependent representations is, arguably, what enables people to flexibly construe the same stimuli in multiple contradictory ways, as called on by task goals (Linhares, 2000).

Perhaps the better question is not why people sometimes mistake 798 for an odd number or fail to classify a scalene triangle as a triangle, but rather how people ever transcend these limitations. After all, although everyone's classification performance is sensitive to typicality in timed presentations, many people's performance becomes essentially perfect when time pressure is removed. Despite thinking that rotated triangles are less triangular, people can be taught to perform geometric computations regardless of orientation.

Notwithstanding the *relative* difficulty of context-independent computations (which are, it should be noted, the bread and butter of any theory positing compositionality as foundational, e.g., Fodor, 2001), people can, under some circumstances *approximate* symbolic, context-free computation. Clearly, it is *possible* for people to operationalize parity in a rigorous way, to build computers that implement perfect parity algorithms, and even simply to ponder these very questions. What enables humans to do this awaits explanation (see Penn, Holyoak, & Povinelli, 2008 and commentaries for related discussion). However, this critically important question is ignored when a field assumes *a priori* the existence of many of the very capacities that are in need of explanation. Cognitive scientists need to take seriously the constraints that the implementational medium—neural networks—place on computations performed by biological organisms (Buonomano & Maass,

[13] Just as a woman who is about to have a grandson is *almost* a grandmother.

2009; Freeman, 2007; Zylberberg, Dehaene, Roelfsema, & Sigman, 2011).[14] Rather than being the building blocks of human cognition, symbolic computation may emerge when "core" human cognition employing similarity-based fuzzy representations is augmented by culture, education, notational systems, and language itself (e.g., Clark, 1998; Gomila, Travieso, & Lobo, 2012; Lupyan, 2012a, 2012b; McClelland, 2010).

14.1. A note on the use of crowdsourcing platforms in cognitive science

The 13 experiments in this paper include a total of 1117 participants of whom all but 35 were tested using the crowdsourcing service Amazon Mechanical Turk (mTurk). A number of good reviews are now available detailing demographics, motivations, and quirks of mTurk's enormously large and diverse user base (Berinsky, Huber, & Lenz, 2011; Buhrmester, Kwang, & Gosling, 2011; Goodman, Cryder, & Cheema, 2012; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). There is also a largely successful effort to replicate classic findings in cognitive psychology through this platform (Crump, McDonnell, & Gureckis, 2013). The rapid rise of crowdsourcing in the social sciences may mean that it will displace a large proportion of lab-based tasks, especially when in-person observation and tight controls over the user's hardware is not required (although clever use of within-subject designs can overcome some of these limitations). Collecting data from much larger and more diverse participant groups will help to address the problematic reliance on "WEIRD" 18–22 year-old college students (Henrich, Heine, & Norenzayan, 2010) and lead to a richer understanding of individual differences that traditional university lab multi-trial/low-*n* tasks may obscure.

References

Abbott, E. A. (1884). *Flatland: A romance of many dimensions*. <http://www.gutenberg.org/ebooks/201>.

Anderson, J. M. (2006). The non-autonomy of syntax. *Folia Linguistica, 39*(3–4), 223–250. http://dx.doi.org/10.1515/flin.2006.39.3-4.223.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). Coherent arbitariness: Stable demand curves without stable preferences. *Quarterly Journal of Economics, 118*(1), 73–106.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13*(3), 263–308. http://dx.doi.org/10.1016/0010-0277(83)90012-4.

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge: Cambridge University Press.

Bates, D., & Maechler, M. (2012). *Package "lme4."* <ftp://ftp.ctex.org/mirrors/CRAN/web/packages/lme4/lme4.pdf>.

Berch, D. B., Foley, E. J., Hill, R. J., & Ryan, P. M. (1999). Extracting parity and magnitude from Arabic numerals: Developmental changes in number processing and mental representation. *Journal of Experimental Child Psychology, 74*(4), 286–308.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2011). *Using Mechanical Turk as a subject recruitment tool for experimental research* (in preparation). <http://web.mit.edu/berinsky/www/files/MT.pdf>.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science, 6*(1), 3–5. http://dx.doi.org/10.1177/1745691610393980.

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews. Neuroscience, 10*(2), 113–125. http://dx.doi.org/10.1038/nrn2558.

Bybee, J., & McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review, 22*(2–4). http://dx.doi.org/10.1515/tlir.2005.22.2-4.381.

Casasanto, D., & Lupyan, G. (in press). All concepts are Ad hoc concepts. In E. Margolis & S. Laurence (Eds.), *Concepts: New directions*. Cambridge: MIT Press.

Chomsky, N. (1995). *The minimalist program*. The MIT Press.

Clark, A. (1998). Magic words: How language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 162–183). New York, NY: Cambridge University Press.

Clark, H. H. (1983). Making sense of nonce sense. In G. B. F. d'Arcais & R. J. Jarvella (Eds.), *The process of language understanding* (pp. 297–331). Wiley-Blackwell.

Clements, D. H., Swaminathan, S., Hannibal, M. A. Z., & Sarama, J. (1999). Young children's concepts of shape. *Journal for Research in Mathematics Education, 30*(2), 192–212. http://dx.doi.org/10.2307/749610.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8*(3), e57410. http://dx.doi.org/10.1371/journal.pone.0057410.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*(3), 371.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science, 33*(4), 547–582. http://dx.doi.org/10.1111/j.1551-6709.2009.01023.x.

Eriksen, C. W. (1995). The Flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition, 2*(2–3), 101–118. http://dx.doi.org/10.1080/13506289508401726.

Fauconnier, G., & Turner, M. (2003). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (2001). Language, thought and compositionality. *Mind & Language, 16*(1), 1–15. http://dx.doi.org/10.1111/1468-0017.00153.

Freeman, W. J. (2007). The place of "codes" in nonlinear neurodynamics. *Progress in Brain Research, 165*, 447–462. http://dx.doi.org/10.1016/S0079-6123(06)65028-0.

Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience* (1st ed.). Wiley-Blackwell.

Geeraerts, D. (1989). Introduction: Prospects and problems of prototype theory. *Linguistics, 27*(4), 587–612.

Gleitman, L., Armstrong, S. L., & Connolly, A. C. (2012). Can prototype representations support composition and decomposition? In M. Werning, W. Hinzen, & E. Machery (Eds.), *Oxford handbook of compositionality*. New York: Oxford University Press.

Gomila, A., Travieso, D., & Lobo, L. (2012). Wherein is human cognition systematic? *Minds and Machines, 22*(2), 101–115.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2012). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*. http://dx.doi.org/10.1002/bdm.1753.

Hampton, J. A. (2006). Concepts as prototypes. *The psychology of learning and motivation: Advances in research and theory* (Vol. 46, pp. 79–113). London: Elsevier.

Hampton, J. A. (2012). Thinking intuitively: The rich (and at times illogical) world of concepts. *Current Directions in Psychological Science, 21*(6), 398–402.

Hasen, R. L. (2012). *Wrong number. Slate*. <http://www.slate.com/articles/news_and_politics/jurisprudence/2012/10/ohio_voter_laws_the_battle_over_disenfranchisement_you_haven_t_heard_about_.html>.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences, 33*(2-3), 61–83. http://dx.doi.org/10.1017/S0140525X0999152. discussion 83–135..

Lakoff, G. (1990). *Women, fire, and dangerous things*. University of Chicago Press.

Larochelle, S., Richard, S., & Souliëres, I. (2000). What some effects might not be: The time to verify membership in "well-defined" categories.

[14] It is interesting to note that the person most widely cited as a proponent for independent levels of analysis–David Marr–in fact remarked that "The choice [of an algorithm], then, may dependent on the type of hardware or machinery in which the algorithm is to be embodied physically" (Marr, 1982, p. 24; see also Rolls, 2011).

The Quarterly Journal of Experimental Psychology Section A, 53(4), 929–961. http://dx.doi.org/10.1080/713755940.

Linhares, A. (2000). A glimpse at the metaphysics of Bongard problems. *Artificial Intelligence, 121*(1), 251–270.

Lupyan, G. (2012a). What do words do? Towards a theory of language-augmented thought. In B. H. Ross (Ed.). *The psychology of learning and motivation* (Vol. 57, pp. 255–297). Academic Press. <http://www.sciencedirect.com/science/article/pii/B9780123942937000078>.

Lupyan, G. (2012b). Language augmented prediction. *Frontiers in Theoretical and Philosophical Psychology, 3*, 422. http://dx.doi.org/10.3389/fpsyg.2012.00422.

Lupyan, G., Mirman, D., Hamilton, R. H., & Thompson-Schill, S. L. (2012). Categorization is modulated by transcranial direct current stimulation over left prefrontal cortex. *Cognition, 124*(1), 36–49. http://dx.doi.org/10.1016/j.cognition.2012.04.002.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology-General, 141*(1), 170–186. http://dx.doi.org/10.1037/a0024904.

Machery, E. (2009). *Doing without concepts*. USA: Oxford University Press.

Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Co.

Marcus, G. F. (1999). Rule learning by seven-month-old infants. *Science, 283*(5398), 77–80. http://dx.doi.org/10.1126/science.283.5398.77.

Margolis, E., & Laurence, S. (2008). Concepts. In S. P. Stich & T. A. Warfield (Eds.), *The Blackwell guide to philosophy of mind* (pp. 190–213). John Wiley & Sons.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129*(4), 592–613.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23. http://dx.doi.org/10.3758/s13428-011-0124-6.

McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science, 2*(4), 751–770. http://dx.doi.org/10.1111/j.1756-8765.2010.01116.x.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception. 1. An account of basic findings. *Psychological Review, 88*(5), 375–407.

Medin, D. L., & Smith, E. E. (1984). Concepts and concept-formation. *Annual Review of Psychology, 35*, 113–138.

Miller, J. (1991). The flanker compatibility effect as a function of visual angle, attentional focus, visual transients, and perceptual load: A search for boundary conditions. *Perception & Psychophysics, 49*(3), 270–288.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215*(5109), 1519–1520. http://dx.doi.org/10.1038/2151519a0.

Murphy, G. L. (2002). *The big book of concepts*. The MIT Press.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411–419. http://dx.doi.org/10.2139/ssrn.1626226.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*(02), 109–130. http://dx.doi.org/10.1017/S0140525X08003543.

Pinker, S. (1999). Out of the minds of babes. *Science, 283*(5398), 40–41. http://dx.doi.org/10.1126/science.283.5398.40.

Pinker, S. (2000). *Words and rules: The ingredients of language*. Harper Perennial.

Pinker, S. (2009). *How the mind works*. W.W. Norton & Company.

Prinz, J. J. (2004). *Furnishing the mind: Concepts and their perceptual basis* (New ed.). The MIT Press.

Rolls, E. T. (2011). David Marr's Vision: floreat computational neuroscience. *Brain, 134*(3), 913–916.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In *Cognitive development and the acquisition of language*. New York: Academic Press.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382–439.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: Shifting demographics in Mechanical Turk. In *Proceedings of the 28th of the international conference extended abstracts on human factors in computing systems* (pp. 2863–2872). New York, NY, USA: ACM.

Sandberg, C., Sebastian, R., & Kiran, S. (2012). Typicality mediates performance during category verification in both ad-hoc and well-defined categories. *Journal of Communication Disorders, 45*(2), 69–83. http://dx.doi.org/10.1016/j.jcomdis.2011.12.004.

Seidenberg, M. S. (1999). Do infants learn grammar with algebra or statistics? *Science, 284*(5413), 434–435. author reply 436–437.

Sleator, D. D. K., & Temperley, D. (1995). *Parsing English with a link grammar*. arXiv:cmp-lg/9508004. <http://arxiv.org/abs/cmp-lg/9508004>.

Tracie Hunter v. Hamilton County Board of Elections (2012). *No. 1:10CV820 (United States District Court for the Souther District of Ohio Western Division 2012)*.

Tsamir, P., Tirosh, D., & Levenson, E. (2008). Intuitive nonexamples: The case of triangles. *Educational Studies in Mathematics, 69*(2), 81–95. http://dx.doi.org/10.1007/s10649-008-9133-5.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124.

Ward, R. A. (2004). An investigation of K-8 Preservice Teachers' concept images and mathematical definitions of polygons. *Issues in Teacher Education, 13*(2), 39–56.

Wasserman, E. A., & Castro, L. (2012). Categorical discrimination in humans and animals. *Psychology of learning and motivation* (Vol. 56, pp. 145–184). Elsevier. <http://linkinghub.elsevier.com/retrieve/pii/B9780123943934000054>.

Zylberberg, A., Dehaene, S., Roelfsema, P. R., & Sigman, M. (2011). The human Turing machine: A neural framework for mental programs. *Trends in Cognitive Sciences, 15*(7), 293–300. http://dx.doi.org/10.1016/j.tics.2011.05.007.