

The Role of “Prominent Numbers” in Open Numerical Judgment: Strained Decision Makers Choose from a Limited Set of Accessible Numbers

Organizational Behavior and Human Decision Processes, Volume 147,
July 2018, pp. 94-107.

SUPPLEMENTAL ONLINE MATERIAL

Contents

1	Robustness of Residuals in Study 1	S3
1.1	Background	S3
1.2	Assessing Robustness	S5
2	Prominent Number Clustering and Investor Sophistication in Study 1	S7
2.1	Background	S7
2.2	Method	S7
2.3	Results	S8
3	Data Collection Plans for Studies 4a—4c	S10
4	Supplemental Studies: The Relative Graininess of Round and Prominent Numbers	S11
4.1	Study S1: Speeded Comparisons	S12
4.1.1	Method	S12
4.1.2	Results	S13
4.2	Study S2a: Explicit Judgments	S14
4.2.1	Method	S15
4.2.2	Results	S16
4.3	Study S2b: Inferred Graininess from Accuracy-Informativeness Tradeoffs . . .	S16
4.3.1	Method	S19
4.3.2	Results	S19
4.4	Summary of Studies S1, S2a, and S2b	S19
4.5	Supplemental Study S3: Task Motivation and Prominent-Number Clustering	S20

4.5.1	Method	S21
4.5.2	Results	S23
5	Supplementary Appendix 1 - Items from Study S2	S25

1 Robustness of Residuals in Study 1

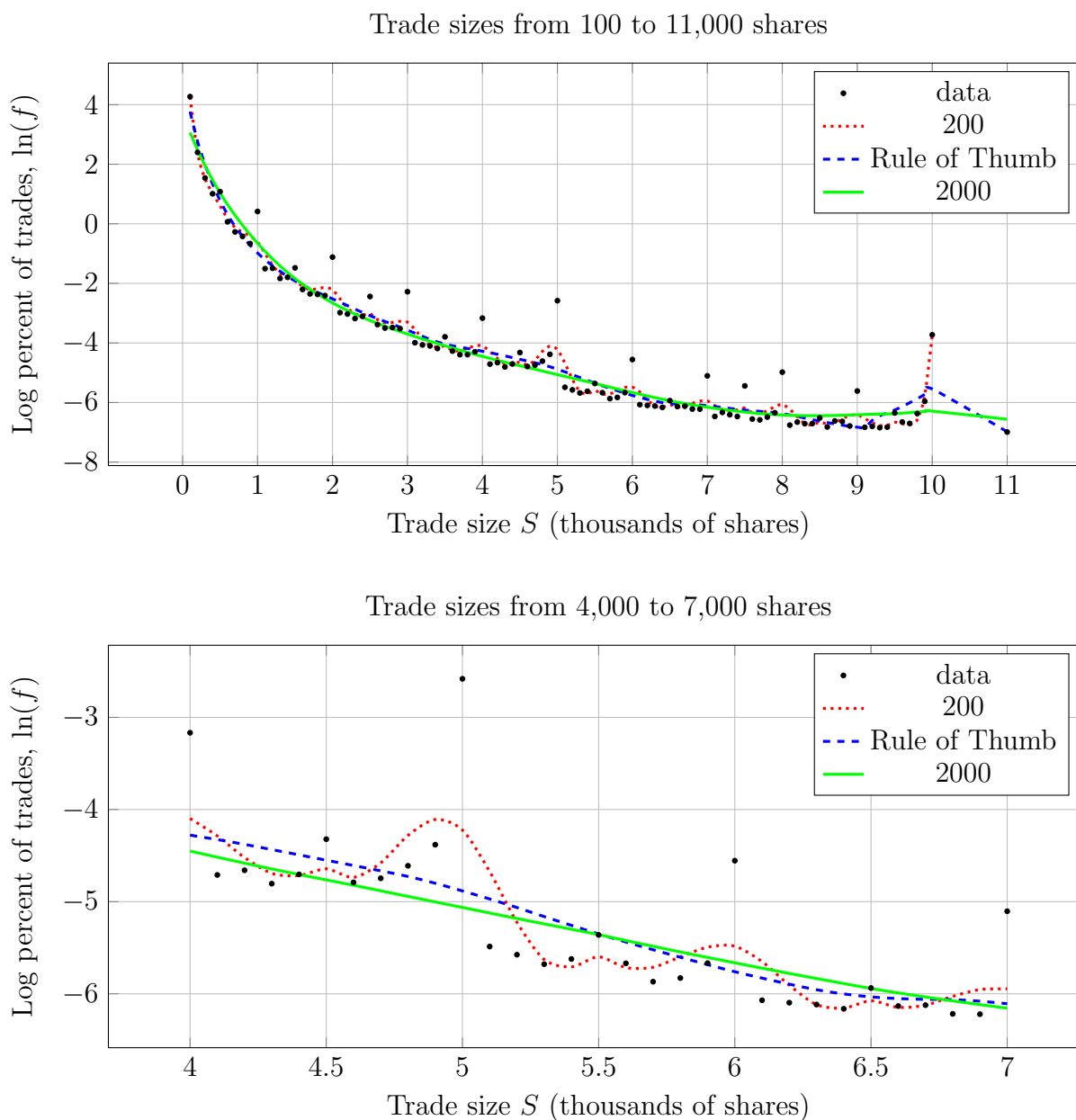
1.1 Background

To test for Prominent-Number clustering in the trading data, we sought to determine if the observed frequency of trades exceeded the expected frequency of trades at each trade size. The challenge lies in computing the expected frequency for a given trade size. Basic economics suggests that larger trades should occur less frequently than smaller trades because all investors are wealth-constrained. Beyond this assumption, there is no other theoretical guidance about what the expected frequency should be. Acknowledging the lack of theory, we chose to use a non-parametric regression rather than trying to fit an arbitrary function to the data.

Specifically, we used a local polynomial smoothing process which fits a smooth line to the data. As with all regressions, there is a trade-off between the fit (i.e., decreasing residuals) and parsimony. Local polynomial smoothing fits a polynomial of degree n around each data point and weights each data point according to a bandwidth. Odd-order polynomials typically reduce bias relative to even-order polynomials (Fan and Gijbels (1996)), and we therefore chose $n = 3$ (the lowest odd-degree polynomial > 1). The more critical choice was bandwidth. Low bandwidths put a lot of weight on nearby data points and less weight on distant data points. Hence, using a low bandwidth is more like “connecting the dots.” High bandwidths, in contrast, put less weight on local points and more weight on distant point. They result in a smoother overall function. There is theory guiding the choice of bandwidth and an explicit formula for what is called the “rule-of-thumb” bandwidth (Fan and Gijbels (1996); StataCorp (2013)). Applying the formula to our data indicates that we should use a bandwidth of 820.

To see the effect of bandwidth selection, refer to figure S1. The first fitted function was estimated using a relatively low bandwidth of 200 and is shown in the figure as a dotted red line. As the plot demonstrates, a lot of weight is placed on nearby data, so the estimated function spikes at trade sizes of 2,000; 5,000; and 10,000 shares. This low bandwidth was biased toward the “connect the dots” approach. The residuals are smaller, but the shape of the function is counter to the basic assumption that trade frequency declines monotonically with trade size. The second fitted function was estimated using the rule-of-thumb bandwidth of 820, shown with the dashed blue line.

Figure S1: Trade frequency vs. trade size for various non-parametric bandwidths



Note: The top plot shows the log of the frequency of daily stock trades for a particular size, $\ln(f)$, vs. trade size S . f is the frequency trade of a particular size for a each firm in a given day, averaged across the 125 trading days between January 1, 2011 and June 30, 2011 in our sample. The fitted function is obtained using a non-parametric fit with a 3rd degree polynomial, an Epanechnikov kernel and a bandwidth of 200 (dotted red line), a bandwidth of 2000 (solid green line), and an optimal bandwidth (dashed blue line). The bottom plot uses the same data as the top plot but restricts trade sizes to between 4,000 and 7,000 shares.

The higher bandwidth places less weight on local points and more weight on distant points and therefore provides a smoother function than the bandwidth of 200. Notably, it still spikes at 10,000 shares, still violating the assumption of monotonically decreasing trade frequency. The last fitted function displayed was estimated using a bandwidth of 2,000; shown as the solid green line. It closely tracks the fitted function using the rule-of-thumb bandwidth and, while it still spikes slightly at 10,000 shares, it makes more economic sense than the suggested bandwidth of 820. Based on these determinations, we present results in the main text using a bandwidth of 2,000.

1.2 Assessing Robustness

While Figure S1 helps to illustrate the intuition behind the choice of bandwidth, it does not address the impact of bandwidth selection on our conclusions. To that end, we reproduce the analyses reported in Table 1 using a bandwidth of 200 and a bandwidth of 820 and display them in Table S1. Panel A in Table S1 contains comparisons between the residuals of Prominent Numbers and those of their nearest neighbors where residuals are computed using a fitted function with a bandwidth of 200. As in Table 1, the comparisons are made using neighboring Round Numbers (i.e., the closest number with one non-zero digit). Also, as with Table 1, we include a comparison of 100-versus-200 and 1,000-versus-2,000 for the sake of symmetry, but note that they are both Prominent Numbers. Looking at the first row, we see that the observed residual frequency of 300-share trades is greater than that of 200-shares ($p < .001$). This differs from the conclusion described in the main text based on Table 1. Recall, however, that this low bandwidth "over-fits" the function. The over-fitted function spikes at 200 more than at 300, hence the residual for 200 shares is probably estimated as lower than it should be. With such a low bandwidth, we could not fit a value to the 11,000-share trade size (using Stata), hence it is not reported.

Aside from these exceptions at the extreme low- and high-end where over-fitting is a particular problem, the results in Table S1 closely mirror those in Table 1. The results of 500 versus 400 and 600; 1,000 versus 900 and 1,100; 2,000 versus 3,000 and 5,000; and 5,000 versus 4,000 and 6,000 are all consistent with those in Table 1. Thus, while some differences are revealed as a result of over-fitting, the results are largely robust to a chosen bandwidth of 200.

Table S1: Trade frequency residuals for different non-parametric bandwidths

Panel A: Bandwidth = 200							
Next Lower		Prominent Number		Next Higher		t -statistic	t -statistic
Trade	Mean	Trade	Mean	Trade	Mean	for $H_0:$	for $H_0:$
Size	Residual	Size	Residual	Size	Residual	$\mu_l - \mu_p = 0$	$\mu_p - \mu_h = 0$
	(μ_l)		(μ_p)		(μ_h)		
100	0.0098 (0.0001)	200	0.0206 (0.0009)	300	0.0286 (0.0012)	-10.67 p < 0.001	-4.02 p < 0.001
400	-0.0163 (0.0015)	500	0.4731 (0.0021)	600	-0.1028 (0.0025)	-159.49 p < 0.001	146.56 p < 0.001
900	-0.2541 (0.0038)	1000	1.0222 (0.0036)	1100	-0.5097 (0.0044)	-219.89 p < 0.001	220.37 p < 0.001
1000	1.0222 (0.0036)	2000	1.0980 (0.0065)	3000	1.0306 (0.0117)	-10.94 p < 0.001	5.23 p < 0.001
4000	0.9364 (0.0133)	5000	1.6619 (0.0147)	6000	0.9604 (0.0227)	-35.80 p < 0.001	27.19 p < 0.001
9000	0.8437 (0.0307)	10000	0.0189 (0.0037)			27.40 p < 0.001	

Panel B: Rule of Thumb Bandwidth = 820							
Next Lower		Prominent Number		Next Higher		t -statistic	t -statistic
Trade	Mean	Trade	Mean	Trade	Mean	for $H_0:$	for $H_0:$
Size	Residual	Size	Residual	Size	Residual	$\mu_l - \mu_p = 0$	$\mu_p - \mu_h = 0$
	(μ_l)		(μ_p)		(μ_h)		
100	0.5202 (0.0036)	200	-0.3618 (0.0015)	300	-0.4488 (0.0019)	219.54 p < 0.001	37.90 p < 0.001
400	-0.3320 (0.0022)	500	0.2929 (0.0027)	600	-0.2449 (0.0032)	-200.38 p < 0.001	124.01 p < 0.001
900	0.0547 (0.0045)	1000	1.3915 (0.0039)	1100	-0.2996 (0.0057)	-226.39 p < 0.001	226.66 p < 0.001
1000	1.3915 (0.0039)	2000	1.4092 (0.0075)	3000	1.3013 (0.0116)	-2.29 p = 0.024	8.42 p < 0.001
4000	1.1188 (0.0155)	5000	2.3425 (0.0136)	6000	1.2329 (0.0237)	-57.12 p < 0.001	40.31 p < 0.001
9000	1.2727 (0.0316)	10000	1.8993 (0.0276)	11000	-0.0092 (0.0003)	-16.64 p < 0.001	68.68 p < 0.001

Note: This table contains summary statistics for the daily residuals from non-parametric regressions of the logarithm of the frequency of trades in each size on trade size. The fitted function is obtained using a non-parametric fit with a 3rd degree polynomial, an Epanechnikov kernel and a bandwidth of 200 in panel A and a “rule of thumb” bandwidth in Panel B. Daily residuals are computed for each trading day between January 1, 2011 and June 30, 2011 resulting in 125 observations. The t -statistics corresponding to the null hypothesis that the mean residuals for the “Next Lower” and “Prominent Number” sizes are different are in the second to last column. Similarly, those corresponding to the null hypothesis that the mean residuals for the “Prominent Number” and “Next Higher” sizes are different are in the last column. Standard errors are in parentheses below the mean residual (μ_l , μ_p , and μ_h). The p-values for the test in difference in means are below the t -statistics.

Panel B in Table S1 parallels Panel A except that the fitted function is estimated using the rule-of-thumb bandwidth of 820. Here, the results are largely consistent with Table 1 except for 2,000 versus 1,000; which are both Prominent Numbers. All other results are consistent with those presented in Table 1 both in sign and statistical significance, including the residual frequency of 200-share trades relative to 300-share trades. In sum, our results are robust to low, rule-of-thumb, and high bandwidths. We chose to report the results for the bandwidth of 2,000 because, of the three, it best aligns with the assumption of a monotonically downward-sloping function. We suggest that this level of robustness supports the key conclusion from Study 1 that trade sizes of Prominent Numbers are chosen more frequently than expected compared to trade sizes of non-Prominent Numbers.

2 Prominent Number Clustering and Investor Sophistication in Study 1

2.1 Background

We tested the prediction that Prominent-Number clustering is more pronounced among relatively less sophisticated retail investors (e.g., your uncle or sister) than among relatively more sophisticated institutional investors (e.g., Goldman Sachs or Fidelity). Retail investors are significantly under-diversified and they under-perform relative to the average market return, indicating that institutional investors must out-perform them ([Barber and Odean \(2000\)](#)). Also, in contrast to retail investors, institutions employ full-time analysts and have access to proprietary data. Institutional holdings are therefore a good proxy for investor sophistication. Compared to institutional investors, retail investors would need to invest more cognitive effort to make good mean-variance optimizing decisions. Assuming that they instead rely even more on cognitive shortcuts, we predicted that retail investors would show stronger Prominent-Number clustering than would institutional investors.

2.2 Method

Using the stock ticker symbol and date from the New York Stock Exchange Trade and Quote database (TAQ), we matched each observation to the corresponding ticker and date from the Center for Research in Security Prices (CRSP) database, which contains the stock price,

the number of shares outstanding, and the firm's CUSIP, a unique identifier assigned to each stock by the Committee on Uniform Security Identification Procedures. Each quarter, institutional money managers such as banks, pension funds and mutual funds must disclose their stock holding on form 13F to the Securities and Exchange Commission. This data is provided by Thomson Reuters and we matched the institutional holdings of each stock to trade size using the firm's CUSIP. This allowed us to determine if the trade is associated with a stock primarily owned by institutions and to test whether sophistication mitigates the tendency to use Prominent Numbers.

2.3 Results

To test whether Prominent-Number clustering is more pronounced among retail investors, we computed the frequency of trades in each trade size as in Study 1 and then sorted firms into deciles based on the fraction of institutional ownership. We computed the average frequency of a given trade size across all firms in each decile, which gave us 125 time-series observations that were used to compute the results in Table S2. In the table, we compare each Prominent-sized trade to that of its Non-Prominent Round neighbor by taking the ratio of the frequency of trades in each size. We used ratios rather than differences due to the extreme variation in frequency; in decile 1, 200 shares occur 13.944% of the time, while 9,000 shares occur .011% of the time. For example, the frequency of 500 to 400 share trades decreased from 1.239 for those stocks that have the lowest institutional ownership to 0.710 for those stocks that have the highest, indicating that retail investors (decile 1) are more likely to use Prominent-Number sized trades than institutional investors (decile 10). The t -statistics in the table correspond to a test of the null hypothesis that the ratio is 1 while p -values are almost universally less than .001 and are not reported in the table to conserve space.

Similar results hold for the ratio of 1,000 to 900 share trades and 2,000 to 3,000 share trades. The pattern for the ratio of 5,000 to 4,000 share trades is not monotonic, beginning at 1.762 for those stocks with the lowest institutional ownership and then decreasing and increasing several times before ending at 2.067 for those stocks with the highest institutional ownership.

Table S2: Trade size ratios sorted by institutional ownership

IOWN Decile	Size				Size				Size			
	200 (a)	300 (b)	ratio (a/b)	t-stat (a/b)	500 (c)	400 (d)	ratio (c/d)	t-stat (c/d)	1000 (e)	900 (f)	ratio (e/f)	t-stat (e/f)
1 (low)	13.944	7.043	1.980	60.42	5.474	4.418	1.239	22.77	3.020	0.981	3.079	47.15
2	12.837	6.387	2.010	72.33	4.869	3.924	1.241	25.07	2.719	0.868	3.133	58.98
3	12.629	6.058	2.085	77.62	4.490	3.683	1.219	23.20	2.511	0.794	3.161	51.74
4	11.941	5.380	2.219	96.91	3.811	3.359	1.135	11.76	2.037	0.630	3.233	38.13
5	11.214	4.840	2.317	97.74	3.239	3.014	1.075	8.53	1.674	0.549	3.050	46.08
6	10.400	4.112	2.529	107.38	2.295	2.372	0.968	-2.95	1.178	0.412	2.856	37.58
7	9.852	3.652	2.698	142.05	1.745	1.995	0.875	-16.62	0.766	0.316	2.424	33.07
8	9.495	3.324	2.857	176.39	1.369	1.778	0.770	-35.55	0.517	0.237	2.182	31.94
9	8.981	3.031	2.963	212.45	1.150	1.535	0.749	-46.25	0.414	0.200	2.067	28.66
10 (high)	8.532	2.727	3.129	197.78	0.948	1.336	0.710	-55.51	0.295	0.163	1.812	22.89

IOWN Decile	Size				Size				Size			
	2000 (a)	3000 (b)	ratio (a/b)	t-stat (a/b)	5000 (c)	4000 (d)	ratio (c/d)	t-stat (c/d)	10000 (e)	9000 (f)	ratio (e/f)	t-stat (e/f)
1 (low)	0.745	0.240	3.105	24.78	0.184	0.104	1.762	11.48	0.060	0.011	5.453	8.24
2	0.593	0.187	3.169	30.14	0.128	0.078	1.637	9.92	0.036	0.006	6.342	8.17
3	0.556	0.177	3.139	23.24	0.133	0.069	1.932	10.46	0.039	0.006	6.765	8.57
4	0.443	0.135	3.294	20.40	0.086	0.049	1.747	11.13	0.027	0.004	7.629	4.26
5	0.333	0.103	3.249	28.70	0.085	0.040	2.153	12.23	0.025	0.004	6.222	11.12
6	0.217	0.071	3.081	22.67	0.054	0.032	1.712	10.59	0.020	0.002	8.285	11.07
7	0.154	0.045	3.458	25.38	0.034	0.018	1.907	14.24	0.012	0.002	7.035	9.21
8	0.103	0.027	3.830	21.29	0.022	0.012	1.757	11.70	0.009	0.001	8.757	11.95
9	0.071	0.020	3.481	28.96	0.018	0.013	1.369	7.40	0.007	0.001	8.639	17.12
10 (high)	0.052	0.018	2.814	17.29	0.013	0.006	2.067	13.27	0.006	0.001	10.031	11.84

Note: This table contains summary statistics for various trade sizes sorted by the decile of institutional ownership. The numbers listed in the column “Size 200” represent the number of 200-share trades expressed as a percentage of all trades for each firm for each day, averaged across all days in the sample. The number of trades as a fraction of all trades for a firm i on day t , averaged across all firms in each institutional ownership decile of day t yields a sample of 125 independent observations. The t -statistic in the column to the right of the trade size columns corresponds to the null hypothesis that the ratio of the proportion of trade sizes is different from 1.

While the ratio generally increases with institutional ownership for 10,000 vs. 9,000 share trades, these results are hard to interpret because not many retail investors trade such share sizes; 10,000 shares at \$20 per share is \$200,000 which is almost 10 times the average household’s stock holdings of \$29,000 ([Federal Reserve Board \(2010\)](#)). Similarly comparing the ratio of the frequency of 200 to 300 shares is difficult because institutions break up large trades into smaller trades to conceal their identity ([Kyle \(1985\)](#)) so that others may not profit from information they possess. Hence the interpretation of more sophisticated investors clustering their trades less on Prominent Numbers is confounded by their preference for small trades. This preference for small trades is illustrated by the ratio of 100 to 200 share trades which increased from 3.9 in the lowest institutional ownership decile to 9.9 in the highest decile, even though both 100 and 200 are Prominent Numbers (these results are not reported in Table S2).

Comparing retail to institutional investors’ trades reveals a pattern that is largely consistent with the conjecture. At most trade sizes, retail investors—those who are less sophisticated and who therefore would have to invest more cognitive effort to make optimal decisions—make decisions that reflect a stronger reliance on Prominent Numbers than do institutional investors. It is notable that retail investors typically have smaller portfolios than institutional investors and therefore more of their wealth allocation is affected by this decision-making bias. The decision to trade 500 rather than 400 shares of a \$10 stock is a difference of \$1,000, a significant portion of the average family’s equity holdings of \$29,000 ([Federal Reserve Board \(2010\)](#)). We have no way to determine the extent to which institutional investors are basing their decisions on computer algorithms, but whether this explains much or little of the difference from retail investors, the basic conclusion is the same.

3 Data Collection Plans for Studies 4a—4c

The sample-size decision for Study 4a was a largely heuristic decision, aiming to be able to detect a “small effect” ($d_z = 0.30$) with 90% power, we pre-determined a target sample of $N = 120$ (using G*Power; [Faul et al. \(2007\)](#)). For each subsequent decision, we incorporated the observed effect size with our expectation that we could refine aspects of the procedure to increase the size of the effect. After observing $d_z = 0.13$ in Study 4a, we thought appropriate refinements could increase the effect size to $d_z = 0.23$. For Study 4b, we were aiming to detect

this effect size with 90% power ($N = 201$, which we rounded to $N = 200$). However, there was a mistake in posting the study and we set it for $N = 300$. When we noticed the sample going above 200, we pulled the study, allowed all ongoing surveys to be completed and decided before examining the data to include the additional responses, yielding the final sample size. For Study 4c, we again expected that refinements would be able to increase the actual effect size, and we were comfortable over-sampling to provide some closure on this series of studies, so we determined the sample size aiming to detect $d_z = 0.25$ with 99% power ($N = 296$, which we rounded to $N = 300$).

4 Supplemental Studies: The Relative Graininess of Round and Prominent Numbers

We have suggested that the critical difference between Prominent Numbers and Round Numbers is chronic accessibility. A corollary to this assumption is that Prominent Numbers and Round Numbers do not differ on perceived graininess. One of our approaches to arguing for the accessibility-based model of open numerical judgment is to use discriminant validity. The logic is that if a given independent variable (e.g., cognitive load) increases Prominent Numbers but not Round Numbers, it is likely due to accessibility. However, if the Prominent Numbers are also perceived as grainier than the Round Numbers, then the effect of a variable like cognitive load could be attributable to the desire to communicate graininess instead.

We present here three studies that tested for differences in perceived graininess between Prominent Numbers and adjacent Non-Prominent Round Numbers. The critical reader who worries that we may be biased to set up experiments that would not produce such a difference should know that we ran these studies expecting to find a difference. At that stage of the project development, we had predicted that “the Prominent Numbers may be rounder than the Round Numbers.” We designed and conducted these studies aiming to show direct evidence of this difference. Obviously our own intent or motivation would not overcome any inherent design flaws, but we note this as one potentially relevant factor for a reader who is concerned about Type-II errors stemming from experimenter biases in these studies.

We took three different approaches to these experiments. In Supplemental Study 1 (“Study S1”), we attempted to assess intuitive or even automatic perceptions of roundness.

In a speeded-judgment task, we presented pairs of numbers and asked participants to evaluate them as quickly as possible on a number of dimensions, including “roundness.” In Study S2a, we moved to explicit ratings. We presented very short vignettes in which one character offered a price estimate for some object and we asked participants to evaluate the precision of the estimate (as the inverse of “roundness”). In Study S2b, we took the most indirect route, relying on the accuracy—informativeness tradeoff demonstrated in previous research (Yaniv and Foster (1995)) to infer the graininess that participants ascribed to different estimates.

4.1 Study S1: Speeded Comparisons

In this study, participants completed a series of rapid number judgments comparing two numbers. In critical trials, we presented two numbers, one Prominent and one Round, and asked participants to choose, “Which number feels ... ROUNDER? [e.g., 50 or 60].” To encourage rapid responding and “gut-feeling” judgments, we interspersed these roundness judgments with filler judgments of which number is larger, which numbers is smaller, and which number feels luckier. If participants perceive Prominent Numbers as grainier than adjacent Round Numbers, then they should choose the Prominent Numbers as “rounder” in head-to-head comparisons more often than would be expected by chance.

4.1.1 Method

We recruited MTurk participants for a judgment study that would take “about 3 minutes” (paying \$0.37). We aimed to recruit 100 participants for this study (and ended up with $N = 101$ complete responses). With this sample size, a given option would need to be selected 60% of the time to be recognized as significantly different from chance (using a two-tailed z-test against the chance rate of 50% at $\alpha = .05$).

Stimuli. We created 66 comparison trials from a list of seven two-digit stimulus numbers [20, 30, 35, 40, 50, 57, 60]. This list gave us two Prominent Numbers (20, 50), their adjacent Round Numbers (30, 40, 60), and two more precise filler numbers (35, 57). For the roundness judgments, we included all 21 possible pairwise comparisons that could be made from the 7 stimuli. To keep the study short, we randomly selected only 14 out of the possible 21 pairwise comparisons for the three filler judgments. Trials were presented in completely randomized order.

Procedure. Participants read the following instructions:

This is a study about rapid number judgments. We will ask you a series of questions comparing two numbers. There will be 66 comparisons, which will change between questions. Some questions have an objectively right answer, for example: "Which number is larger?" Some questions are more about your feeling, for example: "Which number feels luckier?" Please try to answer quickly and accurately.

Please respond as QUICKLY as possible to each question.

We are interested in your RAPID RESPONSES and GUT FEELINGS.

The point of these instructions was to convey a cover story that would prompt participants to use their “gut feelings,” and discourage them from approaching ‘roundness’ as a mathematical construct. (In other words, we were aiming to capture participants’ sense of the graininess of the numbers rather than their roundness.) In addition to the critical judgment, “Which number feels **ROUNDER?**,” we asked participants to make three kinds of (filler) comparisons: (1) Which number is **LARGER?**, (2) Which number is **SMALLER?**, and (3) Which number feels **LUCKIER?** On each trial, the two numbers were presented in randomized order.

Our focal dependent measure was the percentage of trials in which participants selected a Prominent Number over an adjacent Round Number (i.e., 20 over 30, 50 over 40, and 50 over 60). We refer to these as the critical-adjacent trials. As a secondary measure, we also examined the percentage of trials in which participants selected a Prominent Number over any Round Number (i.e., the three critical-adjacent trials, in addition to 20 over 40, 20 over 60, and 50 over 30). We also examined each choice separately to test whether people systematically perceived any of the Prominent Numbers to be “rounder” than any of the non-Prominent Round Numbers.

4.1.2 Results

First, to determine if people systematically identified Prominent Numbers as “rounder” than adjacent Round Numbers (overall), we used a one-sample t-test to determine whether the average percentage of critical-adjacent trials was significantly greater than 0.50 (the proportion that would be expected by chance). We found that the average ($M = 0.48$, $SD = 0.28$) was no different than would be expected by chance, $t(100) = -0.70$, $p = .49$. Casting a slightly broader net, we also asked whether people systematically identified the Prominent Numbers as “rounder” than the Round Numbers in general (not just those that

were adjacent). This average ($M = 0.48$, $SD = .22$) was similarly indistinguishable from chance, $t(100) = -0.92$, $p = .36$.

To get a more complete view, we also examined each of the six comparisons in isolation. Table S3 shows the selection percentage for each of the 6 critical trials. (Two comparisons, 20 versus 30 and 20 versus 60, have only $N = 100$ responses, respectively, because of non-responding.) To be considered significantly different from the chance proportion of 0.50 at $\alpha = .05$ (using a two-tailed test), one of the options would have to be selected with proportion 0.60.

Table S3: Proportion of Participants Selecting the Prominent Number in Each Comparison

	Adjacent Comparisons			Non-Adjacent Comparisons		
	20 vs. 30	50 vs. 40	50 vs. 60	20 vs. 30	50 vs. 40	50 vs. 60
proportion	0.55	0.51	0.38	0.59	0.42	0.43
z	1.00	0.20	-2.41	1.81	-1.60	-1.41
p	.32	.84	.02	.07	.11	.16

Table S3 shows that only one of the comparisons approached that mark in favor of the Prominent Number, with 20 selected as rounder than 40, but 20 was also perceived as less round than 60 at a marginal level of significance, and it did not significantly differ from its best comparison number, the adjacent Round value of 30. Similarly, 50 was not systematically perceived as rounder than any of its comparisons, not differing from the adjacent Round value of 40, and actually being perceived as less round than 60. It is certainly possible that 60 is an “unfair” comparison given its base-status in the sexagesimal system commonly used for time. However, even granting this exception, there was not strong evidence for systematic perception differences in the cleanest comparisons of 20 versus 30, nor 50 versus 40.

4.2 Study S2a: Explicit Judgments

The logic of this study was that if people perceive Prominent Numbers as “rounder” (grainier), then they should report that those numbers are less precise when offered by others as an estimate. To test this, we presented participants with a series of short vignettes in which one character offers a price estimate to another character. We asked participants to rate the

estimate on its precision. We used a between-subjects experimental design, with participants assigned to see a series of Prominent-Number estimates (interspersed with non-Prominent fillers) or a series of non-Prominent Round-Number estimates (interspersed with the same non-Prominent fillers). If participants perceive Prominent Numbers as grainier than adjacent Round Numbers, then Prominent Numbers should be rated as consistently less precise.

4.2.1 Method

We recruited MTurk participants for a judgment study that would take “about 4 minutes” (paying \$0.46). We aimed to recruit 160 participants, and received a total of $N = 162$ complete responses.

After consenting, participants learned that the study was assessing “how people feel about numbers.” In the “first part” (Study S2a), participants learned that they would read seven very short conversations. In each, the first person would ask the second person how much he/she thought something costs, and the second person would make a guess. Participants further learned that we would ask them after each conversation to give their “gut feeling about how precise of an answer the second person is giving.” We explained that we were not asking whether the answer was accurate, but were focused on its precision: “For example, most people would say that an answer of \$100 feels less precise than an answer of \$116.42 regardless of what they think the true value is.” To orient them to the scale, we instructed, “If the answer seems very precise or specific, you would respond by selecting the higher end of the scale. If the answer seems very imprecise, general, or ‘round,’ give an answer toward the lower end of the scale.” Finally, we gave a last note asking them not to “over-think,” but to “just give your own personal feeling based on the little information that you have for each case.”

Participants saw 7 short vignettes (in fixed order). The odd-numbered vignettes (i.e., 1, 3, 5, and 7) were each a critical comparison and the even-numbered vignettes (i.e., 2, 4, 6) were each filler vignettes (using non-Round numbers). As an example, here is vignette 1 in the Prominent [and Round] condition:

Person 1: “I’m thinking about getting a new TV like that one. How much do you think I’d have to pay for a TV that big?”

Person 2: “Probably around \$500 [\$400].”

Participants then responded to the prompt, “Person 2’s estimate feels like ____” on a 101-point slider scale with anchors 0 = a very IMPRECISE estimate for a [TV] to 100 = a very PRECISE estimate for a [TV]. The objects in the anchors were tailored to each question. The full list of conversations appears in Supplementary Appendix 1. The four critical objects with the [Prominent / Round] price were: (i) TV [\$500 / \$400], (ii) plates [\$20 / \$30], (iii) car [\$50,000 / \$60,000], and (iv) house [\$500,000 / \$400,000]. Thus, all of the Round values were adjacent to the respective Prominent Numbers, with two taking the lower adjacent Round value and two taking the higher adjacent Round value.

4.2.2 Results

We started by combining the four critical judgments into a combined precision index (Cronbach’s $\alpha = .77$). There was no difference in the perceived precision of the Prominent-Number estimates ($M = 41.6$, $SD = 21.0$) compared to the Round-Number estimates ($M = 41.3$, $SD = 21.6$), $t(160) = 0.10$, $p = .93$. Table S4 shows that on an item-by-item basis, none of the Prominent Numbers were perceived as significantly ‘rounder’ than their Round counterparts.

Table S4: Precision Ratings (Lower Values Represent ‘Rounder’ Perceptions)

	M (SD) Prominent Estimate	M (SD) Round Estimate	t	p
Television (500 vs. 400)	41.4 (28.4)	34.1 (26.0)	1.70	0.09
Plates (20 vs. 30)	45.5 (27.5)	48.1 (27.7)	-0.61	0.55
Car (50k vs. 60k)	34.5 (25.0)	40.5 (25.5)	-1.52	0.13
House (500k vs. 400k)	45.0 (30.0)	42.7 (30.6)	0.47	0.64

4.3 Study S2b: Inferred Graininess from Accuracy-Informativeness Tradeoffs

The measure in Study S2b was appended to the end of Study S2a. It is not formally a separate study—it is the same group of participants from Study S2a—but we describe them separately for ease of presentation. Two participants did not respond to one of the critical questions in Study S2b and are excluded from these analyses (final $N = 160$). In Study

S2b, we took the logic of the accuracy—informativeness tradeoff (Yaniv and Foster (1995)) one step further than we did for the measure in Study S2a. Yaniv and Foster proposed and empirically supported an additive trade-off model, based on the following basic logic (p.426):

The definition of accuracy as a “ratio of error to precision” implies that coarser judgmental estimates are likely to be more accurate. However, coarser judgments are less informative. This naturally creates a trade-off between accuracy and informativeness.

They go on to propose a formal model to represent this tradeoff. Error is represented by the absolute value of the difference of the true answer and the given estimate, divided by the width (or graininess) of the implied range. Informativeness is represented by the log of the width of the implied range. Adding the two terms gives a preference score, with lower values preferred. The central insight is that a grainier judgment (wider implied width) decreases the error score (a good thing) but also increases the informativeness score (a bad thing).

An example they give involves two persons estimating the amount of money that Michael Jackson received to star in a series of 1987 Pepsi commercials. The correct answer turns out to be \$15mm. Person A’s estimate was \$1mm—\$20mm. Person B’s estimate was \$12mm—\$14mm. Intuitively, one can see that Person A’s range was “correct,” in that it included the correct answer, whereas Person B’s range did not include the answer. Yet, Person A’s answer was so grainy that it may have been relatively uninformative compared to Person B’s more specific answer. Turning to the formula, the error score for Person A indeed represents that she had a lower error score, $((15 - 10.5) / 19) = 0.24$, compared to Person B, $((15 - 13) / 2) = 1.00$; but that Person A’s informativeness score was much higher, $(\ln(19) = 2.94)$, than Person B’s, $(\ln(2) = 0.69)$. Adding the respective scores, Person A has a higher preference score, $(0.24 + 2.94 = 3.18)$, than Person B, $(1.00 + 0.69 = 1.69)$, suggesting that most people would prefer Person B’s judgment.

Attempting to capitalize on this logic, we identified a critical case that should in theory have allowed us to determine, to a relative degree, the graininess people assumed for various single-point estimates. In short, we asked people to evaluate a series of point estimates for a quantity whose true value turned out to be 54mm. The two critical evaluations were estimates of 50mm (a Prominent Number) and 60mm (a Round Number). If the perceived graininess of 50mm and 60mm were about the same, then, according to the accuracy—informativeness tradeoff, people should prefer the estimate of 50mm. (It is closer, after all.)

However, if people perceive 50mm to have a higher graininess than 60mm, they might prefer 60mm overall.

For example (dropping the millions for ease), if 50 implies “45 to 55” and 60 implies “55 to 65,” then 50 would have a lower (preferred) score:

$$\begin{aligned} \text{Error}_{50} &= |50 - 54| / 10 = 0.40 & \text{Error}_{60} &= |60 - 54| / 10 = 0.60 \\ \text{Inform}_{.50} &= \ln(10) = 2.30 & \text{Inform}_{.60} &= \ln(10) = 2.30 \\ \text{Pref}_{50} &= 0.40 + 2.30 = 2.70 & \text{Pref}_{60} &= 0.60 + 2.30 = 2.90 \end{aligned}$$

In contrast, 50 might imply a wider interval than 60. One possibility, following a strict interpretation of Albers and Albers’ model (1983) of how people construct judgments, is that 50 might be taken to imply “35 to 75” because any signal below 35 would be better represented by 20, whereas any signal 75 or above would be better represented by 100. By this same logic, 60 would imply “55 to 65” because any signal below 55 would have stayed at 50 instead of going to 60 = (50 + 10), whereas any signal above 65 would have jumped to 70 = (50 + 20).

$$\begin{aligned} \text{Error}_{50} &= |50 - 54| / 40 = 0.10 & \text{Error}_{60} &= |60 - 54| / 10 = 0.60 \\ \text{Inform}_{.50} &= \ln(40) = 3.69 & \text{Inform}_{.60} &= \ln(10) = 2.30 \\ \text{Pref}_{50} &= 0.10 + 3.69 = 3.79 & \text{Pref}_{60} &= 0.60 + 2.30 = 2.90 \end{aligned}$$

Under these assumptions, people should now prefer 60 even though 50 is closer to the true value. This order holds even if we relax the assumptions so that 60 implies a wider interval of “50 to 70.” This yields a preference score of 3.30, which would still be preferred to 3.79. Notably, it also holds if we take what some might say is a more intuitive implied range of 50 as “25 to 75,” rather than using the strict [Albers and Albers \(1983\)](#) construction model.

To test whether these implied ranges might be plausible, we presented participants with evaluations that were in a similar format as the [Yaniv and Foster \(1995\)](#) problems, but with some adaptations. First, because we were interested in the implied range, the estimators in our scenarios gave a point estimate rather than an explicit range. Second, we did not ask participants to directly rank the estimates based on preference because we thought it would be an unfair standard to ask participants to directly rank 60 over 50 as a better estimate of 54. Instead, we used a continuous measure of the estimator’s competence. The hypothesis here was that, if 50 is perceived as (substantially) grainier than 60, then 60 should be evaluated as a more competent judgment of 54.

4.3.1 Method

Participants completed this measure (“Part 2”) after evaluating the conversations in S2a (“Part 1”). They read the following scenario:

In a meeting, a boss asked employees to write down the following value without looking it up: *How many millions of dollars (USD) did Snapchat pay to acquire the company Scan (for its scanning app)?*

Each employee writes down a number on a slip of paper without hearing others’ estimates.

The boss knows the actual answer is \$54 million.

We will show you the estimates that 6 different employees offered and ask you to evaluate those estimates from “extremely incompetent” to “extremely competent.”

Participants then saw, one by one on separate pages, a prompt that asked “How many millions of dollars did Snapchat pay to acquire the company Scan? Actual Answer: \$54 million. One employee responded: ‘\$[50] million.’ How will the boss evaluate this employee’s answer?” Participants responded on a 7-point scale, from 1 = *Extremely incompetent* to 7 = *Extremely competent*. They saw six different estimates in random order. The critical trials were 50 and 60; and four trials were fillers (12, 30, 42.5, and 55) intended to discourage direct comparison between the two critical trials.

4.3.2 Results

We used a paired-t test to determine whether participants saw \$50mm as a relatively less competent judgment than \$60mm. We found strong evidence of the opposite, indicating that people saw \$50mm as a more competent judgment, $M = 6.00$, $SD = 1.00$, than \$60mm, $M = 5.64$, $SD = 1.04$, $paired-t(159) = 4.77$, $p < .001$. This is not at all surprising if people see the two as approximately equally grainy because then there is little tradeoff in informativeness, but an obvious accuracy advantage for \$50mm over \$60mm.

4.4 Summary of Studies S1, S2a, and S2b

We took three different approaches to testing whether Prominent Numbers are perceived as grainier than adjacent non-Prominent Round Numbers, including a range of more direct and

indirect measures. None yielded systematic support for the assumption. We did not employ an equivalence-test approach that would allow us to more formally support the absence of a meaningful effect (Lakens (2017)), but in each of the three approaches, the observed effect approached zero (S2a) or went in the opposite direction (S1 and S2b). This suggests a fairly strong basis for concluding that the current observations are not merely Type-II errors. That said, other procedures, manipulations, or measurement procedures could well have yielded different results, thus we do not intend to suggest a strong general conclusion that “Prominent Numbers are not grainier than Round Numbers.” We merely cannot identify evidence that they are. To better support a stronger conclusion like this would require a much more extensive array of tests, with more developed procedure validation than we have here.

4.5 Supplemental Study S3: Task Motivation and Prominent-Number Clustering

Supplemental Study S3 is a study that we removed from the main text during the review process due to length considerations. In it, we examined archival data that pair a behavioral measure of an individual’s task motivation with an open numerical judgment, allowing us to test for an association between motivation and the use of Prominent Numbers. We predicted that the effort voluntarily invested in an experimental task would inversely predict the likelihood of using a Prominent Number.

The data come from a set of unpublished pilot studies (Reference blinded for review (2016)), the aim of which was to validate a screening procedure for use with online populations. Participants were “workers” on Amazon Mechanical Turk (MTurk). Within a significantly longer survey were two measures critical for the current analysis. Both came from a section of the survey that was specifically about the MTurk worker experience. The motivation measure came from an idea-listing task in which we solicited participants’ ideas for improving the experience of workers on the site. They were specifically told they could “list as few or as many as you would like” and 22 lines were available for entry. We operationalized task motivation as the number of ideas they listed. Immediately preceding this was the open numerical judgment, which asked “How many HITs would you guess you have done in your career?” “HITs” stands for Human Intelligence Tasks, the basic unit of work on MTurk. We coded whether the HITs estimate was a Prominent Number or not. We also

coded whether the HITs estimate was a Round Number.

4.5.1 Method

We assembled the Study S3 dataset from three waves of data collection, which we will refer to as Study A, Study B, and Study C. These represent all datasets that we could identify in our archives that include the two key measures for Study S3, a measure of voluntary behavioral effort and the open numerical judgment regarding career-work completion on MTurk. We identified 622 participants who had a valid observation for each of these measures.

Overview. The studies were part of an informal lab effort to test different methods of screening MTurk participants based on their personal interests. One concern with using MTurk to recruit participants who have a particular interest, such as “being a football fan,” is that participants may be incentivized to exaggerate their interest, potentially leading to biased assessments of sample characteristics. Studies A, B, and C, were an initial attempt to explore the efficiency of an “embedded” screening mechanism (in contrast to an explicit, self-report screening mechanism such as, “Please only respond if you are a football fan”). Instead of recruiting self-identified fans, we embedded a number of behavioral report questions that we expected would better identify “true fans.” To do this, we posed a variety of filler questions about one’s media habits and, within those longer surveys, we embedded behavioral reports that might identify highly committed NFL fans. Specifically, we asked questions about premium cable subscriptions to football channels (Study A) and weekly football viewing habits (Studies B and C). Using this method, we could use automatic branching in the survey software to send “true fans,” identified by some self-reported behavioral threshold, to the critical block of questions while diverting others to a different study or a set of filler questions that would take about the same amount of time. All participants would be paid the same, reducing incentives to exaggerate interest. The downside for researchers is potentially paying many participants who do not “qualify” for the focal study. In Studies A, B, and C, we aimed to get an initial idea of the percentage of participants who would qualify using different behavioral thresholds and, by comparing across studies (hence, the “informality” of our test), assessing the relative effect sizes of different procedures that used explicit versus embedded recruitment methods.

Procedure. The studies all had four blocks of questions. First, participants responded to a number of questions about their media consumption habits (e.g., “How many hours

of television do you watch on a typical day?” [free response]; “How do you watch most of your television content?” [*traditional cable, via internet, other, not applicable to me*]. The embedded filter was included in this block. In Study A, the critical question was, “Which premium content do you have a paid subscription for? (check as many as apply to you).” There were nine choices (e.g., HBO, Apple Music, MLB.TV). The critical item that we used to surreptitiously identify committed NFL fans was if they subscribed to some premium football package, “NFL Football (Sunday Ticket or other).” In Studies B and C, the critical question was “Which content do you watch regularly? (at least 5 hours per week, typically). There were nine choices (e.g., Premium Movie Channels, Cooking Shows, Major League Baseball (when in-season)). The critical item that we used to identify committed NFL fans was if they indicated watching at least five hours of NFL Football coverage per typical (in-season) week.

The second block varied based on the screening items. Participants who met the designated threshold for being a “committed fan” were directed to a set of questions about their experience as a football fan ($n = 8$ out of 116 (6.9%) in Study A; $n = 69$ out of 209 (33.0%) in Study B; and $n = 83$ out of 312 (28.5%) in Study C). This block followed procedures employed in previous research on rivalry and motivation (Converse and Reinhard (2016)). (The screening and planned exclusion standards proved too strict to provide the needed sample sizes for testing any relevant rivalry and motivation hypotheses.) Participants who were not classified as “committed fans” were directed to a second block of media-consumption questions (e.g., “What are your attitudes about each of these music streaming services? {Apple Music, Spotify, Pandora} from 1 = *very negative* to 4 = *very positive*, or *no opinion*).

All participants then moved to the third block of questions, a survey about Workers’ experience using MTurk. Participants indicated their length of time using MTurk as a Worker, their frequency of using MTurk, and the number of HITs they had completed that day. Next, participants responded to the item that we used as the open numerical judgment: “How many HITs would you guess you have done in your career?” Following that, participants completed the motivation measure. By random assignment, half of the participants were asked about “Maximizing the Positive Aspects of Mechanical Turk.” They were asked to, “Please list the things that you think should be *added* to MTurk in order to ensure that workers gain as many positive experiences as possible while they are using MTurk. *List as few or as many as you would like.*” Participants could enter up to 22

ideas. The other half of participants completed the same items, but framed as, “Eliminating the Negative Aspects of Mechanical Turk.” They were asked to, “Please list the things that you think should be *eliminated* from MTurk in order to ensure that workers avoid as many negative experiences as possible while they are using MTurk. *List as few or as many as you would like.*” The manipulation was based on regulatory focus theory (Higgins et al. (2003)) and included because our screening pilots were run using procedures that included regulatory focus manipulations (as in Converse and Reinhard (2016)).

Finally, participants completed a short demographic block. We asked participants to report their age and gender. We also asked how they learned about the HIT {*MTurk Search*, *Reddit: “HITs worth turking for”*; *Other MTurk blog or forum* [open text box]; *Turk Alert*; *Twitter*; *Other* [open text box]}.

4.5.2 Results

Motivation distribution. The observed range on the motivation measure was 0-14 ($M = 2.47$, $SD = 1.62$, Med. = 2.0). The distribution was strongly right-skewed (Skew = 1.22), not a surprise given that the measure was bounded by 0 on the low side and unbounded on the high side. To avoid excessive influence of outliers, we first calculated the interquartile range (IQR) of the raw values and capped all outliers at 1.5*IQR above the third quartile ($n = 11$; capped Max. = 6).

Open numerical judgments. Estimates for career HITs ranged from 1 to 10,018. Though there are only 13 Prominent Numbers in this range (0.13%), 20.1% of all estimates used one of the Prominent Numbers. There are also 37 Round Numbers in this range (0.37%), and 43.1% of all estimates used one of the Round Numbers.

Motivation and Prominent Numbers. To test whether more motivated workers were less likely to use Prominent Numbers in their estimates, we conducted a binary logistic regression, regressing Prominent-Number usage (no = 0, yes = 1) on motivation. Motivation was a marginally significant negative predictor, $B = -.14$, Wald’s $\chi^2 = 3.73$, $p = .054$. Table S5 illustrates the pattern.

Participants who displayed the least motivation on the idea-listing task by skipping it altogether, were the most likely to use a Prominent Number in their experience estimate. Looking across groups who displayed increasing levels of motivation from there, the groups show an almost perfectly monotonic pattern of decreasing Prominent-Number usage.

Table S5: Prominent-Number clustering as a function of motivation

Motivation	N	Percent of Prominent Estimates
0	54	29.6
1	120	21.7
2	171	19.9
3	136	19.9
4	86	15.1
5	34	17.6
6	21	14.3

Note: The highest level of Motivation includes 11 outliers whose raw scores (range = 7 to 14) were capped using the $Q3 + 1.5 \cdot IQR$ rule. N is the number of participants exhibiting each level of motivation. The last column contains how many estimates of the number of HITs were Prominent Numbers for each level of motivation.

We next examined the specificity of the effect. Was motivation associated with using Prominent Numbers specifically or with Round Numbers more generally? Unlike in the extremely well-powered stock data, we cannot compare the frequency of each Prominent-Number response to its nearest Non-Prominent Round neighbor. However, if judges were selecting estimates based primarily on their graininess, then we would expect an analysis predicting Round-Number usage from motivation to be at least as reliable as the one predicting Prominent-Number usage from motivation. This does not appear to be the case. A binary logistic regression with Round-Number status rather than Prominent-Number status as the outcome is not significant, $B = -.05$, Wald's $\chi^2 < 1$, $p > .25$. (Though largely redundant with this analysis, we can illustrate how the inclusion of the Non-Prominent, Round Numbers dilutes the relationship by noting that motivation is not a significant predictor of producing a Non-Prominent Round estimate, $B = .05$, Wald's $\chi^2 < 1$, $p > .25$.) Together, we interpret these results as evidence that increased task motivation is related to decreased reliance on Prominent Numbers specifically.

Robustness checks. Motivation scores did differ as a function of framing condition, with participants volunteering more ideas in the promotion-framed version ($M = 2.57$, $SD = 1.49$) than in the prevention-framed version ($M = 2.29$, $SD = 1.44$), $t(620) = 2.37$, $p = .018$. This could reflect differences in motivation or differences in ease of the task, but either way warrants confirmation that the focal relationship between motivation and Prominent-

Number usage holds when accounting for the difference. We checked this in two ways. First, we standardized the motivation scores within each framing condition and then used the resulting motivation z-scores as predictors. A binary logistic regression predicting the use of Prominent Numbers from the motivation z-scores found a comparable, marginally significant negative relationship, $B = -0.19$, Wald's $\chi^2 = 3.31$, $p = .069$. A second robustness check used the non-standardized motivation scores and added a dummy-coded framing variable as a control (promotion = 1, prevention = 0) and again found similar results. Motivation was a marginally significant negative predictor of Prominent-Number usage, $B = -0.13$, Wald's $\chi^2 = 3.36$, $p = .067$ (and frame was not a significant predictor, $B = -0.10$, Wald's $\chi^2 = 1.00$, $p > .25$). The focal result seems to be independent of differences in the motivation distributions.

5 Supplementary Appendix 1 - Items from Study S2

- (1) Person 1: "I'm thinking about getting a new TV like that one. How much do you think I'd have to pay for a TV that big?"
Person 2: "Probably around \$500 [\$400]."
- (2) Person 1: "I was considering putting in a pool. How much do you think that would run me?"
Person 2: "It depends on what you want to do, but probably around \$12,500." [filler]
- (3) Person 1: "Those are really nice plates. How much do you think a set costs?"
Person 2: "I bet you could get a set for close to \$20 [\$30]."
- (4) Person 1: "I really like her smart watch. How much do you think it costs?"
Person 2: "I would estimate \$220." [filler]
- (5) Person 1: "I like that car. How much would you have to pay for a car like that?"
Person 2: "A car like that probably runs about \$50,000 [\$60,000]."
- (6) Person 1: "That's a sharp tie. How much do you think he bought that for?"
Person 2: "I'd put that at about \$79." [filler]
- (7) Person 1: "Man, I would love to live in that neighborhood. How much do you think a house there goes for?"
Person 2: "A house in that neighborhood is probably in the \$500,000- [\$400,000-] range."

References

- Albers, Wulf, and Glsela Albers, 1983, On the prominence structure of the decimal system, in *R. W. Scholz (Ed.), Decision Making Under Uncertainty* (Amsterdam: Elsevier).
- Barber, Brad M, and Terrance Odean, 2000, Trading is hazardous to your wealth: The common stock investment performance of individual investors, *Journal of Finance* 55, 773–806.
- Converse, Benjamin A, and David A Reinhard, 2016, On rivalry and goal pursuit: Shared competitive history, legacy concerns, and strategy selection, *Journal of Personality and Social Psychology* 110, 191–213.
- Fan, Jianqing, and Irene Gijbels, 1996, *Local polynomial modelling and its applications: monographs on statistics and applied probability*, volume 66 (CRC Press).
- Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner, 2007, G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, *Behavior Research Methods* 39, 175–191.
- Federal Reserve Board, 2010, Survey of consumer finances.
- Higgins, E Tory, Lorraine Chen Idson, Antonio L Freitas, Scott Spiegel, and Daniel C Molden, 2003, Transfer of value from fit, *Journal of Personality and Social Psychology* 84, 1140–1153.
- Kyle, Albert S, 1985, Continuous auctions and insider trading, *Econometrica* 1315–1335.
- Lakens, Daniël, 2017, Equivalence tests: a practical primer for t tests, correlations, and meta-analyses, *Social Psychological and Personality Science* 8, 355–362.
- Reference blinded for review, 2016, [screening out online participants who say they are interested even when they are not], Unpublished raw data.
- StataCorp, 2013, Stata 13 Base Reference Manual, College Station, TX: StatCorp L.P.
- Yaniv, Ilan, and Dean P. Foster, 1995, Graininess of judgment under uncertainty: An accuracy-informativeness trade-off, *Journal of Experimental Psychology: General* 124, 424–432.