Cognitive cost and information gain trade off in a large-scale number guessing game

Felix J. Binder*
Department of Cognitive Science
UC San Diego,
fbinder@ucsd.edu

Cameron R. Jones*
Department of Cognitive Science
UC San Diego,
c8jones@ucsd.edu

Robert A. Kaufman
Department of Cognitive Science
UC San Diego,
rokaufma@ucsd.edu

Naomi T. Lin
Department of Education
UC San Diego,
ntlin@ucsd.edu

Crystal R. Poole
Department of Cognitive Science
UC San Diego,
c1poole@ucsd.edu

Edward Vul
Department of Psychology
UC San Diego,
evul@ucsd.edu

Abstract

How do people ask questions to zero in on a correct answer? Although we can formally define an optimal query to maximize information gain, algorithms for finding this optimal guess may impose large resource costs in space (memory) and time (computation). To understand how people trade off the information gain and the computational difficulty of choosing the ideal query, we turned to a large dataset of 380,000 guesses made during a number-guessing game with Amazon Alexa. We analyzed whether the arithmetic difficulty of following the optimal strategy predicts how far a guess deviates from theoretically optimal query. We find that when memory load is higher, and when more arithmetic operations need to be performed, human guesses deviate more from the most informative query. These results suggest human computational resource constraints limit how people seek out informative questions.

Keywords: optimal experiment design; resource rationality; mental arithmetic; decision theory; information gain

Introduction

In situations as wide-ranging as medical diagnosis, asking for directions, and debugging software, people must come up with good questions to reach a suitable answer (Coenen, Nelson, & Gureckis, 2019). However, identifying the best question to ask is a difficult computational problem, even before considering the limited cognitive resources available to humans. How do people balance the need to find informative questions with the computational constraints of their cognitive resources?

Research on active learning and optimal experiment design reveals that in a wide range of circumstances, people are fairly effective at asking informative questions (Coenen et al., 2019; Nelson, 2005; Gureckis & Markant, 2012). People ask informative queries when classifying unfamiliar alien creatures (Nelson, 2005), finding battleships (Gureckis & Markant, 2009), and teasing out causal models (Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Cook, Goodman, & Schulz, 2011) Although people are sensitive to the informativeness of questions, they do not always ask the most informative questions. For instance, in a battleship game, participants selected the most optimal choice with a high frequency, but still chose non-optimal spaces often (Gureckis & Markant, 2009). What explains this deviation from optimality?

Finding the most diagnostic question can be a challenging computational problem, so it may be the case that people fail to find optimal questions because of their cognitive resource constraints (Coenen et al., 2019). Cognitive psychology has documented numerous limits in resources including memory (Baddeley, 1997), central processing capacity (Pashler, 1984), attention (Cavanagh & Alvarez, 2005), and processing speed (Ratcliff & Rouder, 1998). Recently, these cognitive constraints have been studied using algorithm analysis from computer science (Lieder & Griffiths, 2020; Gershman, Horvitz, & Tenenbaum, 2015; Dasgupta & Gershman, 2021), clarifying the roles that limitations to memory (space), computational speed (time), and language (communication throughput) have had on shaping human cognition (Griffiths, Lieder, & Goodman, 2015; Griffiths, 2020). Algorithm analysis allows researchers to incorporate resource constraints in models of cognition to postulate "resource rational" algorithms that balance task objectives with the execution costs (Lieder & Griffiths, 2020; Gershman et al., 2015). Such bounded-rational analyses have succeeded in explaining peculiarities of human cognition in domains ranging from sentence parsing (Levy, Reali, & Griffiths, 2009) to hypothesis generation (Dasgupta, Schulz, Goodman, & Gershman, 2018).

Despite thorough analyses of the effect of cognitive limitations on various inference and decision-making algorithms, how these cognitive limitations apply to active learning and optimal experiment design is less clear. Some analyses have indicated that people are able to maximize their information gain given what they already know; however, the relative efficiency of the posed question varies as a function of the search space (Gureckis & Markant, 2009). Cognitive limitations might influence human active learning in at least two ways. First, memory constraints mean that finding diagnostic questions in larger hypothesis spaces is harder, as it requires evaluating the diagnosticity of a question with respect to a larger set of hypotheses. Second, algorithms for finding the most diagnostic question might be limited by time – some hypoth-

* indicates equal contribution

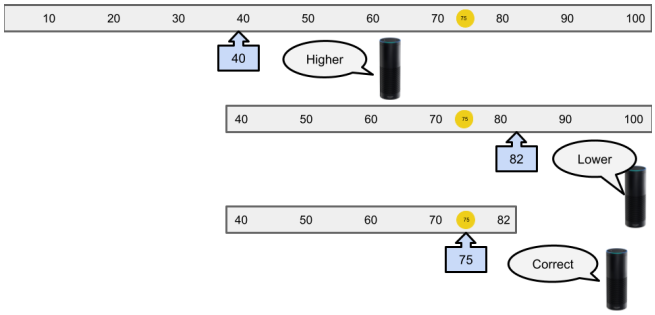

Figure 1: In the number guessing game, players try to guess which number between 1 and 100 a computer voice assistant has chosen. The player repeatedly queries the voice assistant with a guess and receives feedback of either "lower", "higher" or "correct" as the answer. Each guess rules out a part of the number space, and the game ends when the player has guessed the correct number. The optimal strategy—binary search—is guaranteed to halve the interval in which the target number could lie at each guess, and will find the correct answer in at most 7 guesses.

esis spaces might be more conducive to efficient calculations of optimal questions. Although these possible influences of cognitive resources on active learning have long been evident, measuring the influences of these factors require a fairly constrained problem with well-defined cognitive constraints, and with large amounts of data.

To study the role of cognitive constraints on human information seeking, we turn to a large dataset showing human performance on the relatively simple Number Guessing Game on Amazon Alexa (Dobson, 2019). With the number guessing game, the integer interval domain allows us to explicitly map well-studied memory and computation constraints onto the arithmetic properties of the interval underlying each guess. This precision in turn allows us to measure how the efficiency of questions varies as a function of these cognitive constraints, to ask whether guesses for intervals imposing larger memory or computational costs are further from optimal.

Alexa Number Game

We used the Alexa Number Game dataset, which included 380,000 guesses across 50,000 games and 14,000 players, with date, outcome, and unique player IDs for all games. In each game, the computer randomly picks an integer between 1 and 100, and players have to find that number by making sequential guesses of integers (see figure 1). After each guess, players receive feedback about whether the target number matches their guess, or—if not—whether the target is higher or lower than the guessed number. Feedback constrains the plausible interval for the target, so that after guessing "50", and receiving the feedback "higher", the player knows that the target number is in the interval bounded by 51 and 100. The target number can be reached in the fewest number of guesses, on average, by following a binary search strategy: guessing the midpoint of the current interval. Given the configuration of the game, the binary search strategy not only yields the earliest correct answer, on average, but also maximizes the expected information gain of the query.

This dataset offers several key advantages. First, the size of the dataset yields fine-grained measurements of human behavior in a variety of circumstances, rather than limiting analyses to particular pre-designed conditions. Second, this game reflects naturalistic information-seeking by voluntary players, reflecting the behavior of motivated individuals in lifelike situations, rather than in artificial lab settings where people might make different tradeoffs between cognitive cost and performance. Third, the game has a simple optimal solution, yet finding that optimal solution requires a variable amount of cognitive effort, depending on the specific interval in question. Fourth, the game sets up a straight-forward distribution over the target location: it is equally likely to be any point in the interval not yet already ruled out. This allows us to assume that the players correctly represent the uncertainty at each guess, which allows us to look for sources of errors beyond incorrect estimations of the potential locations of the target. Finally, the simple game structure — restricting hypotheses to a set of integers — allows us to use the rich literature on human cognitive limitations in integer arithmetic to define, a priori, the difficulty of information search in different circumstances.

Since this dataset reflects people interacting with their smart speakers in a variety of circumstances, the dataset necessarily contains an admixture of people who do not appear to be intentionally playing the game. Fortunately, the structure of the game means that completely inattentive play can be easily detected and discarded. The available dataset already excluded games that were not completed within 15 guesses, or that contained guesses that could not be parsed by Alexa, and we further excluded 33,442 games (176,107 guesses) that contained guesses outside the bounds of the current interval (45.8%). We also excluded from analysis 9,459 guesses where there was only one candidate within bounds, as no information can be gained in these circumstances. These exclusions left us with a dataset of relatively attentive play, consisting of 12,115 players, playing 32,480 games for a total of 198,487 guesses. To evaluate how well people pose questions in the number guessing game, we need to quantify the efficiency of a given query, as well as the algorithmic difficulty of posing a good question in a given circumstance. The next two sections describe these measures in detail.

Relative Expected Information Gain

How good is a particular guess? Although there is a large number of candidate criteria for optimal experimental design (Nelson, 2005; Coenen et al., 2019), in the case of the number game, their subtle differences are largely irrelevant. We will evaluate the quality of a guess in terms of how much information one expects to gain from the answer: how much uncertainty (as measured by entropy) is expected to be reduced

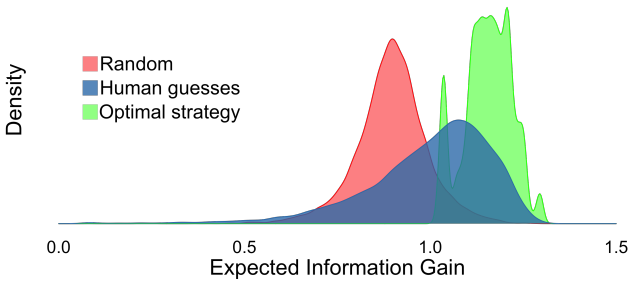

Figure 2: How much can we expect a guess to reduce our uncertainty about the target number? The higher expected information gain (*EIG*), the more informative the guess. The distribution of *EIG* of human players (blue) is compared to simulations of two key baselines: a random guessing strategy (red) and the optimal strategy of binary search (green). On average, human guesses are less efficient than optimal, but much more diagnostic than random.

by getting an answer to the question posed (Shannon, 1948). This "expected information gain" (Lindley, 1956; Fedorov, 1972) not only maps onto the optimal binary search strategy, but also explains human intuitions in a number of querying tasks (Nelson, 2005).[1] In our case, the expected information gain, for a given guess x is defined as:

$$EIG(x) = \mathcal{H}(a) - \sum_{b \in B} \mathcal{H}(b)p(b) \quad (1)$$

Where $\mathcal{H}(a)$ is the entropy (in bits) associated with the current interval (a), so if the interval is 1 to 78, $\mathcal{H}(a) = log_2(78) = 6.28$. B is the set of possible intervals after the feedback is received for guess x. So if one were to guess $x = 53$, the set of possible intervals is $B = \{[1,52],[53,53],[54,78]\}$. We average the entropy of each of these resulting posterior intervals, weighted by their probability, $p(b) = |b|/|a|$, to obtain an expected posterior entropy of the interval: $\sum_{b \in B} \mathcal{H}(b)p(b) = 5.7(0.67) + 0(0.013) + 4.64(0.32) = 5.29$. The expected information gain for guess x amounts to the difference between the current entropy, and the expected posterior entropy: how much the guess is expected to reduce our uncertainty. Here $EIG(x = 53) = 0.997$.

For the number guessing game, *EIG* is maximized for a simple optimal strategy: binary search. Binary search entails repeatedly dividing the list of possible outcomes in half to maximally narrow the list of possibilities. For example, if starting with the interval $[1,78]$, the first guess ought to be 39 or 40, and if given the response "lower", the next guess would be 19, which bisects the "lower" possibilities ($[1,38]$). On average, idealized binary search should be able to guess the number in $\log_2(n)$ guesses, where n is the number of candidates. For this game, n=100; thus, the number of guesses required is at most 7. Because of the possibility of the guess being correct, the optimal guess (39 or 40 for an interval of $[1,78]$) does not have an expected information gain of exactly 1 bit. Instead, in this case the optimal EIG is 1.086 bits. To evaluate the quality of a particular guess, we calculate a *relative* Expected information gain by normalizing the $EIG(x)$ of the guess by the optimal EIG^*: $rEIG = EIG(x)/EIG*$. This Relative Expected Information Gain measure asks what proportion of the maximum available information gain was realized by the participant's guess.

Arithmetic measures of algorithmic difficulty

In many domains, characterizing the cognitive constraints inherent in enumerating, evaluating, and searching over the hypothesis space is an open challenge. Fortunately, in the domain of numbers and arithmetic these constraints are well documented and each factor maps onto known cognitive limitations. These cognitive constraints fall into two general categories: limited space in working memory, and limited time for computations. In arithmetic, the size and quantity of numbers to be kept in working memory indicate the memory demand of the calculation (Dehaene, 2003), and the number of individual operations that need to be performed indicate the time cost.

In order to estimate the processing cost of calculating the optimal guess at each decision point, we make several assumptions about how the player might determine the midpoint of the interval. We assume that users store the lower bound, L, and upper bound, U, of the interval in which the target can lie. They then calculate the interval, $I = U - L$, and divide the interval by 2: $J = I/2$. Finally they add the quotient to the lower bound to find the midpoint of the two numbers, $M = L + J$.

We identify four features of this calculation which might index processing cost: interval size, midpoint, carrying, and number of operations. Formally, we define the interval size as $U - L$, and the midpoint as $(U + L)/2$. We identify 4 potential carrying operations throughout the three steps of mid-point calculation (subtraction, division, and addition). Carrying is required during subtraction if the ones-place of L is greater than the ones-place of U, during division if the tens-place digit and/or the ones-place digit of the I is not divisible by 2 (e.g. $34/2$; $43/2$), and during addition if the sum of the addends in the ones-place is > 10 (e.g. $5 + 7$). There are values of U and L for which all of these conditions are met (e.g. $U = 100$, $L = 23$), and others for which none are met ($U = 69$, $L = 41$). Therefore, the number of carries ranges from 0 to 4.

To define the number of operations, we decompose the midpoint calculation into single-digit operations (e.g. $42 + 12$ can be decomposed into $2 + 2$ and $4 + 1$) (Hitch, 1978). If we assume that adding or subtracting 0 from a single digit number does not constitute an operation (as the value of the other operand does not change), then there are 6 potential op-

[1] It is possible to evaluate actual information gained from each guess, conditional on the answer. However, such a measure would only noisily reflect the quality of questions people pose. Because, absent any clairvoyance, when people make a guess, they cannot know what answer will follow, therefore their choice of query must be made based on the expected, rather than observed, answer.

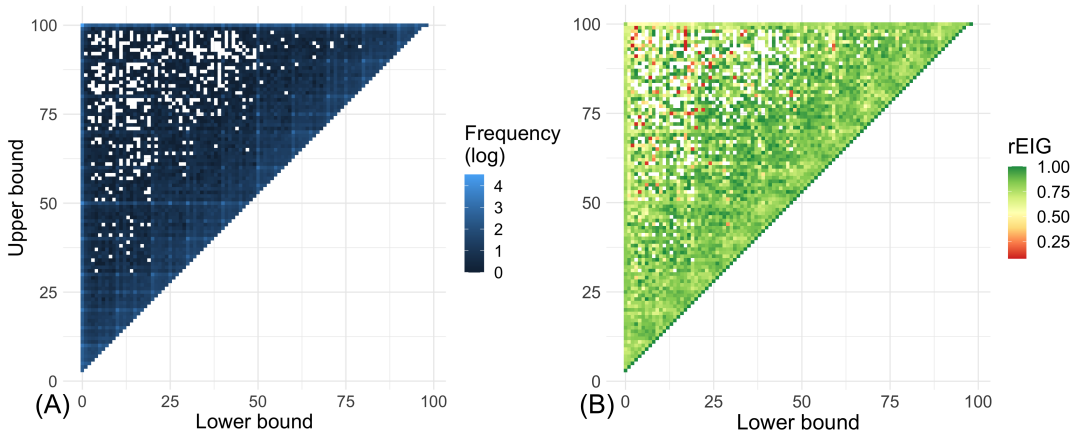

Figure 3: (A) Each pixel indicates the relative frequency of a particular interval the true number must be encountered by the players. Every game starts in the top-left corner. Note the regular patterns on 10s and 5s. (B) Human guess efficiency as a function of interval location. The average relative Expected Information Gain (*rEIG*)—how much information is gained relative to following the optimal strategy?—for each interval is shown.

erations: 2 during subtraction (one if $L > 10$, another if L is greater than zero in the ones-place); 2 during division (1 if $I > 10$, another is always necessary); and 2 during addition (1 if either $J < 10$ or $L < 10$, another if either J or L have a 0 value in the ones-place. The number of operations ranges from 6 (e.g. when $U = 96$, $L = 72$), to 2 (e.g. when $U = 80, L = 0$).

Both the size of the interval and the value of the midpoint place constraints on memory. As the size of the interval between the two numbers increases, the number of available hypotheses about the target number increases (Zbrodoff, 1995). As the magnitude of the bounds of the interval increases, keeping the size of the interval constant, a higher fidelity of representation is necessary to account for the change in relative proportion of size and bounds (Dehaene, 2003)—the midpoint of the interval measures the magnitude of the bounds. The need to carry or borrow terms when calculating requires additional numbers to be kept in working memory, and increases the number of operations that need to be performed (Imbo, Vandierendonck, & Vergauwe, 2007). Finally, each single-digit operation will take some amount of time and processing effort to perform. Therefore, as the number of single-digit operations increases, the total processing cost for the calculation should increase.

We predict that when the processing cost of finding the optimal guess as indexed by these measures is high, users will avoid incurring this cost by using easier but less accurate strategies. Therefore, we predict that as processing cost increases, the optimality of guesses—as measured by *rEIG*—will decrease. In other words, how informative the chosen guess is will depend on how difficult it is to identify the optimal query. This would indicate that users' querying behavior is sensitive to cognitive effort in determining the most informative guess.

Results

To establish a baseline against which to compare the optimality of human guesses, we ran two simulations. To simulate random guessing, we generated numbers from a uniform distribution with limits at the current lower and upper bounds. For optimal guessing, we calculated the optimal guess at each decision point using binary search. Figure 2 shows the distribution of *EIG* across guesses for each simulation and the user data.

User guesses had a significantly higher *EIG* on average (mean=0.997, sd=0.324) than would be expected if they were randomly guessing (mean=0.814, sd=0.164, $z = 30.64, p < 0.001$). However, the *EIG* of human guesses falls well short of what would be expected following the optimal strategy (mean=1.224, sd=0.241; $z = 278.3, p < 0.001$). This demonstrates that human players perform much better than chance, but much worse than optimal. Next, we examined whether the deviation from optimality can be explained by processing costs incurred during calculation of the optimal guess.

Effect of processing cost on optimality

An overview of the frequency and the mean *rEIG* for each interval is shown in Figure 3. Although some patterns are suggested in the heatmap, we aim to characterize the variation pictured, in terms of the arithmetic properties of each interval, as well as the computational difficulty it imposes on the guesser.

To investigate whether the difficulty of computing the midpoint of the interval predicts the *rEIG* of a guess, we constructed linear mixed effects models to predict *rEIG* using our measures of arithmetic difficulty. We constructed separate linear models, which tested the effect of each measure in isolation as well as a full model with all four measures included. All models were constructed using the lmerTest R package (Luke, 2017) and t statistics are computed via Satterthwaite's degrees of freedom method (Satterthwaite, 1946;

Table 1: Standardized coefficients for a full model predicting rEIG from the arithmetic properties of the interval.

	Std. Coef	Std. Error	df	t value	Pr($\|t\|$)
(Intercept)	0.882	0.002	78194.4	442	< 0.001
Interval size	-0.135	0.002	194588.6	-60.3	< 0.001
Midpoint	-0.011	0.003	194421.8	-4.2	< 0.001
Carrying	-0.078	0.003	194417.4	-30.2	< 0.001
No. of operations	-0.002	0.003	196532.3	-0.8	0.411

Luke, 2017). All models used the same random effects structure: random intercepts by user.

There was some ambiguity about how users might update their bounds. A rational user would update their bound to exclude their most recent guess: after guessing "50" and receiving "higher" as feedback, the players would update their lower bound to 51. However, we hypothesized that many players may have stored their most recent guess itself as the bound due to its saliency. We ran a version of our full model with and without adjustment for this effect. Then, we computed the Akaike information criterion for both (AIC adjusted: -25407, AIC unadjusted: -24584), which indicates that the adjusted model better fits the data. All reported results and figures therefore include this adjustment.

Interval Size Larger intervals impose higher memory loads by virtue of requiring that a larger hypothesis space be kept in mind. Insofar as memory limits the efficiency of posed questions, we would expect lower *rEIG* for larger intervals. Figure 4 confirms this effect of interval size on *rEIG*: *rEIG* decreases with increasing interval size $t(191700.8) = -53.89, p < 0.001$; 95% CI $[-7.940 \times 10^{-4}, -7.380 \times 10^{-4}]$. This effect holds in the full model (Table 1), controlling for other arithmetic properties of the interval: $t(194588.6) = -60.312, p < 0.001$; 95% CI $[-1.010 \times 10^{-3}, -9.440 \times 10^{-4}]$. Less informative guesses, from participants when there are large intervals, is consistent with the interval size imposing a working memory load. However, part of the effect is driven by poor performance at interval sizes between 50 and 100. These interval sizes can only contain data from people who had made particularly inefficient early guesses, and thus, ought to be expected to make bad guesses in the future. However, this adverse selection process is not responsible for this effect, for even when we consider only interval sizes less than 50, the negative linear trend still holds ($t(120963.54) = -13.028, p < 0.001$; 95% CI $[-7.100 \times 10^{-4}, -5.240 \times 10^{-4}]$).

Midpoint The magnitude of numbers is known to cause delays in processing during mental arithmetic (Dehaene, 2003), consistent with a role for the approximate number system in such calculations. If the approximate number system influences the efficiency of search, we would expect intervals bounded by larger numbers to yield less efficient guesses. The midpoint of the upper and lower bounds indexes the mean magnitude of both bounds and is orthogonal to interval size. Therefore, as the midpoint increases we expect the cost of calculating the optimal guess to increase and therefore *rEIG* of each guess to decrease. A linear model confirms this negative effect ($t(191472.5) = -11.93, p < 0.001$; 95% CI $[-3.070 \times 10^{-4}, -2.210 \times 10^{-4}]$; Figure 5). This effect holds in the full model, controlling for other predictors, including interval size ($t(194421.8) = -4.17, p < 0.001$; 95% CI $[-1.520 \times 10^{-4}, -5.470 \times 10^{-5}]$). Although these results do support our hypothesis, the effect size is negligible (an expected change of ~ 0.01 *rEIG* across the whole range of possible midpoint values).

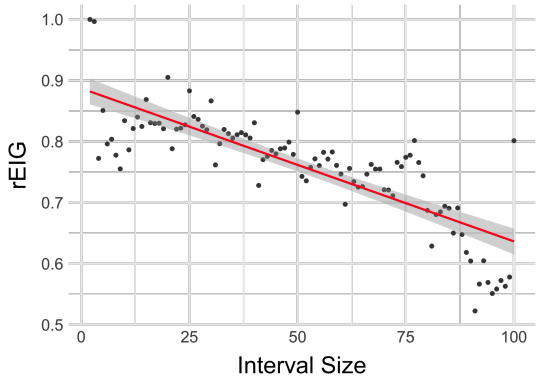

Figure 4: Relative Expected Information Gain as a function of interval size. *rEIG* is lower when the interval is larger ($r = -0.147$), with some notable exceptions at 100 (first guess), 50 and 25 (second and third guess following the optimal strategy), and 2-3 (final guess).

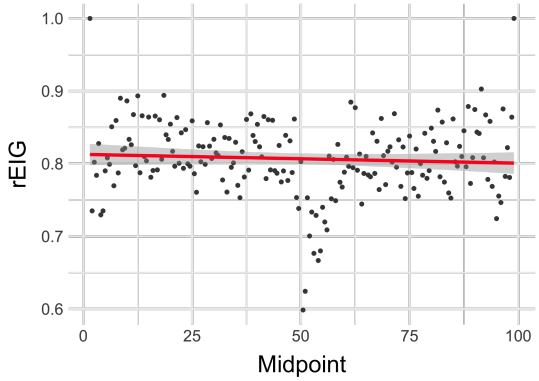

Figure 5: *rEIG* as a function of the magnitude of the interval midpoint. rEIG decreases as midpoint increases ($r = -0.026$).

Carrying When doing multi-digit arithmetic, carrying across decimal places requires one to temporarily store additional pieces of information and perform additional operations on an internal representation of the number. Carrying thereby imposes costs both in terms of memory and the time needed to perform a calculation (Imbo et al., 2007). We therefore expect users' performance to decrease as the number of required carry operations increases. A linear model confirms this effect ($t(195207.03) = -17.52, p < 0.001$; 95% CI $[-0.009, -0.007]$; Figure 6). The marginal effect of carrying remains even when controlling for other factors including midpoint, and thus the overall size of the num-

bers in play ($t(194417.4) = -30.17, p < 0.001$; 95% CI $[-0.018, -0.016]$).

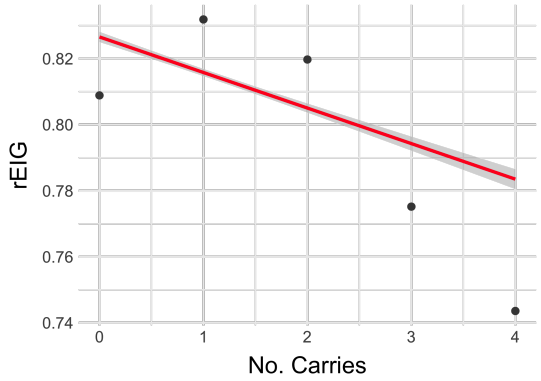

Figure 6: *rEIG* as a function of the number of carries required for a calculation. rEIG decreases as the number of carries increases ($r = -0.049$).

Number of Operations Finally, the total number of arithmetic operations involved in finding the midpoint is expected to decrease efficiency. Indeed, we found an effect on *rEIG* when marginalizing over all other variables ($t(198070.66) = -2.698, p = 0.007$; 95% CI $[-1.760 \times 10^{-3}, -2.783 \times 10^{-4}]$.

However, the effect is not robust to controlling for the effect of other factors. In the full model it is attenuated and no longer significant ($t(196532.3) = -0.822, p = 0.41$; 95% CI $[-1.406 \times 10^{-3}, 5.759 \times 10^{-4}]$). This likely reflects that the negative marginal effect above arises from correlations with factors such as interval size, midpoint, and the number of carry operations.

Discussion

We found that the efficiency of questions posed by humans, compared to the optimal question, varies systematically as a function of the magnitude and arithmetic properties of the current interval. People were less efficient when faced with larger intervals, intervals comprised of bigger numbers, and intervals requiring more elaborate arithmetic operations. Together, these results highlight the role of cognitive limitations and processing constraints in shaping human information search. Understanding how people trade off cognitive ease with their effectiveness in gathering information is crucial to explaining how people can efficiently make sense of an uncertain world.

The finding that the midpoint of the interval predicts reduction in *rEIG* independent of the size of the interval hints at an additional hypothesis on mental arithmetic: that certain computations are not actually computed, but merely retrieved, or estimated from the approximate number system (Dehaene, 2011; Zbrodoff, 1995).

These findings should not be taken to suggest that players are incapable of the simple operations needed to calculate the optimal guess: it is to be expected that given sufficient time and motivation, most players would be able to solve the calculation required to find the optimal guess. Rather, these results indicate that people trade off information gain and cognitive effort. Because the queries were undertaken in a game rather than an explicit study that rewards people for their participation, the reduction in optimality shows how people weigh information gain against effort in play, without an external reward. As such, this finding hints not at absolute limitations of human cognition, but rather at the weighing priorities and effort: people choose to accept less information gained per guess when the alternative is to expend significant cognitive effort. As there is no explicit cost associated with multiple guesses, performing multiple less-than-optimally informative guesses can be a rational strategy when determining the optimal query is time consuming. Without information on how long each guess took, this hypothesis is hard to confirm.

In sum, our paper presents a novel source of evidence for the claim that information gathering behavior is guided by cognitive effort. This connects fundamental measures of information to the human cognitive limitations and suggests that humans balance the cost and gain of gathering information.

Acknowledgments

The authors would like to thank Sam Dobson for the provision of the dataset as well the developers of the Alexa number guessing game.

All code and materials available at:
`https://github.com/felixbinder/number_guessing_game`

References

Baddeley, A. D. (1997). *Human memory: Theory and practice*. Psychology Press.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, *9*(7), 349–354.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*(5), 1548–1587.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349.

Dasgupta, I., & Gershman, S. J. (2021). Memory as a computational resource. *Trends in Cognitive Sciences*.

Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, *178*, 67–81.

Dehaene, S. (2003). The neural basis of the weber–fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences*, *7*(4), 145 - 147.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. Oxford University Press.

Dobson, S. (2019). *Guessing-game-ml-dataset.*

Fedorov, V. V. (1972). *Theory of optimal experiments* (No. 12). New York: Academic Press.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015, April). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Gureckis, T. M., & Markant, D. (2009). Active Learning Strategies in a Spatial Concept Learning Game. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 6.

Gureckis, T. M., & Markant, D. B. (2012, September). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Hitch, G. J. (1978). The role of short-term working memory in mental arithmetic. *Cognitive Psychology*, *10*(3), 302–323.

Imbo, I., Vandierendonck, A., & Vergauwe, E. (2007). The role of working memory in carrying and borrowing. *Psychological Research*, *71*(4), 467–483.

Levy, R. P., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in neural information processing systems* (pp. 937–944).

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

Lindley, D. V. (1956, December). On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005.

Luke, S. G. (2017, August). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502.

Nelson, J. D. (2005). Finding Useful Questions: On Bayesian Diagnosticity, Probability, Impact, and Information Gain. *Psychological review*, 22.

Pashler, H. (1984). Processing stages in overlapping tasks: evidence for a central bottleneck. *Journal of Experimental Psychology: Human perception and performance*, *10*(3), 358.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological science*, *9*(5), 347–356.

Satterthwaite, F. E. (1946, December). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110.

Shannon, C. E. (1948, July). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Zbrodoff, N. J. (1995). Why is 9+ 7 harder than 2+ 3? strength and interference as explanations of the problem-size effect. *Memory & cognition*, *23*(6), 689–700.