The Influence of Statistical Information on Number Estimation Under Uncertainty: Why Do People Present Benford Bias?

Duyi Chi

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (Science) in the School of Psychology, Faculty of Science, the University of Sydney.

September 2024

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Acknowledgements

First and foremost, my deepest appreciation goes to my supervisor, Dr. Bruce Burns, whose guidance and wisdom have been invaluable throughout this process. His expertise, patience, and unwavering support have not only shaped this project but have also significantly contributed to my personal and professional growth.

Additionally, I would like to sincerely thank Dr. Damian Birney, my auxiliary supervisor, for his invaluable input and support throughout this project. I also appreciate Dr. Sabina Kleitman for her time and insightful comments in reviewing the milestone work of my research.

I wish to express my gratitude to my family including my husband, Zhentao Yan, for his assistance in proofreading my work; my parents, for their endless love and care; and my dog Saki, for her spirit to explore and companion.

Finally, my appreciation to everyone I've encountered in my life so far. Their presence and moments shared with me have shaped who I was yesterday, who I am today, and who I will become in the future.

Abstract

People often have to generate numerical answers to questions they are not sure about, such as when police estimate the size of a crowd of protestors, or someone has to estimate the value of an asset. The process of such estimation and how the statistical data people encounter influences it is not well understood, so it is an active area of investigation. Recent research (e.g., Burns & Krygier, 2015; Diekmann, 2007) found a consistent first-digit bias that approximates the logarithm distribution of Benford's law when individuals spontaneously generate unknown numbers, such as the length of a river. They did not obtain a perfect fit of human data to Benford's law, but its pattern accounts for a large amount of variance in human first-digit data. Such findings were extended to other estimation tasks using the visual stimuli of pictured items (Chi & Burns, 2022), suggesting the prevalence and robustness of this phenomenon. Thus, people exhibit a Benford bias in number generation, emphasising a stronger preference towards the smaller leading digits (i.e., 1, 2, 3) and a trend of monotonic decline as the first digit gets larger. Yet, the reasons behind the Benford bias remain unclear. Hence, my project examined how people acquire and utilise statistical information to shape their number estimates by exploring how easily this first-digit distribution can be changed under typical experimental conditions. In four experiments, participants generated numerical estimates in response to factual questions or when assessing large quantities of visual elements, both in response to just exposure and simple feedback. Inconsistent with the optimal statistical principles observed in everyday cognitive judgment (e.g., Griffith & Tenenbaum, 2006), we have not been able to demonstrate that people's estimation followed the underlying distribution of each variable estimated. However, supporting evidence has emerged indicating that the Benford bias can be shifted through experiential learning with feedback, highlighting sensitivity to the first-digit distribution. My investigations with Benford's law seek to enhance the framework of statistical learning and associated problem-

solving. In practice, this also allows us to gain a better understanding of how numbers are generated for use in decision-making.

Table of Contents

List of Tables

List of Figures

The Influence of Statistical Information on Number Estimation Under Uncertainty: Why Do People Present Benford Bias?

It has been over seventy years since the American mathematician Samuel Wilks (1951), paraphrasing Wells's (1903) book *Mankind in the Making*, famously declared that "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." Although his statement might have seemed overstated at the time, the importance it places on statistical knowledge continues to be significant in the current era, replete with information. A recent global application of statistical regularities is evident in the well-documented phenomenon of Benford's law concerning the first digits of numbers.

During the past three pandemic years related to COVID-19, Benford's law has been extensively applied to analyse the reported cases and numbers across various regions, serving as a tool to detect potential anomalies indicative of human manipulation, assuming that naturally occurring datasets typically adhere to this mathematical principle. Over eighty published and unpublished papers have documented the investigation with Benford's law in the context of COVID-19 cases since the start of the global pandemic. This statistical approach has provided insights into the data reporting practices of different regions, flagging irregularities that suggest deviations from expected natural distributions (Negi & Singh, 2023). While the reliability of Benford's law in scrutinising COVID-19 data may be influenced by various factors, such as healthcare policies and systems (Farhadi & Lahooti, 2021; 2022), it has effectively identified certain deviations.

The utility of this statistical pattern lies in the substantial agreement found across a wide range of datasets in both nature and societal activities. Examples include internet traffic data (Arshadi & Jahangir, 2014), distances between the Earth and stars (Alexopoulos & Leontsinis, 2014), brain activity records (Kreuzer et al., 2014), stock turnover (Balamurugan, Harish & Gandhi, 2019), and the size of small ribonucleic acid (sRNA) (Jha, Mendel, Cho &

Choudhary, 2022). More than 700 relevant papers have been published in mathematics and statistics, 600 in finance, auditing and accounting, and 300 in science topics like physics, biology and psychology (estimated by Chi, December 2023). This principle regarding the first digits posits that in any numerical domain covering a wide range of values without confined limits, the leading digits of its data are expected to follow a logarithmic distribution that decreases monotonically (Hassan, 2002). In this distribution, the number 1 appears as the first digit about 30% of the time, while the number 9 appears less frequently, constituting no more than 5% of occurrences. Therefore, it has been proposed that any significant departure from Benford's distribution ought to be reviewed with caution, as it could indicate human intervention. This assumption is based on the premise that data created by humans are unlikely to exhibit the same regularities that are commonly found in naturally occurring datasets.

However, recent studies (e.g., Burns & Krygier, 2015; Chi & Burns, 2022) in Psychology demonstrated that people could spontaneously present a bias towards the smaller first digits, which approximates Benford's law when generating the unknown numbers, such as estimating the area of a region, the cost of electricity consumption, or the quantities of jellybeans. This phenomenon has been replicated in several tests asking for estimates in trivia or factual questions and even extended to visual estimation tasks. Although the data did not show a perfect fit, for example, the leading digit 5 sometimes emerged as a secondary peak, Benford's proportions accounted for a significant proportion of variance in human first-digit data. Therefore, it seems that people exhibit a Benford bias.

This psychological phenomenon, where people also follow the logarithmic distribution of first digits found in naturally occurring settings, challenges the reliability of Benford's law as a lie detector and leads to investigations into why people exhibit Benford bias. To effectively employ Benford's law in fraud detection, it is essential to comprehend

the ways in which individuals generate numbers. Additionally, we must explore why people exhibit a bias that also agrees with this pattern. A potential explanation lies in the pervasive evidence of the unconscious acquisition of statistical information through mere exposure (Aslin, 2017). If people adhere to Benford's law due to long-term exposure in their lifetimes, given its ubiquity in natural data, then it is plausible that this phenomenon could extend to other tasks, such as selecting a number. However, this bias has been exclusively reported when generating meaningful numbers for decision-making. Since people failed to show strong preferences for smaller leading digits when choosing numerical responses as answers, it is argued that the Benford bias might be a product of the cognitive process in number estimation (Burns, 2009).

Nevertheless, the appeal of statistical learning remains strong, given the various documented evidence supported by laboratory observations, such as the behaviour of probability matching. Therefore, building on compelling data about people's sensitivity to statistical information, the present project aims to systematically examine the plausible explanations for Benford bias by exploring how people acquire and utilise statistical regularities in the process of number estimation. Benford bias might represent a regularity in how people generate unknown numbers, thus the relevant investigation could add insights into the framework of statistical learning. Practically, this could illuminate the understanding of how people produce estimates for decision-making and problem-solving under uncertain situations.

CHAPTER 1: INTRODUCING BENFORD BIAS

The inspiration for the current project stems from a well-established phenomenon related to the distribution of first digits in datasets found in natural environments, known as Benford's law. This pattern is widely used today for verifying the authenticity of records across various disciplines, including accounting, economics, and politics. However, its relevance to psychology has recently come to light as studies have shown that humans may also exhibit similar patterns. To lay the foundation for the current study's link to Benford's law, this chapter introduces its historical relevance and significance, along with its examination in Psychology.

Background of Benford's Law

The origins of Benford's law can be traced back to an intriguing observation made by the American astronomer Simon Newcomb (1881). While perusing logarithm tables, he noticed that the pages starting with the digit-1 showed significantly more wear than others. This observation hinted at a frequency pattern where numbers beginning with lower digits appeared more often. Newcomb's brief publication inferred that the occurrence of the leading digit 1 was almost one-third of the time and decreased monotonically until reaching digit-9. Physicist Frank Benford independently reported this phenomenon in 1938 through his extensive observations. Benford collected over 20,000 data points from 20 domains, such as death rates, river areas, and mathematical laws. He found that unrelated datasets from various natural and social resources demonstrated a close fit of their first digits to a logarithmic curve (see Figure 1).

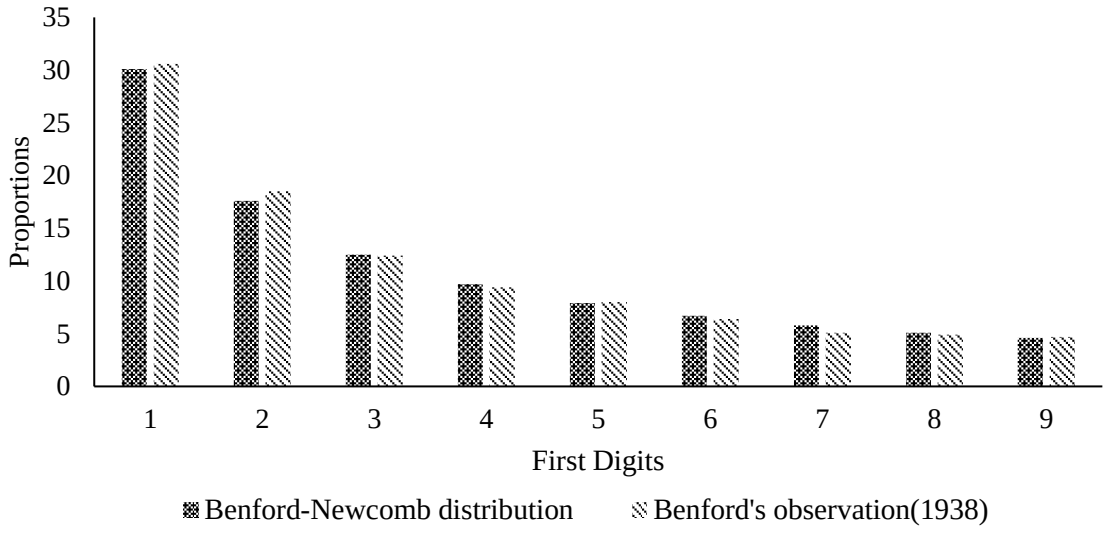


Figure 1. Benford formulated the mathematical expression of the frequency distribution of leading digits and later refined by Hill (1995a): $P(d) = \log_b ((d + 1)/d)$, for $b \geq 2$, $d \in \{1, \ldots, b - 1\}$ where P(d) represents the probability of leading digit "d" in base "b." These theoretical proportions of BL are also commonly referred to as the Benford-Newcomb distribution (BND). Benford's observation showed the percentage of times the natural numbers 1 to 9 were used as first digits in numbers accumulated from 20,229 observations by Frank Benford in 1938.

Initially referred to as "the law of anomalous numbers" by Benford, it later gained widespread acceptance as Benford's law (BL) or the first-digit phenomenon due to its robust skewness compared to other digit distributions. Both Newcomb and Benford noted that differences in the frequencies of the second digits were less pronounced and diminished beyond the third digit place. In essence, the significant differences from a flat distribution are primarily observed within the first significant digits (FSDs), and the refined logarithmic distribution of the first digits was commonly referred to as the Benford-Newcomb distribution (BND). Figure 1 shows that the combined proportions of lower leading digits "1," "2," and "3" account for over 60%, indicating a strong bias toward smaller FSDs. Furthermore, Hill's (1995a, 1995b) mathematical demonstrated that this logarithmic distribution remains invariant to the scale or base of measurement. For example, the law still yields a reasonable approximation when converting miles to kilometres or transferring

between different currencies. These base-invariant and scale-invariant attributes support the generalisation of Benford's law in natural phenomena, making it independent of the units in which numbers are measured.

Benford's law, which claims to capture regularities in natural data, has sparked an increasing number of empirical studies across various disciplines aiming to expand the observation of its logarithmic distribution. This includes mathematical and statistical regularities, finance and accounting datasets, scientific investigation results, and socio-economic and clinical environment observations. Recent literature has accumulated evidence of BL fitting into contemporary topics. Examples include sports events (Park, Choi & Cho, 2016), Facebook features (Striga & Podobnik, 2018), music notes frequencies (Nelson, Wickman, Null & Gazin, 2021), COVID-19 data reported by most countries (Farhadi & Lahooti, 2022), human speech spectra (Hsu & Berisha, 2022), and online auction (Chang & Chang, 2023). These findings reinforce the universality of BL in natural data. However, Durtschi, Hillson, and Pacini (2004) highlighted circumstances that might inhibit the manifestation of BL: 1) when numbers in datasets are randomly determined (e.g., mobile numbers), 2) when they are tailored for a specific purpose (e.g., priced at $9.99), 3) when confined to a narrow range (e.g., satisfactory ratings), or 4) datasets comprise a large volume of pre-set values like $100 gift cards. Nevertheless, BL is a robust phenomenon describing the regularities of leading digits observed in unrelated datasets from naturally occurring settings as long as they span multiple magnitudes without limited boundaries. The underlying message emphasises a stronger bias towards smaller first digits (i.e., 1, 2, 3) and presents a trend of monotonic decline as the first digit increases.

The prevalence of the first-digit phenomenon in nature makes it a valuable resource for use in developing a fraud detection tool. This idea was conceived and refined by Nigrini (1992), a research accountant, during his PhD project, where he explored the question: "What

if the observed first digits from naturally occurring data deviate from the expected frequencies?" Nigrini showed that Benford's law could serve as a test (utilising the first two digits) to screen for financial fraud. This was after examining various datasets deemed genuine, which included taxpayer data, census data, and accounts payable data. Following Nigrini's (2015) work in forensic accounting, the practical application of BL rapidly evolved to identify the authenticity of datasets by contrasting samples from fraudulent and legitimate records. BL is extensively integrated into audit procedures and accounting software as a deception detector nowadays (Miller, 2015). A more recent application involves detecting falsified research records, data exfiltration, and IT sabotage. Given the substantial societal implications of erroneous conclusions from flawed data, these tests offer a critical instrument to validate claims.

Employing Benford's law as a deception detector postulates that deviations from its distribution might signal human interference. Yet, this method assumes that human-generated data typically diverges from Benford's distribution (Bolton & Hand, 2002). In Nelson's check fraud case, fabricated amounts often began with higher primary significant digits, indicating that the first-digit phenomenon contradicted human intuition. Nonetheless, for a long time, it remained an unanswered empirical question whether people's behaviour could follow Benford's distribution, given its pervasive occurrence. Recently, psychologists have observed that, under certain circumstances, people may also inherently exhibit a bias towards smaller leading digits in line with BL. Thus, I will revisit the past and recent psychological investigations concerning the first-digit patterns in the following sections.

Psychological Phenomenon of Benford Bias

Early research in Psychology failed to show conformity to Benford's law, even though the studies on number generation suggested that people do not produce numbers in

the same manner as dice. Investigations into how humans produce random numbers revealed that their results did not display a monotonic decline consistent with Benford's distribution. Yet, promising findings have emerged in recent studies that employed alternative methods (e.g., Diekmann, 2007; Burns, 2009). Mirroring fundamental concepts of BL, a bias towards the smaller leading digits of a number has been consistently reported when producing meaningful numbers in response to decision-making queries, such as the electricity consumption, the household cost, and the quantities of jellybeans.

Early Explorations

One early experiment by Hsü (1948) asked individuals to generate a four-digit number freely. The leading digit-4 appeared most frequently (15.5%) produced, followed by digit-1 (13.3%), with 5 being the least produced first digit. A replication by Hill (1988) with six-digit number generation found 6 to be the most frequent leading digit, with 1 and 9 being less favoured. These findings not only failed to confirm Benford's distribution but also indicated an effect of priming, where requesting a particular number length influenced its likelihood of being the leading digit. Kubovy and Psotka (1976) identified an inverse relationship, where mentioning the digit-7 as an example decreased its occurrence. Afterwards, Kubovy (1977) contrasted various priming stimuli. Asking for one-digit numbers resulted in 1 as the most common response (18.0%), while requesting a number between 0 and 9 led to a prevalence of 7 (28.4%). Using digit-1 as an example reduced its occurrence (5.4%). Scott, Barnard, and May (2001) additionally examined the impact of the central executive function on number production. The degree of elaboration in numbers generated affected the leading digits. Unelaborated responses (e.g., 3,000,000) showed a preference for 5, while elaborated responses (e.g., 7,000,378) peaked at 1. Nonetheless, none of these leading digit patterns aligned with Benford's distribution.

It is important to note the limitations of these study designs. Participants were tasked with producing only one random number per test, making it an observation only through aggregated outcomes. Moreover, results were largely influenced by priming effects. In everyday life, individuals often spontaneously generate numbers for decision-making, a process differing from creating random digits. Hence, patterns observed from randomly generated numbers might not necessarily reflect everyday situations, raising concerns about their broader applicability.

Establishment of Benford Bias

Initial investigations into the pattern of leading digits generated by humans encountered challenges, predominantly due to negative results from random number generation studies. However, promising findings emerged when Diekmann (2007) and Burns and Krygier (2015) introduced alternative methods, yielding supporting evidence.

Diekmann's research focused on testing the applicability of Benford's law as a lie detector by comparing reported regression coefficients with fabricated ones. Rather than examining established fraud cases, Diekmann created data under experimental conditions. His initial examinations on unstandardised regression coefficients from the *American Journal of Sociology* found no notable discrepancy from Benford's distribution. Although a secondary collection displayed a slight deviation, it was mainly due to an uptick at digit-5. Overall, the leading digits of regression coefficients reasonably approximated BL. Diekmann then proceeded to scrutinise spontaneous data generation, asking students to respond to ten questions with four-digit "plausible values." The aggregated frequencies of the first significant digits in the fabricated data closely mirrored Benford's distribution, albeit with deviations in the secondary digits. Diekmann suggested that investigating later digits might provide more insights for fraud detection. Notably, the priming effect evident in prior studies was absent in this research, as digit-4 was not elevated when creating a four-digit number.

Nevertheless, the attempt to form regression coefficients revealed a potential human bias towards smaller leading digits.

In a separate study, Burns and Krygier (2015) identified a similar bias by prompting participants to produce numbers within a meaningful context. They posited that the conformity to Benford's law observed in Diekmann's study resulted from the context of providing meaningful values for decision-making. Nine questions were selected from real-world scenarios, such as national debt, regional population, and infant mortality rates, while avoiding widely recognised answers. They ensured that each digit from 1 to 9 equally represented the leading digit of nine correct responses. The first-digit pattern of these responses closely followed the frequencies of BL, except for an elevation at digit-5. Further separation of elaborated data, wherein a number contains multiple non-zero digits (e.g., 5,000,776), reinforced the agreement with BL due to a suppressed elevation of digit-5. To address the potential limitation that everyone received the same set of nine questions, they designed a follow-up experiment with an item pool of 81 questions by including nine different estimated targets (e.g., Nile River) crossed with nine meaningful domains (e.g., the length of a river). This enabled an adequate randomisation of the test items while maintaining an even distribution of the first digits in the correct answers. The first-digit pattern of responses aggregated in this second attempt also replicated a reasonable monotonic decline, showing a stronger preference over the smaller leading digits with a spike at digit-5.

Additional support came from unpublished research at the University of Sydney by Krygier (2009), which directly compared the impact of context meaningfulness on data generation. It was found that numerical answers from meaningful domains (e.g., drainage area, autonomic weight) more closely adhered to Benford's frequencies than those from meaningless domains where the numbers do not convey quantitative meanings (e.g., random numbers, postcodes) for decision-making. These findings were replicated in subsequent

studies that posed trivial questions in an unpublished project (Tripodi, 2016) and even extended to dimensions primarily answered based on daily experiences and activities, such as time spent on social networking or the weight of a salad bowl (Chi, 2020). Data from meaningful contexts consistently exhibited a close, though not perfect, fit to BL, with a secondary peak at digit-5 in the distribution of leading digits.

Compelling evidence emerged not only from numerical answers to general knowledge or experience questions but also from estimated numbers in response to tasks involving visual stimuli. Chi (2020) presented a set of nine pictured items showing hundreds of jellybeans in a jar or dots in a display area for participants to estimate the quantities. The leading digit distribution of these estimates more closely resembled Benford's proportions than the actual answers, where each leading digit held an equal probability. Spikes at digit-5 were also detected in some conditions. Moreover, this monotonic decline persisted across various information levels presented in the images (i.e., blurry vs. static vs. dynamic).

These studies have offered valuable insights into the relationship between human-generated data and Benford's law. The collective findings emphasised that people's behaviour could approximate Benford's distribution observed in natural settings. The data reported by these studies are selectively summarised in Table 1. A stronger bias towards smaller first digits was consistently reported when generating quantitative answers, and the pattern highlighted a trend of monotonic decline as the leading digit became larger. While the results did not achieve a perfect fit (e.g., elevation at digit-5), they generally concurred with the underlying notion of BL. Burns (2009) speculated that the overproduction of digit-5 might be a common choice when individuals are uncertain about the correct answer, as it represents a midpoint in magnitude. The frequent use of digit-5 as the leading digit has also been noted in a recent analysis of off-balance sheet records from 22 commercial banks, showing manipulation such as excess rounding and deliberate reduction of percentages

(Parnes, 2022). This might indicate a unique human response when coping with uncertainty despite a general approximation to Benford's proportions. Therefore, to both link and distinguish it from BL evident in natural contexts, it was proposed to portray this psychological observation as "Benford bias."

Table 1.

Percentage frequency of each first digit generated by the participants in the studies showing Benford bias (selectively included).

First Digits	1	2	3	4	5	6	7	8	9
Benford's proportions	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Diekmann (2007): Four-digit regression coefficients fabricated in Exp 2 (n=13×10)	26.2	19.2	10.8	5.4	12.3	5.4	5.4	10.8	4.7
Burns and Krygier (2015): Numerical responses in Number Generation, Study 1 (n=127×9)	26	14	10	7	18	9	4.0	6	5
Burns and Krygier (2015): Numerical responses in Number Generation, Study 2 (n=335×9)	23.9	15.1	9.7	7.6	19.0	9.1	5.5	5.8	4.4
Chi (2020): Numerical responses to experience questions, Study 3 (n=76×9)	26.4	17.1	15.7	9.5	13.5	6.4	5.1	1.8	4.6
Chi (2020): Estimates generated to the quantities of dots in the pictures, Study 3 (n=210×9)	26.0	21.4	15.7	9.00	9.5	4.7	4.8	5.6	3.4
Chi & Burns (2022). Estimates generated to the quantities of jellybeans in blurry pictures (n=105×9)	26.8	15.8	12.9	10.8	10.7	6.7	7.2	6.8	2.4

Exploring Explanations of Benford Bias

Consistent with Benford's law found in naturally occurring settings, evidence has also emerged in Psychology, suggesting that people could display strong preferences towards the smaller leading digits. This challenged the validity of using BL as a tool in fraud detection since numbers produced by humans also inherently follow such regularities. However, the reasons behind the observation of Benford bias in human behaviour remained unclear. Several studies have sought to provide explanations based on theories of learning, speculating that individuals might have been exposed to this statistical pattern during their lifetime, given its ubiquitous presence in various datasets.

Evidence of Statistical Learning

Many researchers (e.g., Bargh & Ferguson, 2000; Gigerenzer & Todd, 1999) have emphasised how decision-making can be influenced by knowledge of regularities in the data they encounter in the environment. From infancy to adulthood, the ability to extract statistical information from the surrounding environment has been consistently documented (e.g., Saffran, Johnson, Aslin & Newport, 1999). Early experiments primarily focused on sensitivity to distributional information, such as analysing continuous streams of speech syllables and testing for associative patterns, like word segmentation. However, this ability to utilise statistical cues extends beyond language acquisition and encompasses both auditory (e.g., Saffran, Johnson, Aslin & Newport, 1999) and visual stimuli (e.g., Kirkham, Slemmer & Johnson, 2002; Turk-Browne, Jungé & Scholl, 2005). Moreover, Fiser and Aslin (2002) demonstrated that individuals could even acquire higher-order statistical regularities embedded in spatial positions by experiencing the frequencies of multiple pairs of visual shapes. Collective evidence points to people being sensitive to statistical information in various domains, capable of producing responses that mirror such regularities.

A notable characteristic of these experimental setups is that the passive learning experience is offered in the absence of explicit instruction or feedback, commonly referred to as “mere exposure” (Saffran, Johnson, Aslin & Newport, 1999). This rapid acquisition process typically falls under the construct of implicit learning, where learning occurs without apparent awareness, sharing similar mechanisms of unconscious knowledge (Conway & Christiansen, 2005). Participants in these paradigms of statistical learning often lack explicit and conscious knowledge of the underlying pattern of associations (Turk-Browne et al., 2005). Evidence suggests that the unconscious acquisition of statistical information is substantiated by the difficulty in recalling or recognising sequences despite evident learning (e.g., Willingham & Goedert-Eschmann, 1999) and the lack of correlation between confidence and correctness in answers (e.g., Scott & Dienes, 2008).

Two types of subjective measures are commonly employed to assess the presence of implicit or explicit knowledge. Implicit knowledge can be inferred by examining the “guessing criterion” to determine if participants claim performance above chance levels to mere guessing. On the other hand, the “zero-correlation criterion” examines whether there is a correlation between accuracy and confidence. If consistently lower confidence is reported regardless of answer correctness or if accuracy is unrelated to confidence, it typically indicates implicit knowledge. Recent studies have further illuminated the nature of statistical learning by contrasting performance in direct recognition tests with indirect tests, as measured by novel reaction times (Batterink, Reber, Neville & Paller, 2015). These studies suggest that statistical learning may also involve the optional formation of explicit knowledge, which operates alongside the unconscious process (e.g., Servan-Schreiber & Anderson, 1990; Sanchez & Reber, 2013). While explicit knowledge of the underlying structure was not considered essential for facilitating performance, it might still be developed

during statistical learning, capturing a broader scope of knowledge than traditionally believed (Batterink, Reber, Neville & Paller, 2015).

The ability to acquire statistical information is a prevalent and robust phenomenon in the field of learning. However, its effectiveness can be limited or influenced by various factors. Attention can effectively guide adults through explicit instruction, while infants typically receive indirect instruction. They seem to possess an innate sense of when attention is likely to yield further information, depending on the complexity or simplicity of the available information (Kidd, Piantadosi & Aslin, 2012). Additionally, statistical learning is time-sensitive; learning nonadjacent statistics is more challenging for adults than adjacent statistics unless a common feature binds them together (Aslin, 2017). Therefore, certain associations are formed more easily than others, depending on perceptual cues. Meanwhile, factors such as primacy and familiarity significantly influence the constraints of statistical learning. Once an information structure is established, it often resists change, only altering when a new pattern persistently repeats itself (Gebhart, Aslin & Newport, 2009; Weiss, Gerfen & Mitchel, 2009). Learning unfamiliar new patterns demands additional effort and time, such as lowering the presentation rate, compared to familiar ones (Gebhart, Newport & Aslin, 2009).

Despite these limitations, the power of statistical learning is evident in its ability to generalise to novel situations. A key concern regarding the utility of information acquisition lies in the capability of transfer of knowledge; a crucial question revolves around the computational mechanisms that enable structure extraction and generalisation to new examples. Given that individuals might have restricted exposure to stimuli in daily life, transferring fundamental features of patterns to abstract guidelines is essential for problem-solving and influencing future actions (Aslin, 2017). Encouraging results have been observed across various studies and domains. For instance, adults exposed to sentences in an artificial

language were able to judge novel sentences that shared the same grammatical rules rather than relying solely on the familiarity of nonsense words (Reeder, Newport & Aslin, 2013). There have also been advanced instances of uncovering hidden structures within complex stimuli in visual perception. Here, individuals could form precise and economical visual-spatial information representations, reflecting the optimal predictions of Bayesian learners (Orban, Fiser, Aslin & Lengyel, 2008). Such results confirmed and elucidated the transition learners make from grasping specific examples to understanding abstract computational principles.

Recognition Hypothesis

Given the extensive support for theories of statistical learning and its solid foundation in laboratory practice, if a higher-order statistical relationship, such as conditional probability, can be extracted through just 5 minutes exposure (e.g., Fiser & Aslin, 2002), then it is not surprising that people may act in accordance with a law that holds in various situations where they are likely to encounter it. This potential for unconscious acquisition of the regularities in the environment leads to the proposal of the *Recognition Hypothesis*, which assumes that if individuals are constantly surrounded by data consistent with BL, they may become sensitive to the statistical relationship of the first digit. If the bias towards smaller first digits can be attributed to implicit learning through exposure, then such a preference should not be limited to tasks involving number generation. It should also be present when participants are asked to *recognise* correct answers, such as when they are given numerical answers and asked to *select* the correct one.

To test the Recognition Hypothesis, Burns and Krygier (2015) introduced a selection task in one of their studies. This task presented numerical questions similar to those used in their generation tasks (e.g., national debt, region population, etc.). However, instead of generating a number as a response, participants were asked to choose an answer from nine

numerical options, each with a different first digit. The questions were randomly selected from an 81-item pool (nine domains by nine targets), and the leading digits of the correct responses to the nine questions were equally distributed. If people have Benford bias due to long-term exposure in their environment, the options with lower first digits should be selected more frequently than those with higher first digits. Contrary to this prediction, except for a slight elevation for digit-1, the relative frequencies of the first digits chosen by participants were close to a flat distribution. People appeared to have no strong preference for any first digits when selecting answers. However, it was argued that providing too many choices might demotivate individuals to make a rational judgment (e.g., Iyengar & Lepper, 2000) by exhibiting undesired behaviour such as selecting the default option (Shafir & Tversky, 1992), relying on heuristics (Timmermans, 1993), and processing partial information (Hauser & Wernerfelt, 1990).

To avoid distorting responses by overwhelming participants with too many options, Tripodi (2016) modified the task in his honours project. Participants were presented with a single numerical value for each question and asked to indicate whether they believed it was correct or not. For example, “In 2013, was the Protestant population of Italy 755,328? Choose ‘yes’ or ‘no’.” Participants answered 18 questions from non-arbitrary domains (e.g., areas of countries) and 18 from arbitrary domains (e.g., specific phone numbers). Again, neither non-arbitrary nor arbitrary domains produced evidence of any first-digit preference.

Chi and Burns (2022) undertook a third attempt. They argued that previous selection tasks offered either nine options or a single potential answer; thus, neither directly contrasted the first digits. So before ruling out the utility of the Recognition Hypothesis, they tested it with what was considered to be the most sensitive paradigm for a selection task by directly pitting a smaller and a larger FSD of a number against each other, for example, a forced-choice item was offered between 2xx and 8xx where x is a random digit. Suppose people

implicitly learned that the lower first digits occurred more frequently than the higher ones due to exposure to this pattern in the environment. In that case, selected responses should have strongly favoured the smaller first digits of a number. Unfortunately, similar to the previous studies by Burns and Krygier (2015) and Tripodi (2016), the data did not show a systematic difference in preferences for any first digit. There was no evidence of Benford bias in the selected responses.

Summary of Chapter 1

So far, the evidence from three attempts has not supported the reliability of the recognition hypothesis in explaining Benford bias. They all emphasised the distinctions in the patterns when it comes to selecting numerical responses compared to generating them, particularly with regard to their first-digit distributions. The result that the chosen numbers did not exhibit a bias toward smaller leading digits was surprising, given the existing findings that point to people's unconscious acquisition of statistical relationships within controlled laboratory conditions (e.g., Fiser & Aslin, 2002) and the widespread prevalence of BL. This suggested the need for a more comprehensive grasp of the constraints associated with implicit statistical learning to fully understand the mechanisms of automatic relationship acquisition (Fiser & Aslin, 2002).

Furthermore, these findings have highlighted a divergence between the processes of number generation and number selection. If the preferences towards the smaller leading digits can only be detected through the act of producing numerical answers, then the exploration of alternative explanations may be best directed towards the potential mechanisms linked to number generation. As speculated by Burns (2009), Benford bias could be a consequence of cognitive processes when quantities are produced. In that regard, comprehending this

phenomenon should illuminate the cognitive procedures that underlie the spontaneous formulation of numerical responses.

However, investigations into how individuals generate quantities with limited information, particularly within the theoretical framework of how estimates are generated for decision-making and judgment, have not been systematically explored. Thus, the following chapter is poised to present a comprehensive review of empirical research conducted within Psychology that delves into the realm of number estimation under uncertainty. By acquiring a more profound understanding of the general process through which individuals generate estimates for unknown questions, we stand to gain valuable insights into our quest for the alternative explanation for Benford bias.

CHAPTER 2: A REVIEW OF LITERATURE IN NUMBER ESTIMATION

The exploration of the first-digit pattern in Psychology has established that people tend to show a stronger preference for smaller leading digits, which mirrors the regularities of Benford's law present in naturally occurring settings. However, such bias has been consistently emerged only in studies that prompted participants to spontaneously generate numerical responses to meaningful but unknown questions, such as national debt, electricity consumption, and the quantities of pictured items (e.g., Chi & Burns, 2022). These findings suggested that Benford bias is exclusively linked to the estimation process in number generation.

To further investigate alternative explanations for the Benford bias, it was essential to comprehensively understand the relevant body of work related to how numbers are predicted, estimated, or produced under uncertainty. Tests assessing individuals' ability to generate reasonable estimates, as a component of cognitive test batteries, span various research disciplines and practical settings, such as those for patients with mental health illness (e.g., Barabassy, Beinhoff & Riepe, 2000). Despite the extensive use of estimation tasks, the core mechanisms guiding people in making reasonable estimates remain unexplored. This chapter is dedicated to a systematic review and presentation of the current insights into number estimation. As the exploration delved deeper, it was inevitably drawn into a debate juxtaposing the idea of whether that individual behaviour is driven by intuitional judgment like heuristic and mental shortcuts (e.g., Tversky & Kahneman, 1974) or if they align with statistical regularities (e.g., Gigerenzer & Hoffrage, 1999). While resolving this debate was not the primary aim of the current project, nor was a resolution expected from a mere review, the goal is to shed light on the quest to understand the Benford bias in number generation.

This chapter aims to systematically present the existing knowledge about the cognitive mechanisms involved in number estimation, specifically focusing on searching for plausible explanations for Benford bias. The extensive exploration has examined the diverse ways in which people make estimates. Although certain topics explored might appear irrelevant to the primary focus of the current project, their investigation still holds merit and value. This body of knowledge has informed the design of the current research, aided in the interpretation of the results, and laid the groundwork for future inquiries. It provides context and depth to the present study, even if not all aspects directly clarify the nature of Benford bias.

The Mental Number Line

The present research began with an intensive search for keywords and tests in empirical literature related to number estimation. It was noted that a subset of studies used a task called number line estimation (NLE), where participants estimate a target number (e.g., 988) and mark its position on a horizontal line spanning a wide range (e.g., 0 to 100,000). The concept of NLE traces back to the mental number line theory (Dehaene, 1997), which posits that numbers are mentally represented on a logarithmic scale, causing the perceived distance between numbers to compress exponentially as they increase. For example, the perceived difference between 80 and 95 is generally more pronounced than between 180 and 195 despite the actual numerical distance being the same. This logarithmic perception suggests larger numbers are packed more closely, making them less accessible (Dehaene, 2003). Dehaene and Mehler (1992) additionally used this logarithmic representation of numbers to explain why smaller numbers like 10, 20, and 50 are more frequent in language across cultures, which reflects how the brain processes and represents numbers. However, recent studies indicate that this logarithmic representation, especially for larger numbers,

depends on factors like age and number range (Siegler & Opfer, 2003). A progressive shift from logarithmic to linear representation was observed in older children's groups within large number ranges, such as the median estimates made in the NLE by sixth-grade students for numbers up to a hundred thousand (Thompson & Opfer, 2010). Moreover, the logarithmic model does not strictly apply in cases involving extremely unfamiliar and large numbers influenced by surface structures like the use of number words (e.g., "million," "trillion") in undergraduate samples (Landy, Silbert & Goldin, 2013).

Dehaene's (1997) work suggested a general preference for smaller digits, somewhat echoing Benford's law, where digit-1 is more common than digit-9. However, Benford bias focuses on the frequency of leftmost digits in numbers, irrespective of their magnitude. Dehaene's experiments also explored whether people were more attentive to leading digits, hypothesising quicker responses in numbers with different first digits. Contrary to his expectations, his study revealed that judging that 71 is larger than 65 was not significantly faster than comparing numbers with the same leading digit, like 69 and 65. As a result, he discounted the explanatory power of BL, attributing the phenomenon instead to "the grammatical structure of our numerical notations" (p.99).

Tests demonstrating the compressive nature of numerical representation primarily assessed the speed and accuracy of recognition and comparison tasks. However, these tests of judging relative sizes should not be equivalent to number generation behaviour. Although both the mental number line and Benford bias involve a logarithmic scale, it is argued that spontaneously generating numbers differs from selection tasks.

Dehaene and his colleagues' work is instrumental in understanding the fundamental mechanisms of cognitive development concerning number acquisition and representation. Developing a robust and appropriate number sense is widely believed to be crucial for handling numerical tasks. For example, children excelling in NLE tasks were often more

competent in numerical tasks like remembering numbers (Thompson & Siegler, 2010), counting (Hamdan & Gunderson, 2017) and standard math tests (Booth & Siegler, 2008). However, it is essential to differentiate between providing a horizontal line to identify a number's position and spontaneously generating digits for an estimate. Placing a number using a slider (e.g., pinpoint the location of 778) should be considered differently from the process of free number estimation (e.g., $778 is an estimated cost for the monthly household expenses). This differentiation parallels the well-established debate concerning recall and recognition in memory tasks, which was explained through different theories with respect to the number of cognitive stages (Anderson & Bower, 1972) or cue information (Tulving, 1976). Regardless of the justifications provided, it was widely accepted that the process involved in recalling and recognising differ. The research on the mental number line appears to offer limited insights into how people generate estimates freely for unknown questions where the Benford bias emerged from. Nevertheless, the studies supporting the framework of the mental number line highlight an intriguing observation that people might be prone to make biased judgments in numerical representation.

The Left Digit Effect

While Dehaene's work (1997) refuted the speculation that the first digit is unique in numerical judgment, as evidenced by tests on recognition speed, a different observation has emerged. Referred to as the "left digit effect," this phenomenon was recently established under the testing paradigm of number line estimation tasks. This effect is noteworthy not only for its similarity in name to the "first digit phenomenon" explored in this research but also because it reveals a potential bias concerning the leftmost digit.

The left digit effect was initially documented by Lai, Zax and Barth (2018), showing that the distance perceived between two three-digit numbers with different leftmost digits

(e.g., 398 vs. 403) was considerably greater than the distance between pairs with the same leftmost digit (e.g., 398 vs. 393) in the placement tasks. Similar effects were also detected with two-digit numbers (Williams, Zax, Patalano & Barth, 2022). According to the theory of mental number line, close larger numbers are expected to be placed in similar locations due to compressed spaces. However, those findings addressed an additional influence of digit-level information, potentially biasing how numbers are perceived in terms of magnitude. This emphasis on the first-digit perception was also observed across many other domains of research; for example, judgment about price is sensitive to changes in the first digit (e.g., an increment from $4.99 to $5.01 is judged more costly compared to the same increment from $4.97 to $4.99) (Thomas & Morwitz, 2005). This phenomenon appears to be quite robust, as even precise feedback failed to eliminate its effects (Kayton et al., 2022). Recent research, however, has uncovered its interaction with linguistic structures. The left digit effect could be mitigated among individuals using inverted number words, as seen in Dutch, where “29” is articulated as “nine and twenty” in verbal communication (Savelkouls, Williams & Barth, 2020). This insight implied that language properties could exert influence on multi-digit number processing.

The sensitivity to the leftmost digit is crucial in daily activities because it provides the first piece of numerical information. The explorations with the left digit effect show some similarities with the idea from Benford bias that people find it more important to talk about the leftmost digit. It supports the idea that the first digit is special, which appears consistent with the notion of BL about the phenomenon observed at the first-digit place. However, this is still unable to provide a reliable explanation to justify the stronger preferences over the lower leading digits when generating unknown numbers.

Number Estimation Using Heuristics and Mental Shortcuts

Previous studies (e.g., Burns & Krygier, 2015; Chi & Burns, 2022) consistently demonstrated that the numerical options selected in recognition tasks did not exhibit the Benford bias, highlighting that the preference for smaller leading digits is likely a result of the number generation process. Thus, we turned our attention to empirical work centred on the spontaneous generation of estimates as answers under uncertainty. The initial collection that caught our attention consistently referenced a well-established framework regarding the use of heuristics in decision-making and judgments. The exploration of heuristics, seen as rules of thumb or mental shortcuts used due to cognitive ease, has been ongoing for over fifty years since Tversky and Kahneman introduced such concepts in 1974. This body of work encompasses several tasks requiring a numerical response estimate, particularly for certain mental shortcuts, such as the representative and availability heuristics, anchoring effect, and the use of mental reference values like prominent or round numbers. These shortcuts may be utilised in very similar types of tasks as those in studies where the Benford bias emerged, such as producing estimates to the factual knowledge questions (e.g., Burns, 2009).

Representative and Availability Bias

One of the classic test items presented in previous studies to illustrate Benford bias involved estimating frequencies, such as the infant mortality rate (Burns, 2009). Decision-making often requires judgments of the occurrences of events. Kahneman and Tversky (1972) demonstrated that, unlike rational judgment following Bayes' theorem, people's estimates regarding frequencies or likelihood largely deviated from the norms of probabilistic reasoning. One typical behaviour they observed was that people tend to overweight the individuating information even when base rates were provided. For example, when participants were asked to determine the probability of a person's occupation (e.g., engineer), they typically relied more on attributes like family status, personality, and hobbies rather than

factual statistics (e.g., 30% of the group are engineers). This phenomenon, known as "base-rate neglect," was further interpreted as a shortcut using representativeness, where judgments favour dispositional attributions of a category rather than the consensus data when encountering uncertain situations.

Another closely related heuristic often used for estimating the frequencies is based on the availability of an event. For instance, when judging relative death rates for different causes, deaths due to murder were significantly overrated compared to deaths due to suicide, even when the opposite was true (Lichtenstein, Slovic, Fischhoff, Layman & Combs, 1978). Several explanations were proposed to illustrate how the availability bias works. One theory argued that frequencies were determined by the number of incidents that can be recalled from memory (Kahneman & Tversky, 1974). In this context, incidents involving murder might be reported more frequently to the public through the media, leading to overestimation. Conversely, Schwarz et al. (1991) rebutted it by showing that recalling twelve incidents failed to result in higher ratings of the associated behaviour compared to six items. Instead, they argued that the judgment of probabilities was affected by the extent to which the example can be easily retrieved. Perhaps incidents associated with murder were more pronounced and easily recalled due to the influence of public media, making them more accessible. Both explanations provide insights into the mechanisms of probability estimation that are susceptible to error.

However, tasks that typically involve representativeness and availability biases often do not explicitly require producing a numerical estimate as an output. Consequently, such heuristic approaches might not fully account for the phenomena observed in the process of generating numbers. Furthermore, Benford bias has been identified through the aggregation of estimates across various items, extending beyond mere frequency estimation. While individuals might resort to using analogies or familiar examples as a strategy to answer trivia

questions, the direct connection between these approaches and the collective tendency to favour smaller first digits, particularly 1, 2, and 3, in number generation remains unclear. This gap in understanding highlights a potential area for further exploration of how specific heuristics translate into observed patterns in numerical estimation.

Anchoring Effect

Decision-making not only involves generating estimates associated with probabilities but also requires estimating values in response to other general knowledge or factual questions, such as the external debt and population in a region (e.g., Burns & Krygier, 2015; Chi & Burns, 2022). These sets of questions, serving as key components in the test materials, consistently demonstrated the emergence of Benford bias. Additionally, it was observed that such items were frequently utilised in numerous studies investigating a pervasive heuristic that influences the estimation process: the anchoring effect. In this phenomenon, the first piece of information encountered serves as an anchor or reference point and influences the subsequent estimates of a value, even if they are not related (Kahneman & Tversky, 1974). For instance, recalling the last two digits of a social security number could impact the estimated cost for an item with an unknown value (Ariely et al., 2003). The group with higher social security numbers placed higher bids than those with lower digits. The effect has been observed across various domains, including legal judgment (e.g., Enough & Mussweiler, 2001), purchasing decisions (e.g., Mussweiler, Strack & Pfeiffer, 2000), predicting and negotiation (e.g., Galinsky & Mussweiler, 2001).

The anchoring effect appears to be a very robust phenomenon in human judgment. Various types of anchors have been tested to see if they could make a difference, such as references from external providers (e.g., the information given by experimenters), uninformative and irrelevant information (e.g., digits from a random number generator), and non-numerical concept (e.g., the wording for the description of the speed of actions).

Oppenheimer, Leboeuf and Brewer (2008) designed four experiments that extended the boundaries of such applications by showing a transfer across modalities and dimensions. They found that manipulating the perception of relative sizes of physical items (e.g., length of drawing lines) or the concept of magnitude (texts with semantic meaning regarding largeness and smallness) not only influenced the same measurement dimensional judgment but also dramatically changed the way people estimated the values on other scales deemed irrelevant, like temperature. Moreover, the impact of providing a frame of reference also appeared to be resistant to expert knowledge. Studies showed that the recruitment of experts like experienced car dealers (Mussweiler, Strack & Pfeiffer, 2000) and legal teams (Englich, Mussweiler & Strack, 2006) were also vulnerable to anchor information and their estimation performance was no better than those of inexperienced populations. Similarly, such impact was indiscriminate between the groups of people with different personalities, traits (Furnham & Boo, 2011) or levels of cognitive abilities (Oechssler, Roider & Schmitz, 2009). Although the relevant investigation of individual differences was inadequate or sometimes controversial, the anchoring effect remained fairly robust.

The existing literature regarding the anchoring effect seems evidently to support its pervasive and ubiquitous influence, demonstrating a universal behaviour associated with number estimation under uncertainty. The impact of this effect was profound, both in terms of its depth and breadth, and it even has a temporal dimension, with some studies suggesting it lasts up to several days (e.g., Blankenship et al., 2008). The anchoring effect was commonly recognised as a specific and perhaps extraordinary instance of *priming*, sharing basic features such as being an automatic and unconscious process and activating a broader conceptual construct (Newell & Shanks, 2014). The influence of priming on the process of number generation was introduced in Chapter 1 when participants were asked to generate four or six-digit numbers randomly. However, the anchoring effect discussed here

predominately tested the numerical responses that are supposed to aid decision-making, unlike producing random numbers.

Researchers have tried to mitigate this effect by explicitly informing individuals about the influence of the anchor or by offering financial rewards as motivation. However, the findings were inconsistent. While most early studies failed to observe positive outcomes (e.g., Chapman & Johnson, 2002), more recent efforts using monetary incentives for tasks with self-generated anchors did show some levels of effective interference (Epley & Gilovich, 2005). Nevertheless, it was widely agreed that people find it challenging to escape the robust influence of anchors. In that regard, several explanations have been proposed to understand the cognitive processes that lead to this ubiquitous biased judgment within a frame of reference.

When the anchoring effect was introduced by Kahneman and Tversky in 1974, they described it as a result of insufficient adjustments from the initial value offered. When individuals are expected to generate an estimate with limited information, they tend to adjust the value from the initially encountered anchor. However, this effortless modification is inadequate, leading to estimates closer to the anchor than they should be. While some studies using self-generated anchors supported this theory (e.g., Epley & Gilovich, 2001), others argued that it only applies when the anchor is outside a reasonable range. For instance, most would agree that 153 is not the likely answer for Mahatma Gandhi's life expectancy, so people would adjust away from that number. However, when an anchor is within a plausible range or provided externally, the *Anchoring-and-Adjusting* model might not fully explain the outcome. An alternative theory, *Selective Accessibility*, derived from confirmatory hypothesis testing, has emerged as another dominant explanation (e.g., Mussweiler & Strack, 1999). This theory suggests that people assess the plausibility of an anchor being the correct answer, accessing relevant features between the anchor and target, which results in anchor-consistent

information biasing the estimates. Recently, a third perspective, *Attitude Change*, has been proposed to address inconsistencies that neither of the first two theories can explain. Traditionally, it was believed that a more extreme anchor would produce a stronger anchoring effect, which is supported by both the *Anchoring-and-Adjusting* and *Selective Accessibility* models. However, Wegener and colleagues (2001) found that extremely high or low anchors did not produce a stronger effect than moderate anchors. They suggested that people might produce counterarguments against the validity of extreme anchors, reducing the anchoring effect due to less change in attitude. In general, the model of attitude change embraces the idea of "multi-process" that the same outcome might be a result of different reasons. Hence, when it comes to a specific scenario like anchoring, both agreement and rejection of the reference point are likely to occur depending on the personal preferences and judges of the attributes of an anchor given the specific situation. While no single theory fully captures the nuances of the anchoring effect, they collectively provide valuable insights into the cognitive processes associated with this bias in estimation.

The testing paradigm for anchoring bias appears to share similarities with that used for examining Benford bias, particularly in tasks like generating numerical estimates for factual knowledge questions. However, the anchoring effect is highly dependent on the initial piece of information provided, a factor typically absent in tests of the first digit phenomenon in psychology. Unless a firm link between Benford bias and anchoring can be established, this effect might be excluded as a potential explanation for the logarithmic distribution of leading digits observed in Benford bias at this stage.

Open Judgment with Preferential Representation

The anchoring effect emphasises the influence of initial data on the process of estimation. Yet, when people face a situation where no values come to mind as "close but wrong" starting points, either from internal self-generated numbers (e.g., memories) or the

external environment (Epley & Gilovich, 2001), how do they arrive at a value that represents their rough estimations? Open numerical judgment explores the use of potential mental shortcuts when judges are not influenced by strong anchors or an evident value to adjust to. It suggests that these processes are streamlined through the preferential representation of certain numbers, such as “Prominent numbers” or “Round numbers” (Converse & Dennis, 2018).

Certain types of numbers like 10, 50, and 200 frequently appear in various contexts, such as written languages in books and stories (Dehaene & Mehler, 1992), estimates of willingness to pay (Whynes, Philips & Frew, 2005), and stock trading transaction values (Converse & Dennis, 2018). These numbers can be categorised as Prominent numbers where the collections of values are based on the powers of tens, their doubles, or halves (e.g., 5, 10, 20, 50, 100, 1000), which can be mathematically represented as $n10^i$ (where i is any integers, and n = ½, 1 or, 2). In addition, part of its number collection is also manifested with the rounding nature of a number containing one significant figure only ($n10^{i,}$ where i is any integer and n is any integer ranging between 1 and 9). Prominent and non-prominent round numbers are pervasive in daily life, serving as goals or motives (Pope & Simonsohn, 2011) and linguistic shortcuts (Zhang & Schwarz, 2013). For example, telling a friend about a thousand-dollar for a new bike purchased when precisely spent $960. Round numbers were also frequently generated as answers to general knowledge questions (e.g., The height of Mount Blanc, between 0-100,000 m) and in random-dots displays (e.g., The number of dots in the frame, between 0-1,000), compared to the answers using a scale for response (Honda, Kagawa & Shirasuna, 2022).

In a study investigating the role of prominent numbers in open numerical judgment, Converse and Dennis (2018) argued that it was unrealistic to mentally screen every possible value on the number line to find a reasonable match. The development of preferential representation of selected numbers results from an interplay between culture, education

evolution, and human cognition. Most educational services in different cultures adopt a ten-base system for numerical and mathematical operations. People might use their body parts, like two hands and five fingers, to navigate quantities. People's cognitive representation of numerosity is in a logarithmic fashion, meaning the numbers between powers of tens are theoretically equally spaced along the log-scaled line. That is why major units or benchmarks like 100 and 10,000 stand out due to their accessibility. In addition, halving and doubling are typically considered the most straightforward calculations, which may easily fulfil the fundamental needs of cognitive efficiency when necessary. This generally aligns with the notion of using heuristic cues when facing an unknown situation (Honda, Kagawa & Shirasuna, 2022). This accessibility-based account for prominent number clustering was detected across people and contexts. Increasing cognitive load, such as following movement instructions while making numerical estimates, appeared to rely more heavily on prominent numbers, especially when cognitive efforts are limited for judgment. Meanwhile, the accessibility-based shortcut was further supported by findings that the clustering of prominent numbers generally exceeded round numbers produced in the open numerical judgment from both observational and experimental results. This preferential representation of certain numbers facilitates unanchored judgments in various contexts, ranging from trivial knowledge questions, such as estimating the distance between two planets, to quantifying visual information, like the number of people on a party boat (Converse & Dennis, 2018).

The accessibility of certain prominent numbers, such as 100, 2,000, or 500,000, in uncertain situations may shed light on why smaller leading digits frequently appear in estimates for unknown quantities, including the notable prevalence of the digit-5. These numbers are likely to be readily produced when the exact value is unknown. However, the preferences for such specific numbers might also manifest in selection tasks, which does not entirely explain why Benford bias is predominantly observed in number generation tasks.

Moreover, while the frequent production of these prominent numbers might be related to the preferences for smaller leading digits, it cannot directly account for the monotonic decline as the first digits increase. Therefore, although it initially seems promising as an explanation, comprehensively understanding the phenomenon of Benford bias requires a more nuanced explanation than just the occurrence of numbers with smaller initial digits.

Conclusion on Number Estimation Using Heuristics and Mental Shortcuts

The heuristics discussed above illustrate how people generate estimated values in responses to the questions under the influence of cognitive bias. While heuristics can be effective and effortless strategies for arriving at conclusions when facing unknown problems, the estimates may largely deviate from the facts, compromising decision quality.

These processes are of particular interest because the questions asked in such tasks are generally consistent with the items used where Benford bias emerged, such as the height of a mountain, the cost of an item, or the likelihood of an event. Throughout this review of representative bias, availability heuristic and anchoring effect, none of them seem directly associated with the bias pursued in the current project. Nevertheless, these discussions progressively uncover a potentially biased pattern of human judgment and behaviour, sharing some similarities with the underlying messages found in Benford bias regarding its unequal preferences for the first digit of numerical answers. Particularly, the cognitive process of estimation under uncertainty could become vulnerable and easily influenced or even distorted, ultimately biasing the estimates for decision-making.

Number Estimation Following Regularities in Nature

Demonstrations using heuristics in number estimation tasks highlight the potentially biased processes in judgment and reasoning. However, the precise conditions that elicit the use of heuristics and the relationship among those mental shortcuts remain systematically

unexplored, lacking support from theoretical frameworks (Eysenck & Keane, 2015). This approach predominantly focuses on the cognitive process but detaches from other factors that are vital for decision-making, such as motivations. Newell and Shanks (2014) argued that inadequate effort has been given to some key findings; for example, the cross-modality anchoring effect was only established through the experiments by Oppenheimer, LeBoeuf and Brewer (2008), with no further replication since then. Some observations suggested that simple changes in information quality or framing of descriptions could mitigate the bias. Base-rate neglect could be suppressed to some extent by adopting the terms describing natural "frequencies" compared to the use of more complex relative terms (e.g., probabilities) (Lesage, Navarrete & De Neys, 2013). Considering the limitations and critiques regarding the traditional view of the heuristic-and-biases approach, Kahneman (2003) proposed a dual-process theory to account for a more complex cognitive process in judgment and reasoning. Two systems are activated sequentially: System 1 provides a quick, involuntary, effortless and implicit judgment (e.g., representative bias), which is then redirected to System 2 for a more cautious, time-consuming, and regulated assessment.

The dual-process theory seems to acknowledge that decision-making also involves part of rational judgment and reasoning. Thus, this brings us to a new direction for examining the relevant literature about number estimation that follows the norms of optimal reasoning. Again, to address the correspondence to the purposes of the current study, we specifically looked for the studies using a paradigm similar to those on which Benford bias focused.

Behaviour Guided by Statistical Learning

It was intriguing to see that base-rate neglect could be attenuated to some extent by simply reframing it with more straightforward terms for describing frequencies. The effectiveness of this strategy perhaps potentially lies in a well-documented theory of learning about people's sensitivity to the likelihood of natural events happening in their everyday

lives. Gigerenzer and Hoffrage (1999) argued that figuring out complex mathematical computations like fractions leads to poor judgment because people are usually exposed to the mere frequencies of the associations of events through natural sampling. Attention to frequencies has been inherently observed in humans and animals, and it is in line with the evolutional perspectives for survival in the environment. Newborn infants were found to present longer focus when seeing a continuous drawing of items with low probabilities (Xu & Garcia, 2008). Additionally, the acquisition regarding the frequencies of the events was not only limited to the patterns linked to visual experience; other forms of association, such as music notes and linguistic experiences (e.g., Gebhart, Newport & Aslin, 2009), were also detected. The evidence of statistical learning and sensitivity to the distributional information has been introduced in detail in Chapter 1 when discussing the development of the Recognition Hypothesis for Benford bias, so the relevant empirical findings are not repeated here.

In addition to the points made in the first chapter, individuals not only track the probabilities, but also tend to exhibit relevant response patterns according to the regularities detected. Toddlers showed stronger preferences for the toys repeatedly associated with sounds and lights after being exposed to such a relationship (Waismeyer, Melzoff & Gopnik, 2015). Their choices of a particular toy depended on the probabilistic pattern of the causal evidence manipulated by adults, suggesting the basic reasoning of the cause-effect relationship without additional instructions (Kushnir, Xu & Wellman, 2010). However, most tasks are well-informed with all the relevant information about their patterns before the reaction phase. In real-world situations, people often encounter partial or limited exemplars in the initial stage, and their decisions or judgments require constant updating with incoming information.

Hence, some studies were dedicated to understanding how people react to the dynamic environment when integrating more probability data through experience over time. Two types of behaviour were typically identified by extent research, either showing a pattern of "probability matching" or "maximising" (Vulkan, 2000). Probability matching is best illustrated by the multiarmed bandit task, where the odds of rewarded outcomes require explorations based on personal observation of one's own choice. That is, even if people are aware of the distributions of the probabilities, they still exhibit a manner that follows the probabilities of the options at the cost of time and payoffs to accumulate new information that may improve future performance (Russo, Roy, Kazerouni, Osband & Wen, 2017). On the contrary, the maximising strategy predominately focuses on optimising the gains by constantly choosing the option that is more likely to maximise the profit. Recently, it was found that both children and adults were capable of smoothly transitioning the strategy from matching to maximising throughout experimental sessions when they gained more experience with the pattern structure, though it took longer to crossover for the kids (Plate, Fulvio, Shutts, Green & Pollak, 2018). In fact, probability matching should not be considered a failure for decision-making; instead, continuous sampling information about the environment, as one of the optimal strategies, reflects rational judgment when coping with uncertainty (Green, Benson, Kersten & Schrater, 2010).

Many studies emphasise the unconscious acquisition of probability information. In one study, without explicit instruction for participants to acquire any probability patterns, researchers still found that their reactions tended to follow the regularities of the visual stimuli presented in a temporal order (Fiser & Aslin, 2002). Many lab observations have documented statistical information in the environment without apparent awareness (e.g., Campbell, Zimerman, Healey, Lee & Hasher, 2012), often discussed as examples of everyday automatics. On the other hand, people tend to seek out surrounding information actively and

constantly update it with the existing knowledge as they navigate this dynamically changing world. People need to balance between exploiting what is known to maximise immediate performance and investing to accumulate new information that may benefit them in the future.

In exploring statistical learning, it was noted that few tasks explicitly required participants to make numerical estimates as an outcome measure. Although not overtly specified, it is likely that most decisions and behavioural patterns in these tasks involved participants making rough estimates and adjusting their perceptions based on their experiences. However, the reviewed studies did not directly explore or detail the process of making these estimates. Nevertheless, considering that statistical learning is often linked to rational judgment and reasoning in controlled laboratory environments, it seems plausible that individuals would naturally conform to patterns prevalent in nature, such as Benford's law. This realisation prompted a re-evaluation of the Recognition Hypothesis, initially proposed to explain Benford bias, as discussed in Chapter 1. Previously, it was presumed that participants came into the laboratory with pre-existing knowledge of first-digit distribution, gained through long-term exposure, and were directly tested with estimation tasks. However, statistical learning can also occur through active exploration and experiential learning. This insight suggests that it might be beneficial to immerse participants in scenarios where they can directly encounter these statistical regularities and assess whether they can utilise this learned information in problem-solving contexts. Such an approach proposes an alternate methodology to investigate whether Benford bias arises from an innate sensitivity to statistical patterns in one's environment. These reflections and theoretical considerations have significantly influenced the design of current study. More detailed information about this will be provided in the upcoming chapters.

Optimal Statistical Inferences in Cognitive Judgments

The literature concerning the relationship between probability acquisition and associated behaviour addresses people's sensitivity to the statistical information present in their environment. This partly recognises that the cognitive judgment under uncertainty is not entirely driven by the error-prone approaches observed in the lab using artificial patterns. A recent study by Griffith and Tenenbaum (2006) strongly supported the idea that people's cognitive judgment can be explained in terms of optimal statistical inferences that follow prior probabilities with evidence from estimation in real-life scenarios. Similar to the systems of human memory and perception (e.g., Weiss, Simoncelli & Adelson, 2002), they argued that people's judgment constantly embraces new data and combines it with past statistical information for updates.

Their experiments captured our attention not only because the estimated values were collected as outcomes, but also because of the identical domains of the testing items used, as those found in Benford bias. Their questions were selected from familiar activities in daily life, such as the movie gross and run time, poem lengths, lifespan, lengths of marriages, terms of U.S. representatives, baking time for cakes and telephone waiting times. Having multiple variables ensured that the underlying distributions followed different classes of regularities, based on the statistical observations in nature and society. Participants were asked to make predictions given a reference point by a brief description, for instance, estimating a person's lifespan knowing his current age at 65 for an insurance case. The estimates for the questions then were contrasted against predictions made by the optimal Bayesian model, where its probability is proportional to the product of likelihood and prior distributions, given the assumption of random sampling and reasonable expectations about how it should be distributed. Most of their data showed substantial conformity, where the relevant judgment almost replicated the optimal prediction from Bayesian calculations. Nevertheless, a variation

in performance was observed when participants faced unfamiliar questions, that is, estimating the reigns of pharaohs. This was explained as a reasonable deviation due to their limited direct experience and knowledge of ancient Egypt. Such errors were consistent with their further inspections when assuming participants utilised an analogy such as modern monarchs and adjusted means insufficiently when encountering an inexperienced situation. Despite this, it appeared that people were aware that data come from different distributions, and their responses agreed with the appropriate prior distribution depending on the context. For instance, when estimating a person's total lifetime, the responses followed the pattern of appropriate Gaussian prior, but when predicting the total revenue of a movie, their answers mimicked the behaviour of a power law. It seems quite sensible to assume the use of the multiplication rule for the current intake to approximate the total movie gross (e.g., 1.5 × 7 million), but it should not be applicable to multiply a constant with the current age to estimate a person's lifetime (age 75 × 1.5), according to Griffith and Tenenbaum (2006). This suggested sufficient but (perhaps) implicit knowledge about the distributions of everyday quantities. Their conclusions challenged the traditional view that people are prone to heuristic errors and insensitive to prior probabilities (Kahneman & Tversky, 1974). Moreover, the innovative design of providing a starting point in these studies not only highlights people's sensitivity to the distributional data but also emphasises their capacity to accurately combine new information with existing knowledge about the world around us.

However, the observation that people could make an optimal prediction in their everyday cognition was criticised as an artifact due to aggregation, as individuals only produced a single estimate for each domain of the question. Mozer, Pashler and Homaei (2008) employed a competing model, Min*K*, for contrast, assuming people merely utilised a few instances in memory for estimation. They replicated the identical findings by aggregating individual responses. They even demonstrated a more precise prediction within this

alternative model, suggesting the accumulation effect regardless of the level of knowledge about the distributions of everyday quantities. However, this was still insufficient to empirically conclude that individuals lacked sufficient knowledge. Considering the limitations of the experimental design by Griffith and Tenenbaum (2006) as well as to explore people's estimation performance at the individual level, Lewandowsky, Griffiths, and Kalish (2009) decided to collect repeated measures from participants and discriminate between two models explaining humans' judgment. To repeatedly sample individual reactions, iterated learning was introduced so that the outputs of previous learning trials could randomly become the inputs for the new contexts, resulting in a total of 160 responses (20 trials × 8 targets). They found that most estimation performance strongly favoured the predictions calculated from the Bayesian model rather than Min*K*, supporting the claim that accurate prediction about everyday events is a wisdom of individuals that relies on sophisticated knowledge. This systematic test powerfully demonstrated that individuals' optimal judgment regarding real-life estimation is due to their sensitivity to the underlying distribution of a variable, further rebutting the opposing view that relies on a limited number of instances in everyday prediction.

The demonstration of the relevant knowledge of the underlying distribution regarding the estimated variables in everyday prediction offered additional support for individuals' sensitivity to the statistical regularities surrounding them. In contrast to the recognition hypothesis proposed in explaining the Benford bias concerning the acquisition of the first-digit distribution, people also seem aware of how each variable should be distributed. Such conjecture, combined with the mathematical fact that BL emerges from the numbers aggregated from the variables following the power law (Berger & Hill, 2015; 2020), directly built up the basis for establishing an alternative explanation as one of the major testable directions. This new hypothesis, based on the observation of individuals' sensitivity to the

underlying distribution of a variable, will be discussed in detail in Chapter 3 when systematically introducing the framework of the present project.

Conclusion on Number Estimation Following Regularities in Nature

Determining the level of knowledge utilised in everyday judgment is fundamental to understanding how people make decisions and solve unknown problems. Unlike the previous section, which heavily focused on the cognitive bias, the established optimality of human judgment in real-world situations provides alternative insights into how estimates are formed under uncertainty. However, relying on mental shortcuts or statistical regularities to answer questions does not necessarily have to be mutually exclusive. For instance, estimates influenced by an irrelevant anchor may still adhere to the underlying distributions of that variable to some degree. Each method offers a unique viewpoint that highlights a possible cognitive process aiding individual's judgment when making reasonable predictions or estimates. The relevant body of work discussed here is valued for inspiring the proposal of potential hypotheses regarding the emergence of Benford bias and reliable methods to test them, which will be elaborated on in later chapters.

Quantity Estimation in Visual Perception

The estimation tasks discussed in the previous sections primarily focused on understanding how people produce estimates under certainty that require specific or general real-world knowledge (e.g., how long it takes to travel from Paris to London). In contrast, *pure numerical estimation* demands minimum reliance on real-world knowledge for processing. Instead, it predominantly targets approximating quantitative values that use numbers as inputs, outputs, or both (Booth & Siegler, 2006). Examples include estimating the results of arithmetic calculations (e.g., 101× 25), the quantities of items in a visual scene (e.g., candies in a jar), or determining the relative position of a value on a number line.

However, the purpose of the present project is specifically related to the spontaneous generation of numbers as answers to unknown or even unknowable questions. Therefore, it was chosen to avoid expanding the scope into realms where estimation involves computational skills that can be readily assisted and verified promptly by tools (e.g., calculators), or where the process of free estimation was absent, such as only identifying a plausible spatial position for a number within defined boundaries. In that regard, this section exclusively reviewed the literature about how people estimate the quantities of visual elements, where Benford bias has also been reported based on existing observations from pictured jellybeans or dots (e.g., Chi & Burns, 2022). One of the major test components where Benford bias typically emerged is estimating the quantities of visual elements. These materials include showing nine different pictures of hundreds of dots or jellybeans in a jar. Moreover, it was found that changing the level of information from presenting blurry to dynamic pictures did not yield a different degree of fit to Benford's law. All the existing evidence reported by the visual estimation tasks strongly supported the robustness of Benford bias.

Uncertainty in estimation tasks can arise from inadequate information presented in a visual scene. Additionally, time constraints can also introduce uncertainty, as achieving precise counts within a short period of time may be impractical. Given that many visual stimuli in the environment are either fleeting or too numerous to count, rapidly and effortlessly estimating quantities in a scene is a significant challenge. Understanding how the visual perception system and associated cognitive processes in approximate quantities of visual targets within time constraints is crucial not only for everyday activities, such as estimating traffic volume to plan a route, but also offers broader insights beneficial to society and the natural world. For example, a biologist needs to monitor the number of rare birds in a flock at a conservation sanctuary, or a social scientist intends to study population distribution

during a public event. These scenarios underline the importance of precise quantity estimation.

So far, the framework of statistical learning has shown promising potential in relation to the underlying reason for Benford bias. Nonetheless, we remain keen on exploring further insights that could enlighten us on this journey. Thus, this part of the review is dedicated to presenting the relevant mechanisms associated with number estimation in visual perception, where Benford bias also emerged. The past sections searched for any clues specifically linked to the process of answering the trivia questions. Thus, we hoped the investigation into the visual estimation tasks might shed some light as well.

Numerosity vs. Number

To clarify the mechanism behind how people approximate the quantities of visual elements, it is essential to distinguish between two frequently used terms in these studies: "numerosity" and "number." "Numerosity" typically represents the quantitative information of the items in terms of size and magnitude as visually perceived during physical experiences in the world. But the word "number" is usually used to denote a numeral representation in linguistic format, such as Arabic numbers (i.e., 1, 2, 3), commonly used in mathematics today (Harvey, 2016). Different cultures also had their own number words to verbally quantify the numerals. It has been found that the symbolic acquisition and processing of numerals can be facilitated by having a more transparent verbal representation, such as the place-value system ("十二" ten two for twelve) adopted in China (Miller, Kelly & Zhou, 2005). To address the observations rooted in the Benford bias, which primarily relies on Arabic numerals, this project has chosen to concentrate solely on the paradigm that utilises Arabic numbers for symbolic representation. This not only helps to sidestep potential linguistic discrepancies in verbal explanations but also aligns with the globally accepted standards for conveying numerical data in the contemporary context.

Cognitive Mechanism of Enumerating Small Sets

It was generally recognised that visual perception involves two separate and remarkable pathways for rapidly judging the quantities of visual items (e.g., dots arrays) without counting, especially given limited viewing time. When people were required to quantify a very small set of objects at a glance, almost no errors occurred for sets below five (e.g., Jevons, 1871). This fast and precise enumeration, occurring within a fraction of a second (i.e., 150ms) in response to low numerosities up to four, is known as subitising (Dehaene, 1992; Kaufman, Lord, Reese & Volkmann, 1949). The normal subitising range is limited to four items; however, the accuracy on average was also found to be impressively precise between four and eight (e.g., Kaufman et al., 1949). Moreover, a set of 20 items was reported as the lowest numerosity that prevents the utilisation of rapid enumeration, such as subitising without counting (Revkin, Piazza, Izard, Cohen & Dehaene, 2008).

Subitising is considered distinct from counting because the time consumed to subitise is considerably shorter than counting one item (i.e., 300ms) (Choo & Franconeri, 2014). Although both share common neurological pathways, activating the parietal lobe and intraparietal sulcus, PET scans have indicated specialised brain areas involved in counting, such as the inferior frontal cortex and anterior cingulate area, which are typically responsible for attentional shifts (Corbetta, Shulman, Miezin & Petersen,1995).

The ability to discriminate sets of 1, 2, and 3 items, like small dots, from one another was observed in early infants, suggesting an evolutional and developmental predisposition for rapid identification of quantities up to four in space (e.g., Antell & Keating, 1983). Children in elementary schools were generally able to demonstrate the ability to subitise, similar to the performance of adults (Moore & Ashcraft, 2015). The mechanism of the Object Tracking System (OTC) has been proposed to explain this. The core principles of this nonverbal system depict a mechanism that allows people to track individual entities in parallel to

perceive their boundaries and predict their movements (Piazza, 2011). However, this tracking system has a limited capacity, up to four entities at a time. Such cognitive constraint was not only found when enumerating quantities but was also reported by other studies involving memory tasks like accurately holding simple features of a small set of objects (Alvarez & Cavanagh, 2004). Supporting evidence for such a perception limit of OTC was also indicated by reaction time (RT) analysis. When assessing RT for identifying a set of objects with more than four, it became increasingly slower compared to a relatively flat and unchanged RT when the objects remained within the range of one to three (Geary & Moore, 2016). This highlights a notion that a separate system might be activated rather than subitising when presenting a large set of objects for quantity judgments.

The current project focused on investigating how individuals estimate the number of items under conditions of uncertainty. The Benford bias is commonly observed when estimating large quantities of items. In contrast, the phenomenon of subitising, the rapid and accurate recognition of small numbers of objects, may not be directly relevant to our research scope. Furthermore, the relative precision in estimating small quantities may not represent an uncertain situation, thus offering limited insight into how people navigate truly unknown circumstances. Nevertheless, the concept of subitising provides a valuable understanding of the visual processing mechanisms used to quantify small groups of items in specific spatial and temporal contexts. It serves as an informative point of comparison for estimations involving larger quantities.

Cognitive Mechanism of Approximating Large Sets

Contrary to the numerical judgment of a small set of physical objects up to four, estimating the quantity of a large volume of visual elements is distinctively regulated and governed by the Approximate Number System (ANS), which mentally represents and processes numerical magnitude information (Dehaene, 1997). Estimation performance using

ANS is typically slower and less accurate than the processes guided by OTC because they are recognised as different and separated systems in terms of their cognitive and neurological mechanisms (Piazza, 2011). EEG data have shown that subitising triggered early posterior brain activities, while the ANS corresponded to the modulation of a later posterior wave when dealing with large numerosities (Hyde & Spelke, 2009). From a developmental perspective, ANS is typically regarded as an important predictor (though not the only one) longitudinally associated with mathematical skills throughout one's lifetime (Chen & Li, 2014; Starr, Libertus & Brannon, 2013). Moreover, children with dyscalculia, a disability in symbolic number processing and calculation, demonstrated intact ability to subitise but impaired performance in large numerosity comparison tasks (Decarli, Paris, Tencati, Nardelli, Vescovi & Piazza, 2020). The evidence clearly addressed the distinction between these two systems across both adulthood (e.g., Piazza, Fumarola, Chinello & Melcher, 2011) and infancy (e.g., Oakes, Hurley, Ross-Sheehy & Luck, 2011).

ANS is manifested by two properties in relation to the estimation of numerosity: 1) the estimates generally become more variational as the number of physical items gets larger; for instance, the variation of estimation for 80 dots is in scalar proportion, almost twice as big as that for 40 dots; 2) the discrimination performance of numerosity is ratio-dependent; in other words, people find it harder to distinguish 100 dots from 90 dots, compared to 100 dots from 50 dots, which was typically captured by Weber's law (Odic & Starr, 2018). So, the discriminability of the numerosity is defined as an index of one's ANS acuity. This logarithmic representation of numerosity agrees with the notion of the mental number line theory about the compressed spaces in larger numbers as a cognitive apparatus. However, recent studies reported a collective observation of a developmental shift from the logarithmic representation to the linear representation of numerical magnitude, occurring progressively at different ages for different ranges of numbers. For example, the comparison between toddlers

and second-year school children for 0-100 on number lines showed that a shift of ANS acuity occurred with increased exposure and experience over time (Siegler & Booth, 2004). However, when facing unfamiliar numbers, the intuitive behaviour immediately returned, evidently following the logarithmic representation (Booth & Siegler, 2006).

Numerical estimation in visual perception normally involves two aspects: non-symbolic estimation and symbolic estimation. Non-symbolic processing primarily deals with numerosity without using numerical representations, such as the tasks that compare the relative size of two streams of dots quantities (Smets, Sasanguie, Szücs & Reynvoet, 2015), thus was directly supported by ANS. Infants, as a few months old, have shown sensitivity to differences in visual magnitude given a large ratio between two arrays, even before the development of linguistic representations (e.g., 10:30, Coubart, Izard, Spelke, Marie & Streri, 2014; Xu & Spelke, 2000). However, numerosity judgment based on the discrimination tasks has also been found to covariate with other dimensions such as item density, size, luminance, the area occupied, and time, leading to a systematic over or under-estimation (Castronovo, Crollen & Seron, 2010). It is reasonable to expect that some properties of objects inherently vary as quantity changes; for instance, space generally becomes more occupied with more dots. Several recent studies have broadened their scope to incorporate more real-world variables like depth perception and the movement status of objects. With judgment or reproduction tests, there was a noticeable trend of overestimation when using 3D stimuli in contrast to 2D targets (Aida, Kusano, Shimono & Tam, 2015). On the other hand, a robust underestimation was observed when employing dynamic targets as opposed to static ones (Au & Watanabe, 2013; Poom, Lindskog, Winman & van den Berg, 2019). This evidence suggests that the interplay between numerical judgment and non-numeric dimensions deviates from the independent nature of ANS. To reconcile the discrepancies between domain-specific and domain-general properties of this representation system, an integration

hypothesis was proposed to emphasise the nature of integrating visual features based on numerosity approximation (Gebuis, Kadosh & Gevers, 2016). A more recent account also speculated that the difference in estimation performance could result from a competition between ANS and other perceptual or sensory representation systems. This was evident in neurophysiological studies showing conflicts at the encoding and selection stages (Odic & Starr, 2018).

However, to explicitly generate symbolic quantities that reflect the numerosity of physical objects, one also needs to map the symbolic terms (e.g., Arabic numbers or number words) to the non-symbolic numerosity processed. That is where the ANS mapping account comes into play. Symbolic estimation places more emphasis on the mechanism of mapping numerals to the quantity dimension, such as approximating the number of dots in the arrays (e.g., Booth & Siegler, 2006). However, the interaction between non-symbolic and symbolic representation leads to another ongoing debate regarding whether the non-symbolic processing of numerosity provides foundations to the symbolic processing when extracting the numbers to represent the quantities of numerosity of physical elements viewed. The traditional view claims that the symbolic processing of numerosity is deeply rooted in the non-symbolic estimation. Many researchers have found that the estimated numbers to the dots arrays not only reflect the logarithmic properties of ANS, such as the ratio and distant effect (e.g., Dehaene, 1997; Holloway & Ansari, 2009), but also a systematic *underestimation* for larger quantities, consistent with the compressive fashion of numerosity representation (e.g., Chesney & Matthews, 2018; Crollen, Castronovo & Seron, 2011). Furthermore, improved ANS acuity seems to enhance and facilitate better symbolic processing in visual estimation (e.g., Mundy & Gilmore, 2009; Wong, Ho & Tang, 2016). The ANS also plays a vital role in acquiring numerical words and symbols for the associated meaning as a referent, at least for the smaller numbers learned early on (Dehaene, 2001;

Feigenson, Dehaene & Spelke, 2004). For instance, it is common to see kids establish a relationship between the number words (1, 2 and 3) acquired and their body parts (e.g., counting fingers). The ANS mapping mechanism was also echoed through brain imaging evidence, where the same structure activated for numerosity was later recycled for the representation of number symbols (Dehaene, 2005; Geary, 2013).

Recent observation, however, challenged this unidirectional relation by arguing that the tasks related to symbolic processing might be independent of non-symbolic processing. The distant effect detected in the symbolic estimation could be a simple product of the task-related rather than a numerosity-related phenomenon, as such a finding was also replicated by the tasks based on the order of letters in an alphabet table (Van Opstal, Gevers, De Moor & Verguts, 2008). Moreover, the symbolic processing of numbers is not always accompanied by the activation of numerosity representation in adults, as indicated by EEG results (Van Hoogmoed & Kroesbergen, 2018). The acquisition or manipulation of larger numbers relies more on the semantic network of symbolic processing than the ANS mode, as illustrated by a diminishing association between these two systems with age (Kolkman, Kroesbergen & Leseman, 2013). In addition, the studies supporting the bi-directional mapping hypothesis even found an overestimation effect (rather than underestimation) in tasks that required an interaction between non-symbolic and symbolic processing (Crollen, Castronovo & Seron, 2011).

However, it was noted that such evidence was heavily task-dependent, with most conclusions drawn based on the behaviour specifically related to the ability to discriminate, compare, or reproduce. When the paradigm was limited to a simple perception task with free estimation (e.g., Chesney & Matthews, 2018), that is, generating symbolic answers (e.g., Arabic numbers) to represent the numerosity of a non-symbolic collection, the results generally supported the ANS mapping account, showing consistent underestimation.

Although the estimation of the absolute quantity might be skewed, the perception of relative proportions remained unaffected, as the ratio judgment of numerosities was surprisingly precise (e.g., Chesney & Matthews, 2018; 2022). This might illustrate the inherent nature of relational ANS; thus, experiencing a sense of proportion provides a better grounding for symbolic processing.

The systematic underestimation of large quantities, as explained by the Approximate Number System, has been noted in studies related to Benford's law, such as estimating hundreds of jellybeans in a jar (Chi & Burns, 2022). However, the connection between this underestimation performance and a preference for smaller leading digits remains unclear. Earlier research showed that using clearer, more informative images of jellybeans diminished underestimation compared to blurry images. However, improving the visual clarity did not markedly affect compliance with BL. Thus, paralleling insights from the mental number line, the compressive representation of numerosity may not provide a solid foundation for understanding the occurrence of Benford bias in large quantity estimations.

Visual Processing in Approximate Number Estimation

The conceptual framework of ANS and ANS mapping account for the symbolic processing and non-symbolic estimation of numerosity and illustrates a cognitive mechanism for how people generate estimated numbers corresponding to the magnitude of visual objects with large sets. Yet, this does not offer additional insights into how the approximate number is derived from visual experiences. One dominant model proposes that approximate number estimation relies on the automatic visual grouping of objects in rapid, parallel, and global processing of numerosities.

The concept of "visual grouping" was first emphasised in Gestalt psychology when studying human perception (Wertheimer, 1924), where grouping equal parts by proximity into a higher unit was considered a pre-attentive process (Neisser, 1967), benefiting from

simplified estimation procedures due to reduced free parameters (Compton & logan, 1999). The idea was based on the fact that the visual system is adaptive and flexible in organising the global structure of the visual scene (Pomerantz, 1981). While few empirical works have been done to clarify the underlying mechanism, research on the relationship between approximate number estimation and spatial organisation of the numerosity provides some clues. It was found that the clustered objects were generally perceived as less numerous than the randomised patterns (Ginsburg & Goldstein, 1987), while the randomly distributed items were perceived as less numerous than those with a regular arrangement (Ginsburg, 1976). Moreover, underestimation performance could be significantly enhanced by grouping similar elements in a higher order (He, Zhang, Zhou & Chen, 2009). Meanwhile, presenting groups of items on the periphery of the display encouraged faster numerical enumeration than those shown in the centre, indicating that visual grouping of elements potentially occurred before the enumeration stages (van Oeffelen & Vos, 1982). The automatic activation of visual grouping was also evident in a study using random dot arrays (without explicit grouping cues) when introducing voluntary control under the influence of feedback. However, underestimation performance still persisted (though reduced) due to perceptual clustering (Im, Zhong & Halberda, 2015). Their study found a default process for grouping items within a 4-degree visual angle that was consistent across individuals and experimental image trials. Other studies investigating subitising also extended their examination by proposing that people may group larger sets of items (e.g., dots) into smaller collections of three or four, allowing quick, parallel and effortless processing through mental calculation, such as multiplication (Ciccione & Dehaene, 2020; Wender & Rothkegel, 2000). This process is introduced as “groupitising.” Regardless of different perspectives, the findings of the grouping effect in explaining the visual process of approximate number estimation seem

consistent with the ANS account, which can be manifested by rapid and global processing across the entire view of the visual scene.

On the contrary, recent evidence that emerged from visual perception studies with the help of an eye-tracker argued that numerical extraction in the approximate number estimation is a “serial, foveal accumulator” (Cheyette & Piantadosi, 2019). Their observations primarily considered the roles of eye movement and attention, which past research largely disregarded. The findings recognised that visual estimation could be a product of serial accumulation determined by the number of visual fixations. People appeared to collect the numerical information based on what was seen in the centre of the visual field while potentially ignoring the numerosity in the peripheral vision. That is, quantities were accumulated by moving the centre of gaze in sequences and aggregating nonverbally to an approximate number. Hence, estimation performance was strongly improved as the number of points captured in the fovea increased over time. This alternative view seems to refute the rapid and parallel visual processing concluded by the existing literature with pure numerical measures. Still, in fact, as supplementary evidence, it enriches the understanding of how the approximate number estimation is extracted from visual experiences.

The research on visual processing in approximate number estimation brought up a debate regarding the parallel process of grouping items versus the serial and foveal accumulation of items. However, neither approach directly explains how people quantify numerosity with numerical representation in a detailed process. While general strategies like multiplying small groups of quantities provide some direction, they fall short of elaborating on the process that connects these strategies to numerical outputs. Therefore, without a detailed explanation of the exact calculation steps, comprehending the relationship between these processes and the logarithmic distribution of the first digits in visual estimation tasks remains obscure.

Conclusion on Quantity Estimation in Visual Perception

Unlike subitising, which involves accurately perceiving a small set of visual targets, generating estimates in response to the numerosity of large groups of visually experienced elements is generally supported by the Approximate Number System (ANS). The logarithmic nature of this non-symbolic representation of the numerosity can also be transferred to the symbolic representation, leading to a systematic biased answer - *underestimation.* Such performance is prone to covary with different nonnumerical properties of the elements perceived, such as duration and density, although there is still an ongoing debate about whether this is a task-dependent effect.

Estimating the quantities of visual elements is far from the optimal judgment, but it is recognised as an adaptive and effective mechanism that functions reasonably well for everyday activities. Even if this cognitive or visual apparatus is not perfect for quantifying the exact physical items, the holistic and rapid process of extracting the approximate number impressively supports human beings in surviving and thriving in an environment flooded with numerical information.

Estimating large quantities, akin to the theories of the mental number line, demonstrates a compressive effect on the scale, impacting estimation behaviour. While numerosity perception involves judging magnitudes, the Benford bias is a phenomenon that occurs irrespective of the quantity size. Although the numerosity representation also presents a logarithmic scale as Benford's law, this does not necessarily encapsulate the full essence of the first digit phenomenon observed in number estimation.

Brain Mechanism for Number Estimation

While this chapter is dedicated to understanding the cognitive mechanisms behind number estimation, we believe that reviewing the neurological underpinnings shall provide

supplementary insights into aspects not previously thoroughly explored. Although the cognitive mechanism of number estimation for decision-making lacks a systematic theoretical framework, the assessment of generating reasonable estimates to unknown questions has been widely adopted in clinical settings to examine patients' cognitive abilities.

The development of such utility can be traced back to 1978 when Shallice and Evans investigated cognitive deficits in patients with unilateral focal cortical lesions on the front lobes. They were asked to make estimations in response to fifteen general knowledge questions regarding the length, width, weight, and height of various items (e.g., post office tower), the speed or quantities of animals, and the income of an occupation. The precise answers were generally not expected to be known. Still, a reasonable response can be generated if individuals appropriately identify relevant knowledge, retrieve basic facts, monitor rationality, and repeat those procedures when necessary (Bullard, Fein, Gleeson, Tischer, Mapou & Kaplan, 2004). Their examination revealed that the estimated values by the patients with lesions in the anterior brain region deviated dramatically from the expected range compared to those generated by other patients (posterior group), showing more extreme and bizarre responses without awareness. This implied that damage to the front lobes, which is typically associated with executive functioning, directly impacted the process of number estimation (Stuss & Levine, 2002).

Since this initial finding, the use of Cognitive Estimation Tasks (CETs) has become prevalent as one of the major assessment packages for clinical interviews and research. Numerous studies have demonstrated the utility of the revised version of CETs through its application in multiple domains. Consistent results were also observed in patients with neurological or psychological disorders such as autism spectrum disorder (Liss, Fein, Bullard, Robins & Waterhouse, 2000), Alzheimer's disease (Della Sala, MacPherson, Phillips, Sacco & Spinnler, 2004), and schizophrenia (Jackson, Fein, Essock & Mueser,

2001). Their explorations established that poorer performance in number estimation was attributed to impaired executive function, which involved a range of mental processes such as planning, cognitive adaptation, inhibition control, self-monitoring, and self-correction (Silverman & Ashkenazi, 2016). Moreover, it was generally agreed that this entire set of mental abilities is vital in the estimation process, as no single specific component could be identified as the primary source of cognitive estimates. A review of this model additionally suggested that estimation tasks not only required the activation of central processing control, but also involved working memory to interact with long-term declarative memory (Brand, Fujiwara, Kalbe, Steingass, Kessler & Markowitsch, 2003). Hence, it can be partly considered an alternative instrument to assess crystallised intelligence and cognitive reserve (D'Aniello, Castelnuovo & Scarpina, 2015). In short, a collective body of clinical evidence emphasises that generating reasonable estimates to general knowledge questions demands coordination among complex cognitive processes.

The role of executive functions seems not only to impact the way people generate numerical responses that require general knowledge of the world and experiences, but also to participate in the visual perception process of approximating large quantities of physical items. For example, to suppress the nonnumerical feature of physical elements while perceiving the numerosity, one needs to activate inhibition control to carry out such a task (Starr, DeWind & Brannon, 2017). However, approximation number estimation is primarily supported by ANS, which involves the coordination between the non-symbolic code (" **: .**") of the numerosity in the real-world experiences and the other two symbolic codes of numerals as linguistic format (i.e., Arabic numbers "3" or verbal terms "three").

The Triple Code Model (TCM) of numerical cognition was established to illustrate their distinct but overlapping neurocognitive mechanism (Dehaene, 1992; Dehaene & Cohen, 1995). It is generally agreed that the activation of ANS by experiencing numerosity in the

real world is primarily correlated with the posterior parietal cortex and the intraparietal sulcus (IPS) (Nieder & Miller, 2004). In addition, IPS was commonly recruited for all the discrimination tasks regardless of representation codes (Dehaene, Sergent & Changeux, 2003). However, some research also reported involvement from the early visual system at the retinotopic level in V2 and V3 for the non-symbolic magnitude discrimination, highlighting a role proportionated from lower-level neural substrates (e.g., Cavdaroglu, Katz & Knops, 2015; Fornaciai, Brannon, Woldorff & Park, 2017; Roggeman, Santens, Fias & Verguts, 2011).

Distinct neural substrates also function separately for two types of symbolic representation: visual Arabic number forms and auditory verbal word frames. It was found that children need to acquire a well-established association between the non-symbolic and verbal codes before comprehending the Arabic number system (Fayol, Seron & Campbell, 2005). Patients with impaired processing of Arabic numbers showed intact ability to perform tasks related to numerosity discrimination or verbal number words (Cipolotti, Warrington & Butterworth, 1995). The Visual Number Form Area (VNFA) is commonly recognised to support the processing of Arabic numbers, yet the specific regions involved, with candidates like the inferior temporal gyrus, remain debated (Yeo, Wilkey & Price, 2017). An alternative observation detected that the left angular gyrus might correlate more with the VNFA (Price & Ansari, 2011). For the process of verbal representation, the corresponding neural substrates are primarily located in the left hemisphere of language networks, such as the inferior frontal gyrus, supramarginal gyrus, and angular gyrus (e.g., Schmithorst & Brown, 2004).

The studies on Cognitive Estimation Tasks and the proposed TCM provide a comprehensive overview of the neurological and biological foundations for number cognition and representation, including the workings of the approximate number system. This offers additional insights, complementing the cognitive mechanisms discussed in earlier sections,

particularly in understanding how people produce estimates under uncertainty. For instance, the distinct neural substrates identified for the representation of Arabic numbers and number words highlight the importance of the specific output format used during assessments. So far, tests demonstrating the Benford bias have primarily focused on Arabic numbers as the outcome measure, leaving the robustness of this phenomenon in relation to verbal terms yet to be clarified. Although these aspects related to brain mechanisms, such as the role of executive functioning, have not been directly involved in the test of the current project, they provide directions for interpreting the results and shaping future studies, which will be elaborated in the final discussion.

Improving the Quality of Number Estimation

Research on how individuals produce estimates in response to general knowledge questions and quantities of visual targets has shown that these responses can be largely distorted by heuristics or mental representations of numerosity. Hence, it is essential to explore effective strategies to improve the quality of number estimation and produce more precise answers. Although this topic is not directly in line with the primary purposes of the current study, understanding the techniques utilised for making better estimations shall also be informative in revealing the underlying process of how an estimate is generated. Several methods have shown effectiveness in enhancing judgment quality. However, this review focused on the techniques specifically aimed at improving the performance of numerical responses during estimation. These topics are instrumental in shaping the experimental design for the current project, interpreting results from free estimation tasks, and guiding future tests to explore explanations for Benford bias.

The Wisdom of Crowds

Benford bias is a phenomenon that occurs through aggregating the estimates across individuals. A similar aggregation effect that is well known in Psychology and Sociology, the "wisdom of crowds," suggests that accumulating estimates from multiple individuals can lead to more accurate results. This principle posits that data from independent and diverse individuals can offset individual biases, neutralise extreme values, and yield an unbiased average answer (Yi, Steyvers, Lee & Dry, 2012). This originated from a task asking eight hundred people to estimate the weight of a pictured ox and found that the sample median value was almost identical to the actual weight (Galton, 1907). Such an effect was observed and exploited in many areas demanding quality judgment, such as the business environment and political and judicial procedures, where collective performance was usually found superior to personal choice. A more recent approach, the Delphi method, even incorporated the benefit of social learning into this aggregation process by asking people not only to submit anonymous estimates with reasons but also to view and comment on other's entries anonymously (Kahneman, Sibony & Sunstein, 2021). However, the effectiveness of this technique diminishes when faced with cognitive biases like the anchoring effect. The random errors could be eliminated by averaging, but the collective judgment cannot easily compensate for the systemic errors that influence everyone's behaviour. Furthermore, its generalisation ability has been challenged when the correct answer is unknown or even unknowable. Other factors, like response format, can also influence the effectiveness of crowd wisdom, especially in smaller groups. For instance, freely generated estimates can be less effective than scale-based estimates due to tendencies towards round number clustering (Honda, Kagawa & Shirasuna, 2022). Thus, relying on collective judgment for better decision-making might not always be practical.

The Benford bias primarily focuses on the effects of aggregation across multiple estimates in terms of first-digit frequencies. Previous literature has not empirically established a link between estimation performance and Benford bias. However, drawing inspiration from this phenomenon of collective judgment, we were also interested in examining how estimation accuracy in the current project may affect the occurrence of Benford's law when data is combined. This investigation aimed to discern whether and how the precision of predictions plays a role in the manifestation of BL in aggregated data.

Techniques for Individual Estimation

While aggregated data often outperforms individual responses, strategies for individual use remain valuable. After all, individuals bear the consequences of their decisions, necessitating accurate and quality judgments in daily activities. Several methods have been proposed to improve individual estimates.

As Weinstein and Adam (2008) articulated in their book "*Guestimation*," estimation is a skill that can be improved through proper practice and training. Extensive past experiences have been reported to boost associated estimation performance, such as predicting the total time for running 500 meters (Tobin & Grondin, 2015). This is also consistent with the notion of the learning effect observed in psychology, where people frequently tested on their cognitive abilities showed better scores (e.g., Tao, Yang & Liu, 2019). Similarly, more precise estimates can be facilitated by offering feedback on the relevant event experienced by oneself or others (e.g., Roy, Mitten & Christenfeld, 2008). Learning through interaction and experience appears to be a more effective method of acquiring knowledge. For example, behavioural scientists (e.g., Minturn & Reese, 1951) have supported the utility of using combined types of feedback to enhance task performance. Feedback has been demonstrated to impact visual numeracy perception by merely revealing the true value of the first visual items (e.g., number of dots) displayed after participants'

responses, though the anchoring effect was involved in the judgment, the errors of the estimates and the variability between the participants were sharply reduced (Minturn & Reese, 1951). Such feedback effects could last for several months and have shown similar impact on both over and under-estimators. Gasper, Roy and Flowe (2019) further explained these observations as a result of supplying a reference point because people are terrible at producing an absolute judgment but can make reasonable adjustments according to a correct anchor, such as the average value. Explicit feedback was also reported as a good intervention to calibrate the mapping between non-symbolic and symbolic processing when estimating the quantities of visual elements, allowing people to adjust the amount of underestimation (Izard & Dehaene, 2008; Minturn & Reese, 1951).

Previously, we explored existing literature on number estimation following the regularities in nature, and proposed using alternative approaches, such as experiential learning, to re-examine individuals' sensitivity to statistical distributions. However, the initial proposal did not specify a practical method for implementing this learning. Recognising the role of feedback in modifying estimation behaviour, the idea of integrating feedback mechanisms into the existing experimental framework was considered. This integration is designed to immerse participants in situations where they can actively engage with and learn from statistical patterns. Such immersion is crucial for assessing their ability to use the knowledge gained in practical problem-solving contexts. This approach has consequently evolved to become one of the key test paradigms in the current study, reshaping the experimental design.

Apart from providing a frame of proper references, another technique called "unpacking" also attracted the researchers' attention, perhaps due to the prevalent use of Fermi's intervention to solve unknown estimation tasks in science, engineering and education settings. The core process of this strategy is breaking down a numerical estimation question

into relevant subcategories with smaller steps (Kruger & Evans, 2004). For example, in Fermi's classic problem, most people found it challenging to find the answer to the number of piano tuners in Chicago. But when dividing the question into smaller sections by figuring out the estimates to some relevant subcomponents like the population in Chicago, the proportion of families with a piano, the average frequency required for checking up a piano, and the length of the required work, they became more confident in answering the original question. Some studies found that encouraging thinking in terms of the subcomponents of an event, like the different stages involved in preparing a meal, led to less biased estimates of the time interval judged (Kruger & Evans, 2004). Summing up the responses to a set of simplified questions for a time estimation task generally presented a more precise answer because it facilitated the process by considering all the relevant aspects and assigning proper weight (Macan, 1994). This encourages and agrees with the use of System 2 thinking as proposed by Kahneman (2003). Similar conclusions were advocated in other experiments with their successful replications, particularly for the time estimation tasks (e.g., Forsyth & Burt, 2008; Liu, Li & Sun, 2014). In the visual estimation tasks, identical strategies were also promoted as the technique called "grouping" or "groupitising" (Moscoso, Castaldi, Burr, Arrighi & Anobile, 2020). Evidence showed the advantage of making smaller clusters of groups by improving people's sensory precision in the approximate number estimation tasks regardless of the spatial or temporal arrangement (Anobile, Castaldi, Moscoso, Burr & Arrighi, 2020). Given the global ability to precisely subitise a smaller number of items, up to four, on the visual field across individuals, grouping the items into smaller sizes was an effective and effortless approach for easy mental representation and calculation (Ciccione & Dehaene, 2020).

However, there were some controversial results regarding the level of effectiveness among those strategies. For instance, the technique "summing" was detected as a more useful

tool for improving the witness's judgment about the crime duration than a reference demonstration (Gasper, Roy & Flowe, 2019). Attention and mental calculation ability was found to mediate the estimation precision in the groupitising effect (Moscoso, Castaldi, Burr, Arrighi & Anobile, 2020). This calls for more investigation into the generalisation ability of such approaches.

Conclusion on Improving the Quality of Number Estimation

Improving the quality of decisions and judgment is commonly considered and desired as one of the ultimate aims in the practices associated with number estimation. Although the process of estimation seems susceptible to influence due to some cognitive constraints, it is evident that the estimation performance can be enhanced by several techniques, such as averaging the responses from independent and diversified samples, providing feedback as a frame of reference, or unpacking the original estimation task into subcategories for processing. While the current project has no specific intention to boost the estimation performance, some strategies, like the feedback information, offer valuable insights when the present study aims to facilitate the learning process for the test in the experiments. Furthermore, this has potentially shed light on the concepts for the future studies, which will be introduced in the final chapters.

Summary of Chapter 2

The review of cognitive mechanisms involved in generating explicit numerical estimates for decision-making and judgments can be broadly categorised into two main types within the current investigation scope (see Figure 2): numerical answers produced for 1) general knowledge questions and 2) quantities of visual elements. It is worth noting that there are several other dimensions of estimation that have not been explicitly addressed here, such as estimations related to experiencing time-interval, which also exhibit properties similar to

Weber's Law (Gibbon & Church, 1990). Given the existing observations and the framework for examining Benford bias, the present study specifically focuses on the two types of estimations reviewed above. The cognitive processes underlying the generation of such responses may share some common instrumental pathways. For example, the tendency to round numbers has been observed in both types of questions, whether asking for the height of a mountain or the number of dots in a visual display (Honda, Kagawa & Shirasuna, 2022). Thus, it seems the use of preferential representation of certain groups of numbers, such as 50, 100, or 2000, can be generalised across situations.

However, capturing quantities of visual elements across space is primarily processed through the non-symbolic representation of numerosity with the support of ANS mapping to its symbolic representation. Visual processing apparatus, such as the holistic and rapid processing of multiple groups of items, also facilitates the enumeration of a plausible number. Although ongoing debates exist regarding the coordination between different mental representations for numerical cognition, like whether the symbolic processing is deeply rooted in non-symbolic processing, systematic *underestimation* was a prominent feature consistently reported in free estimation tasks where an explicit number was required (e.g., Chesney & Matthews, 2018). This phenomenon agrees with the compressive fashion in which people mentally represent numbers, as known so far.

The process of producing numerical answers to general knowledge questions can be partly attributed to using heuristics and mental shortcuts to navigate unknown situations. On the other hand, such responses also reflect optimal statistical judgment, demonstrating that people are sensitive to the underlying distributions of the estimated variables. While biased answers and optimal behaviour patterns may seem contradictory, they are not conceptually against each other. Biases are typically considered patterns that are "off the target," so it is a measure in terms of precision (Kahneman, Sibony & Sunstein, 2021, p. 8). But, the

knowledge about the distribution of a variable focuses more on the variability, though the measure of central tendency also matters; it is about the shape of dispersion rather than the precision. Thus, it is likely that the estimates were systematically deviated from the true mean under the influence of cognitive biases (e.g., anchoring effect), but the distribution remains as it should be. However, few studies have attempted to explore the interaction between each framework in understanding how an estimate is produced.

Most mechanisms reviewed for the process of number estimation clearly showed that people's ability to estimate is limited by cognitive constraints such as base-rate neglect and logarithm representation of numerals. That may justify why the estimated values sometimes largely deviate from the facts, resulting in under or over-estimation. However, if individual's judgment was such a fragile and flawed process guiding the behaviour and decision-making, human beings would not have survived for thousands of years. Even though people's estimates may not perfectly describe the quantities as they are supposed to be, this efficient and effective nature should not be undervalued. For example, although many non-numerical features might covariate with numerical representation of visual elements, without the support of other perceptual cues, it would be challenging to make an optimal decision (Odic & Starr, 2018), especially for children who are still in the development stage while exploring the world (Van Hoogmoed, Huijsmans & Kroesbergen, 2021). The size of the occupied area might assist them in recognising more or fewer candies possessed before acquiring the exact number. The use of rules of thumb, though prone to errors, could be a good short-term solution under time pressure by minimising the influence of redundant information and simplifying the process (e.g., Bettinger, Long, Oreopoulos & Sanbonmatsu, 2009; Hales, Terblanche, Fowler & Sibbald, 2008).

In the current search for collections of empirical work about how estimate values are produced, we felt like we were collecting puzzle pieces from multiple areas to complete a

picture. Most research, such as studies on the ANS for numerical recognition, included the free generation task as a small part of tests, or sometimes, the estimation judgment only required simple discrimination without expressing an explicit number. Despite its frequent use in practice, a systematic framework of the free production of estimates still stands outside the spotlight and remains underexplored. Nevertheless, this journey sheds light on our path to uncovering why people exhibit Benford bias by providing theoretical guidance and technical support for the current research design. Specifically, the framework that emphasises the statistical acquisition of distributional information holds promise as the primary focus for our investigation. However, other perspectives should not be disregarded, as they offer substantial support for the experimental design, interpretation of results, and future directions for the present project.

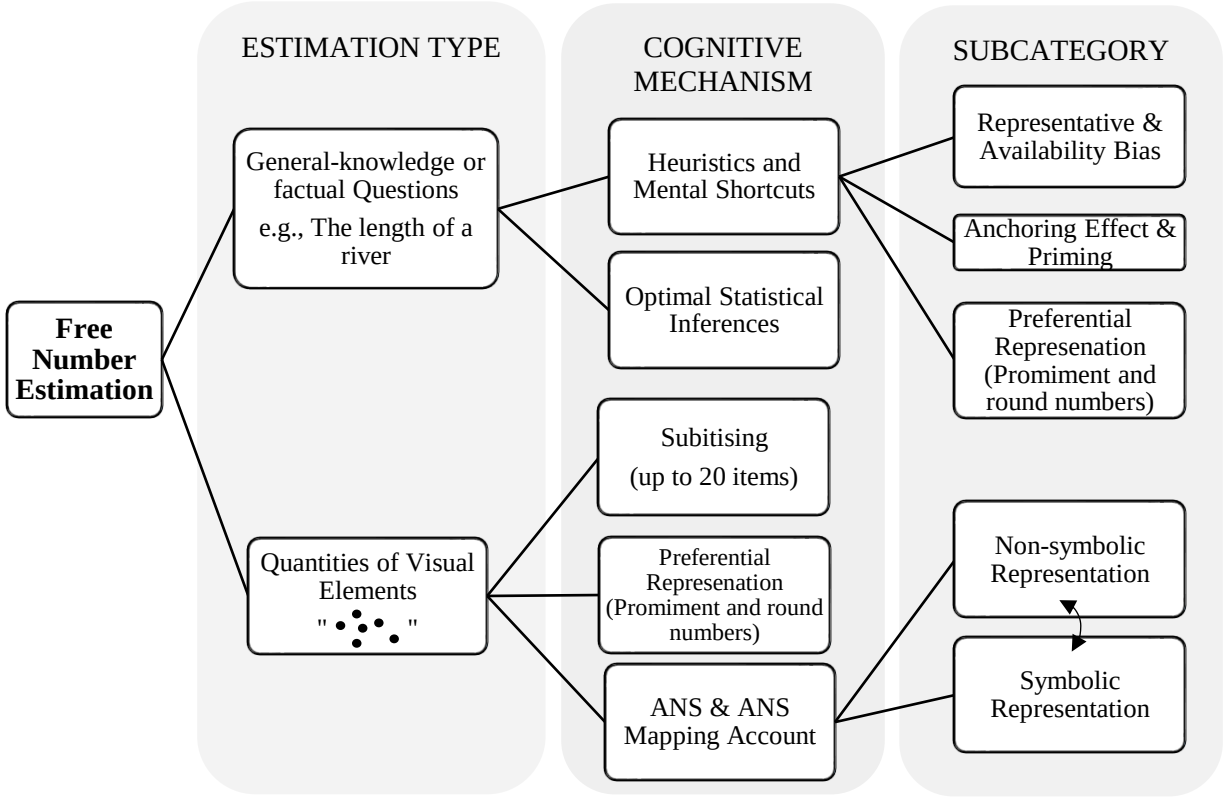


Figure 2: A roadmap illustrating the cognitive mechanisms related to the free number estimation during the review of Chapter 2 for searching the explanations to Benford bias.

CHAPTER 3: CONCEPTUALISATION AND RESEARCH CONSTRUCT OF THE PRESENT PROJECT

The demonstration of Benford bias in Chapter 1 revealed a stronger preference towards the smaller leading digits of numbers, which was exclusively observed when generating meaningful numbers to unknown questions, such as the electricity consumption (e.g., Burns, 2009), and the quantities of jellybeans (Chi & Burns, 2022). Results from the recognition tasks all pointed out that the behaviour of *selecting* yielded no differential first-digit pattern of the answers compared to *generating* a number. Such discrepancies potentially imply that Benford bias could be a product of the process by which people generate responses (Burns, 2009; Chi & Burns, 2022) rather than sensitivity to the first-digit relationship itself, thus challenging the Recognition Hypothesis. Therefore, it is necessary to explore alternative explanations for Benford bias, leading to a review of empirical work on the cognitive process of producing estimates under uncertainty discussed in Chapter 2. Several mechanisms underlying the process of how individuals generate estimates have been systematically investigated. Within this context, the concept of statistical learning emerges as a particularly notable approach. This approach is notable not only because it shares similar testing materials with studies demonstrating Benford bias, but also for its establishment of a pertinent link to the statistical patterns embodied by this bias, considering that both are connected to the nature of occurrences.

This chapter is dedicated to outlining the conceptual framework of the current project, focusing on examining the influence of statistical learning on number estimation. The formulation of research questions, study design, and the hypotheses will be explored. This project uses Benford's law as a tool, a universal rule about the first-digit patterns in naturally occurring settings, to study human responses in number estimation. The study aims to enhance the knowledge of statistical learning by applying it to number estimation tasks,

creating a unique intersection between real-world regularities and experimental manipulations. This also helps in understanding how people generate numbers in decision-making processes.

Development of Research Questions and Objectives

In the search for alternative accounts for explaining Benford bias from number generation, this project also revisited the mathematical explanation of Benford's law found in naturally occurring settings to understand better how statisticians originally justified it. Early studies (Goudsmit & Furry, 1944; Weaver, 1963), which were later refuted by Berger and Hill (2011), argued that BL might be a consequence of the numerical language system employed as a "built-in characteristic." In this system, larger numbers are less frequently employed than smaller ones. This statistical speculation seems consistent with the idea that people represent numbers in a compressive fashion, such as the mental number line and ANS. However, the robust findings of the logarithmic representation of numerals in the cognitive process address a phenomenon about the magnitude of numerosity, while the Benford bias emphasises more about the regularities of the leading digits regardless of the smallness or largeness of a number (Berger & Hill, 2011). Meanwhile, BL has been mathematically validated as being resistant to both changes of scale and base (Hill, 1995), further distancing it from interpretations relying heavily on the base-10 system, such as preferential representation of prominent or round numbers (Converse & Dennis, 2018). Berger and Hill (2011) questioned whether BL could be explained by a simple explanation. Later, Berger and Hill (2015; 2020) refined and reported three basic facts about sequences of random variables converging BL: 1) powers of every continuous variable; 2) products of random samples from every continuous distribution; 3) combined samples of randomly selected sample from random distributions (in an unbiased way). The first fact, which highlights that BL comes

into play when values of selected variables follow a power law, was of particular interest to us. This is because it seamlessly translates to the study of Benford bias in psychology. This mathematical observation essentially outlines that the bias favouring smaller leading digits results from accumulating values of variables that agree with a power law distribution.

Assuming that individuals can produce estimates consistent with a power law distribution, Benford bias should emerge as anticipated. This conjecture happens to support one observation proposed under the framework of statistical learning introduced in Chapter 2. Those studies showing optimal statistical inferences in relation to people's cognitive judgments in everyday prediction (e.g., Griffith & Tenenbaum, 2006; Lewandowsky, Griffiths & Kalish, 2009) demonstrated that people are aware that data come from different distributions, and their responses can follow the appropriate prior distribution depending on the context. For instance, when estimating a person's total lifetime, the responses were consistent with the use of the appropriate Gaussian prior, while predictions of movie gross tended to follow a power law prior. Considering the widespread presence of the power law characteristic in numerous variables around us, the frequent emergence of Benford bias should not be a surprise.

In contrast to the Recognition Hypothesis, an alternative explanation, the Distribution Hypothesis, has been developed, speculating that sensitivity to environmental regularities lies in the underlying distribution of variables rather than the occurrence of the first digits themselves. This hypothesis was rooted in the findings of optimal judgment in everyday prediction, where individuals are sensitive to the underlying distributions of variables and capable of utilising such knowledge in problem-solving. In addition, mathematical evidence supports that datasets adhering to power laws tend to follow BL (Berger & Hill, 2015; 2020). This implies that data derived from other distributions, such as Gaussian variables, should not conform to BL. Taken together, the Distribution Hypothesis posits that the Benford bias

arises only from generating values of underlying distributions that follow a power law. Thus, supposing people can learn the distribution of the variables, this makes a testable prediction that varying what people think the distribution of estimates will affect the degree of Benford bias.

On the other hand, the Recognition Hypothesis concerning sensitivity to first-digit distributions remains compelling despite the lack of support from selection tasks. This is because it resonates well with substantial evidence demonstrating the impact of acquiring statistical relationships from the environment on decision-making processes. Thus, we decided to test its explanatory power again by pushing it to the limit. Instead of directly posting the estimation questions, we returned to the traditional paradigm of learning, such as exposure. Engaging in interactive learning and immediate experience seems to be a more effective way of acquiring knowledge, as people continuously explore their surroundings to integrate new information into their existing knowledge (Russo, Roy, Kazerouni, Osband & Zheng, 2017), as demonstrated in Chapter 2. For example, behavioural scientists have demonstrated the usefulness of employing various types of feedback to enhance task performance (e.g., Roy et al., 2008), such as improving precision and minimising the variation of visual perception of quantities (Minturn & Reese, 1951). Moreover, the absence of observed replication of environmental regularities by individuals should not automatically lead to the conclusion that they lack knowledge or ability because this could also be a consequence of choice, similar to the behaviour found between “probability matching” and “maximising” (Plate, Shutts, Green & Pollak, 2018). Therefore, in light of the compelling evidence from statistical learning and the limitations of prior studies, I decided to run another attempt in this project to test whether people are sensitive and capable of learning first-digit frequencies through interactive engagement with the relevant information.

In summary, drawing from existing literature and findings, the current research aimed to systematically explore and understand the reasons behind the Benford bias. This was achieved by investigating two potential explanations within the framework of statistical learning: 1) the Benford bias as a consequence of sensitivity to the underlying distribution of variables that follow a power law, or 2) the bias as a result of learning the first-digit distribution. The Benford bias is a phenomenon observed primarily in the context of number estimation and is thus generally considered a product of cognitive processes when generating unknown numbers. These hypotheses are not mutually exclusive; instead, they provide distinct testable avenues for understanding behaviour in number estimation.

General Methods, Designs and Materials

General Methods and Construct

To pursue the research questions proposed in this project, the experimental paradigm where Benford bias emerged in number estimation was followed. The direct manipulation of the independent variables with random assignment under controlled conditions enables us to closely observe the changes in outcomes, hence allowing for the establishment of causal relationships (Chambliss & Schutt, 2006). Thus, sticking with the experimental approach was the optimal choice to explore the underlying causes of Benford bias.

In this project, I used the terms “estimates” or “estimation” to distinguish between numbers generated as labels or symbols for items like passwords and order numbers, and numbers generated for decision-making and judgment. The APA Dictionary of Psychology defines an “estimate” as “the best guess of the value of a parameter of a distribution based on a set of empirical observations,” which aligns with the scope of the current research. While random number generators might assign numbers arbitrarily and could even replace them with other symbols like letters for identification purposes, estimates are different because

they carry quantitative meaning. Changing a single digit in an estimate can lead to different judgments and outcomes.

One of the key constructs embedded in the previous investigation of what causes Benford bias was the attempt to influence the degree of fit and understand which process could lead to a stronger bias. However, this proved to be a technical challenge, as most baseline data already showed a good fit to the distribution proposed by BL. For example, one of the studies sought to increase the amount of visual information present in the picture to see if this could achieve a better fit. However, the findings revealed that showing a dynamic picture of jellybeans did not boost a stronger Benford bias than the use of a static picture (Chi & Burns, 2022), both showing substantial agreement. Theoretically, a stronger Benford bias shall encourage the behaviour of producing smaller leading digits such as 1, 2 and 3, but over-representation can also introduce deviations. For instance, a result showing 37% of generated numbers starting with the digit-1 might suggest a pronounced bias towards smaller leading digits, which agrees with the underlying message of Benford bias. But this also significantly diverges from the expected 30.1% frequency in the BL model from the point of statistical analysis. Therefore, the current project shifted its focus from seeking a potential catalyst for enhancing the Benford bias to exploring the methods to disrupt the Benford bias. The feasibility of interrupting the statistical pattern originally acquired was discussed in Chapter 1, such as repeatedly experiencing the new regularities, as well as in Chapter 2 when presenting the benefit of interactive learning, like the use of feedback. In that regard, the present research specifically investigated the conditions under which this well-established distribution regarding the first digit can be altered under experimental manipulations in light of previous demonstrations.

General Designs

Building upon the previous paradigm of experiments examining Benford bias, the key elements in the free estimation tasks involved a simple description of short questions in English, coupled with the presentation of visual elements. To distinguish it from other paradigms studying estimation judgment, such as placement tasks that pose limits on magnitude, the estimation in the current project specifically refers to the process where participants could freely generate the digits of a number (e.g., Chesney & Matthews, 2018) using Arabic number system (e.g., integers) for an estimated answer. Generating outputs without apparent boundaries is consistent with the assumptions for where BL is typically found in naturally occurring settings. The use of Arabic numbers as outputs simplifies the process of producing explicit numerical answers and mitigates potential linguistic differences in number words that may arise due to cultural diversity (e.g., Savelkoul, Williams & Barth, 2020) within the sample. The primary dependent variable in the current study was the proportions of nine leading digits, which were then compared against the theoretical frequencies posited by BL.

Four experiments were conducted online, using web pages controlled by a JavaScript program. Participants were first or second-year undergraduates from the University of Sydney recruited either during the tutorials as a part of their assignment work or via SONA, an internal research recruitment platform in the School of Psychology. The online experiment was designed to be completed within a short timeframe, either 15 or 30 minutes. Students who participated through SONA received partial course credit. The targeted sample size was calculated using G*Power to achieve a small effect size with a power of 0.8 by the specific design of each experiment. The majority of tasks employed a 3 * 2 mixed or 2 * 2 between-subject design, thus the minimum required sample size from the calculations is 120 or 128, respectively.

General Materials and Tasks

To examine the reliability of the hypotheses proposed to justify Benford bias, the existing experiments maintained the basic structure of estimation tasks used in prior studies (e.g., Chi & Burns, 2022; Chi, 2020) while incorporating innovative materials and techniques into the present project for the current testing purposes. Three estimation tasks as the primary paradigm were developed for subsequent experiments.

Animal Estimation. The Distribution Hypothesis proposed that Benford bias could be a consequence of producing values that follow the power law of the variable estimated due to sensitivity to its underlying distribution. To test this, it was essential to make a contrast by offering the variables characterised with different shapes of distributions to see if there is an ability to discriminate between them and produce corresponding responses. Since this project aimed to understand how people generate numerical answers under uncertainty, the estimated variables were carefully selected to avoid relying on very common knowledge or tight boundaries. Various options were considered and scrutinised, such as the mass of planets and items (e.g., telephone waiting time) from the studies showing optimal statistical inferences. Eventually, my focus was narrowed down to the attributes of animals, as they possessed the desired qualities of the study that could be experimentally manipulated once and for all.

Statistically, it was a well-captured phenomenon that the weight of animals typically follows a normal distribution (e.g., Uchmański, 1983), their lifespan tends to be left-skewed (Muradian, 1989), and their group-size mostly agrees with a power law (e.g., Griesser, Ma, Webber, Bowger & Sumpter, 2011). Thus, by estimating these three variables for a particular animal, I could collect responses with different shapes of underlying distributions without changing the topic. This design facilitated a concise and efficient online experiment.

Dots Displays. To assess the reliability of the explanation that Benford bias is a product of people's sensitivity to the first-digit distribution, the design developed in Chi's

(2020) MPhil program where a robust Benford bias emerged, but with an update by including feedback after each response. The benefit of using dot arrays was frequently established by numerous studies when understanding the cognitive and perception processing of numerosity, at least for small numbers. The paradigm of presenting dots was simple and straightforward to perceive during the test and re-test sessions. However, considering the influence of physical features on the judgment of numerosity reviewed in Chapter 2, the colour, size and illuminance of the dots were controlled. As the density and the area occupied are inherently associated with the number of elements (e.g., Morgan, 2013), some attempts have been made to test these effects on the first-digit pattern, but no significant difference has been reported (Chi, 2020). Given this, the dots display designed in the current task consistently adopted the same presentation configuration by covering the entire display area with randomly located dots so that the density varies with the number of dots for each display. I generated random dots using a scatterplot in Microsoft Excel, with a certain number of coordinates produced by the formula "=RANDBETWEEN(1,10000)." To avoid the process of counting and subitising, as well as to simulate uncertain situations, I showed at least hundreds of items briefly, spanning just a few seconds.

The interactive learning process involved here allowed us to explore if Benford bias could be disrupted by experiencing feedback on performance during estimation, hence a stronger test for examining people's sensitivity to the first-digit distribution compared to the passive acquisition. Such a disruptive process was expected to be achieved through repeated exposure to the feedback given the findings that the shift of original acquisition of statistical pattern was dependent on the frequencies of reoccurrence of a new structure (Gebhart, Aslin & Newport, 2009; Weiss, Gerfen & Mitchel, 2009). Hence, the repeated measure under the influence of feedback data shall provide an ideal paradigm for us to run tests and re-test sessions. As improving accuracy was not the primary focus of the study and to avoid the

anchoring effect, only vague feedback on task performance was provided rather than the correct numbers. As pointed out by Weinstein and Adams (2008) in their book chapters, the estimates provided in response to a question typically fall into one of three "Goldilocks" groups: 1) too large, 2) too small, or 3) just about right. I applied this categorisation to create vague feedback prompts to offer respondents additional hints about the correct answer.

Video Estimation. Past literature on estimation judgment tasks predominantly used a number of dots below 300, often not exceeding 100. The task of Dots Displays in the current project restricted the items to hundreds of visual targets up to 1050. Thus, any effect detected could be limited by the magnitude used as the first digit (except 1) was perfectly correlated with the amount of dots. To disentangle the impact of size on visual estimation, we decided to develop a follow-up task to extend our observation beyond hundreds of elements, expanding the testing boundaries to tens of thousands. This task was primarily exploited in the last stage to provide extra observation in addition to the findings from the Dots Displays.

The animation was considered an ideal solution to manage the display of thousands of targets in a scene, as it allowed for conveying information in depth, which optimised the displayed space for abundant objects. To vary the size of the magnitude, three size groups were selected: three-digit length (hundreds), four-digit length (thousands) and five-digit length (ten-thousands) values. To minimise the potential impact of the display view on the perceived number of targets, different angles of viewing (e.g., top, side, and insider view) options were chosen to be included in this presentation. The estimated targets were displayed dynamically in a 10-13 second video clip. The inclusion of depth and movement of a visual object allowed us to mimic the real-world scenario, as well as to extend the scope of current tests to align with the recent updates of the numerosity representations regarding depth and dynamic features (e.g., Aida, Kusano, Shimono & Tam, 2015; Au & Watananbe, 2013).

The visual estimation of quantities in the Dots Displays tasks was confined to a single target within a static display. To provide contrast and investigate estimation patterns across a range of domains, I crafted a series of video clips. Each clip displays a unique visual element engaged in densely populated, dynamic movements. The graphic of these targets was simplified and produced with common scenarios such as people gathering (e.g., at a stadium), flying birds, and swimming whales. The duplication of targets was not a simple copy and paste; instead, scenarios were simulated to represent real-life situations where objects were in motion at a given time. The game engine development platform Unity 2021.3.8f1 was used to simulate these scenes and bake dynamic objects in an animation model. Meanwhile, dots remained as one of the target domains, dynamically displayed by rapidly updating their coordinates in scatterplots. The coordinators were generated using the random number generator formula in Excel. Examples of video clips are illustrated in Experiment 4.

The video estimation was considered an extension and follow-up task of the Dots Displays. Thus, the feedback was also available for the estimated answers to see how it interacts with the process of number estimation, particularly the first-digit pattern. By varying the magnitude of the quantities, it offered a more comprehensive understanding of the relationship between the visual numerosity and Benford bias, thereby expanding the scope of the existing observations.

Major Hypotheses

This project was designed to assess the explanatory power of plausible explanations in justifying the occurrence of Benford bias observed in number estimation tasks. We were particularly interested in testing the impact of statistical regularities, hence examining individuals' sensitivity to two distinct types of patterns: the underlying distribution of the

estimated variable and the first-digit distribution of true values. I formed the general hypothesis for each account to test its reliability, given the current study design.

Drawing upon the optimal statistical inferences in everyday prediction (Griffith & Tenenbaum, 2006) and one of the mathematical facts regarding Benford's law (Berger & Hill, 2015; 2020), Benford bias could be perceived as a consequence of producing variables following the power law due to sensitivity to such underlying distributions. Hence, as proposed by the Distribution hypothesis, if individuals are capable of acquiring knowledge regarding how a variable is distributed, it was expected to obtain Benford bias only when generating the answers in response to questions characterised by power law, but not from other patterns such as Gaussian variables.

Alternatively, given the well-established foundation of acquiring statistical regularities in the environment (Saffran, Johnson, Aslin & Newport, 1999; Fiser & Aslin, 2002), if people are sensitive to the first-digit pattern, I expect to see that providing vague feedback repeatedly regarding the estimation performance shall suppress Benford bias when the first digits of the true values disagree with it. Through repeated exposure and interaction with additional data about the correct distribution, individuals are expected to exhibit a pattern closer to the truth, thereby deviating from Benford bias.

I anticipate that no single explanation could fully account for all observations, meaning that both tests could be supported or rejected. However, each hypothesis was assessed independently, offering a unique perspective in demonstrating the potential impact of statistical learning.

A Review of Statistical Analysis

In this project, data was organised using Microsoft Excel and analysed with SPSS 25. The study required not only the calculation of the proportions of the first digits of the

estimates produced, but also an analysis of their overall degree of goodness of fit to Benford's law. In this section, several common statistical approaches used to test the goodness of fit to a theoretical model are discussed, along with their respective benefits and limitations. Finally, I outlined the specific method selected for this analysis.

When considering the methods for comparing the observed model to theoretical ones, Pearson's chi-squared test (χ^2) often comes to mind as the first option. Indeed, this test is a popular choice in many studies that seek to assess the degree of agreement with BL. However, the goodness-of-fit test employed in this context has limited statistical power due to its eight degrees of freedom (nine leading digits) in the test.

Moreover, the chi-squared test mainly draws conclusions based on the failure to reject the null hypothesis. As a result, any research conducted with a sufficiently large dataset should always show some deviation from BL. In that regard, having a smaller sample often results in a nonsignificant difference, leading to the acceptance of the null hypothesis (Geyer & Williamson, 2004). Studies that reported a reasonable fit, such as Diekmann (2007), generally involved relatively few respondents. Therefore, unless there is a substantial deviation from BL, such studies would consistently support conformity.

In the review of papers related to fraud detection using accounting and economics data, an alternative examination often used is the z-test, introduced by Nigrini (1996). The z-statistic indicates whether a digit's proportion deviates from its expected value. Significant deviations are identified in the tails of the distribution, occurring when the z-score exceeds 1.96, which corresponds to a 95% confidence level. However, this test can only analyse one leading digit at a time. The z-test has faced two major concerns in fraud detection regarding its use (Durtschi, Hillson & Pacini, 2004). The first concern argued that the z-test might not be very sensitive to the significant difference when dealing with a few fraudulent transactions among piles of datasets. Secondly, while testing one digit at a time using z-statistics allows

for a detailed examination, it may need more insight when drawing conclusions based on the overall pattern of FSDs. In other words, it may not provide a comprehensive understanding of the extent to which the observed distribution agrees with the expected one.

Another statistical approach for assessing the degree of conformity between observations and the theoretical model was derived from the Bayesian probability theory (Geyer & Williamson, 2004). This method tests the fit between two proposed patterns: the proportions of the nine leading digits, represented as the vector θ, are compared to Benford's distribution, θ_0, assuming the null hypothesis of equivalence. Although this approach remains a viable option, further research is required to fully understand its practical application in the present studies.

The above approaches introduced may be a good tool in assessing the degree of fitness between the data observed and Benford's proportions. However, this study design additionally requires an analysis to compare the goodness of fit among the conditions where none of those could offer a solution.

In recent studies regarding the psychological investigation of Benford bias (e.g., Burns, 2009; Chi, 2020), an alternative measure called the root-mean-square error (*RMSE*) was introduced to assess the conformity of observed responses to the expected proportions for each leading digit. The *RMSE* is computed as the square root of the sum of squared deviations of the differences between the observed and expected proportions, which has been extensively employed in many domains of research area. For example, economists employ it to assess how well an economic model corresponds with specific indicators (e.g., Armstrong & Fred, 1992) or to compare the forecasting errors of different models for a particular dataset (e.g., Hyndman & Koehler, 2006).

In examining the degree of fit to Benford's law, the *RMSE* relative to the frequencies of BL (*RMSE-Benford*) can be easily computed for each individual. As the *RMSE* measures

errors, a smaller value indicates a better fit. Nevertheless, contrasting with another RMSE value is typically desired to draw meaningful statistical inferences. Based on the studies reporting Benford bias, the initial digits of the correct nine answers selected were usually uniformly distributed, so a flat distribution was adopted as a baseline model for a contrast. The *RMSE-Benford* is then compared against the equal proportions of the flat distribution (*RMSE-Flat*) to see whether the observed data was closer to the true distribution or Benford's frequencies. If *RMSE-Benford* was smaller than *RMSE-Flat*, it implies that the first-digit proportions obtained agreed more with the pattern of BL than the uniform distribution. Transforming the nine estimates into a single value provided a solution to contrast how close an individual is to the Benford or baseline models. Moreover, the mean of *RMSE-Benford* in one condition can be further discriminated against another to tell which group achieved a closer agreement to BL, further demonstrating its utility for multiple testing purposes. Nevertheless, such computation relies on the relative comparison between the two models. Thus, it was concerned with its limited generalisation ability when the true distribution is unknown or even unknowable.

During the 2020 Australia Mathematical Psychology Conference, Burns presented a novel approach for quantifying Benford bias by computing the effect size. This was achieved by calculating the eta-squared (η^2) for the digit liner contrast weighted according to the proportions specified by BL. A contrast is a weighted sum of parameters or statistics, representing a specific type of linear combination. The coefficients (a, b, c, etc.) in the linear combination are chosen to reflect the particular comparison of interest.

Previously, standard methods for testing goodness of fit, such as the chi-squared (χ^2) test, provided a way to determine whether the datasets conform to a hypothetical model or not. However, as researchers, we were more interested in quantifying the degree of fit rather than just obtaining a binary answer of yes or no. Having an η^2 measures how much variances

can be explained by Benford's proportions, with a larger size indicating a stronger fit to the expected model. Thus, it allows us to evaluate the extent to which the experimental results align with desired data. This method has gained support from past research involving visual estimation. For instance, a study by Chi and Burns (2022) found that Benford's proportions accounted for a significant portion of the variance (58%) when estimating the number of jellybeans in static pictures, thereby demonstrating a strong Benford bias. Moreover, this contrast analysis can also be employed to assess the impact of manipulated variables on the extent of bias. For example, one could compare the η^2 values derived from estimates in the static picture group versus those in the dynamic condition to examine the picture effect on the first-digit pattern. Given these advantages, the current project utilised this analytical approach, incorporating the calculation of eta-squared (η^2) for the digit linear contrast, weighted by the proportions of BL.

Summary of Chapter 3

The current research aims to explore the potential explanations for the Benford bias observed in number estimation by investigating how this well-established phenomenon can be altered or even whether it can be changed at all under the typical experimental paradigm of statistical learning. This offers a valuable opportunity to bridge the gap between observations based on real-world regularities and those derived from artificial laboratory settings. Much of the existing literature on statistical learning was primarily derived from controlled laboratory experiments involving artificial patterns or schemes created by scientists (de Finetti, 1974). Such acquisition was commonly achieved or updated following short-term exposure to the stimuli. Comparatively, little research in psychology has explored how the probability distributions portrayed the real-world scenarios were progressively acquired and integrated with new information over time. This gap may exist due to the

inherent challenge of uncovering and establishing a valid connection between universal models and specific human judgment. Consequently, the degree to which test materials created in the lab resemble real-world problems remains a point of contention. However, the establishment of Benford bias offered us a chance to introduce and observe the correspondence between a prevalent rule regarding the first-digit patterns of natural events and human reactions in number estimation. Supposing researchers know that the participants come to the tasks with behaviour patterns that match a universal rule (i.e., Benford's law), then any subsequent changes must be attributed to the alterations they have made in the process of undertaking such tasks. By investigating the Benford bias and the related cognitive process, the current research aimed to enrich the framework of statistical learning, extending its application to the process of number estimation tasks. The inclusion of BL as a tool additionally offered a unique intersection between real-world regularities and short-term experimental manipulations. In practice, it provided us with a better understanding of how individuals generate numbers for use in decision-making.

CHAPTER 4: EXPERIMENT 1

To examine the reliability of the possible explanations proposed to justify Benford bias, the first experiment consisted of two tasks: Animal Estimation and Dots Displays. The questions asked in the Animals Estimation task primarily investigated the responses to the general knowledge questions. Participants were asked to estimate three types of variables for the animals regarding their weight (that is normally distributed), lifespan (which is left-skewed), and group-size (following a power law). To facilitate learning, similar to what has been implemented in the standard paradigm of statistical learning studies, an exposure stage was conducted prior to the testing phase. During this stage, participants were acquainted with statistical data illustrating the distribution of each variable using human-related examples. As proposed by the distribution hypothesis, if individuals are sensitive to the underlying distribution of a variable, it was expected to observe Benford bias when generating the answers in response to questions characterised by a power law distribution such as group-size, but not from those Gaussian variables like weight.

The second task, Dots Displays, was designed to investigate individuals' sensitivity to the distribution of leading digits and to determine if exposure to feedback could influence their ability to learn and adapt to the new first-digit pattern in numerical estimates of visual elements. Previous studies (Chi & Burns, 2022; Chi, 2020) have consistently found a strong preference for smaller leading digits in quantity estimations with visual elements. If individuals indeed possess this sensitivity, it was anticipated that providing feedback across successive trials might disrupt the existing Benford bias, particularly when the first-digit frequencies of true values disagree with Benford’s law. Moreover, I expect to see the first-digit pattern shift towards the true distribution progressively when learning from the true values through feedback.

In general, Experiment 1 comprised the Animal Estimation and Dots Displays tasks. The Animal Estimation task was a within-subject design that manipulated the underlying distributions of the estimated variable. Participants were asked to estimate three types of variables: weight (normal distribution), lifespan (left-skewed distribution), and group-size (power law distribution). The Dots Displays task was a repeated measures design that consistently provided feedback on numerical answers across two rounds. The dependent variables measured the frequencies or proportions of the first digits of the estimated values.

Methods of Experiment 1

Participants

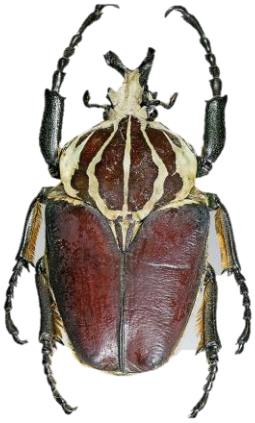
185 first-year psychology students at the University of Sydney were recruited from SONA. The responses of 183 were finally included in the analysis after data cleaning, with an average age of 19.58 (*SD* = 2.63). Of these, 65.6% identified as female and 34.4% as male, predominantly English (53.8%) and Chinese (36.3%) speakers as their native language.

Procedure and Materials

The Animal Estimation task included 27 estimation items, while the Dots Displays task comprised 18 items. The sequence of the two tasks was randomised for each participant.

Animal Estimation. This task involved two stages. During the exposure phase, participants were introduced to visual examples representing three different statistical distributions (see Appendix A), accompanied by thorough explanations to ensure comprehension of distribution concepts. These examples utilised human-related data, illustrating normal distribution through human birth weight, left-skewed distribution via female life expectancy, and power law distribution through the size of WhatsApp groups (see Appendix B). In the subsequent testing phase, participants were individually presented with images of nine animals and asked to estimate their weight, lifespan, and group-size. The

selection of the animals aimed to encompass a diverse array of wild social animals across varied species and habitats. This was designed to ensure a broad spectrum of estimated values by selecting animals with significantly different sizes and lifespans. The animals presented were Capybara, Atlantic Herring, Cape Buffalo, Red-winged Blackbird, Purple Sea Urchin, Plains Garter Snake, Fiddler Crab, Goliath Beetle, and Greater Flamingo (see Appendix C). These images were randomised in the order of presentation. An illustrative example of an estimation item is provided in Figure 3.



It is a picture of a Goliath beetle.
Please give an estimate to the questions below:

The weight of this Goliath beetle is _____(g).
The lifespan of this Goliath beetle is _____(days)
How many Goliath beetles are usually grouped in the wild? _____

Figure 3. This is an example of questions asked to estimate the weight, the lifespan, and the group-size of a Goliath beetle in the task of Animal Estimation in Exp.1. The units were kept consistent across items; that is, gram(s) used for the questions asking for the weight, and day(s) used for the questions asking for the lifespan.

Dots Displays. This task consisted of two rounds, each showing the same set of nine pictures of dots displayed in random order. The task required participants to estimate the quantity of dots presented, with true values of 250, 350, 450, 550, 650, 750, 850, 950, and 1050 (see Appendix D). The duration for which each image was displayed varied randomly between 1500, 2500, 3500, 4500, 5500, 6500, 7500, 8500, and 9500 milliseconds. Upon making their estimation and submitting a response, participants were immediately provided with feedback indicating whether their guess was “too low,” “too high,” or “very close” to

the actual number of dots. An estimate was classified as “very close” if it fell within a 10% margin of the true value, a criterion cited from similar datasets utilised in the Dots Displays task from Chi’s (2020) MPhil thesis. Additionally, following each guess, participants were asked to rate their confidence in their estimate on a 5-point scale ranging from “Strongly not confident” to “Strongly confident.” It was unknown whether the feedback provided in the first round would immediately change the Benford bias. Hence, participants again experienced the same set of nine items in a different random order in Round 2 without being explicitly informed that the testing materials repeated themselves. An example of the Dots Displays is presented in Figure 4.

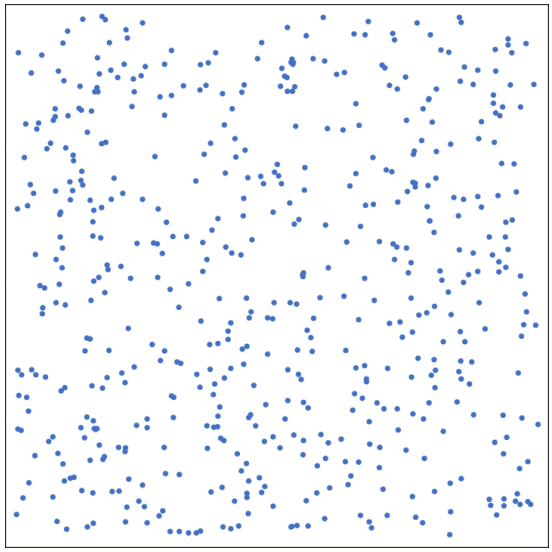


Figure 4. The sample of dots displays contains 550 dots. As the results in Chi’s (2020) study suggested, the dots presentation arrangement did not significantly affect the degree of fit to BL. Hence, the dots presentation chosen in this project kept the display area fully occupied by the dots while varying the dots’ density when the quantity increased.

Questionnaires and general instructions. The study incorporated a questionnaire to gather participants’ demographic data, that is, gender, age, first language, and enrolled degree program (see Appendix E). Alongside this, general instructions were provided to ensure optimal technical setup and to emphasise the importance of individual effort in responding to questions without external help (see Appendix F). A demonstration for the Dots Displays was supplied before the start of the task questions (see Appendix G).

Catch pages. A Catch page was incorporated within the general instructions to ensure participants paid close attention to the guidelines and tasks. Unlike other sections where a "Continue" button triggers the next page, here, participants were explicitly directed to click on the page's title to proceed (see Appendix H). Should they not perform this specific action, a pop-up reminder would prompt them to go through the instructions again carefully. The frequency of such errors was tracked and factored into the data cleaning stage, serving as an indicator of the participants' engagement and comprehension.

Results of Experiment 1

The outputs of Animal Estimation and Dots Displays tasks were discussed separately. This investigation primarily focused on the pattern of Benford bias in number estimation. Therefore, for each task, the findings consistently presented the distributions of the leading digits produced under different experimental conditions, followed by an examination of its degree of fit to BL. Additional analyses, such as the test of response accuracies, confidence, and cultural differences, were also performed when necessary.

Results for Animal Estimation

Benford bias in Animal Estimation. The proportions of the first digits for the estimated variables of weight, lifespan, and group-size were graphed in Figures 5, 6 and 7, respectively. A reasonable monotonic decline of the leading digits of these estimates can be perceived with a few spikes. A significant interaction between the first-digit pattern and the variables questioned was reported by the within-subject contrast analysis weighted by Benford's proportions, suggesting that these patterns were different from each other, $F(1, 182) = 33.131$, $p < .001$, $\eta^2 = .154$. Some unexpected deviations were observed in the estimates from lifespan questions (see Figure 6), such as the elevation at digit-3 ($M = .205$, $SD = .154$) and digit-7 ($M = .085$, $SD = .102$). Further investigation suggested that more than

one-third of such responses were generated corresponding to the exact counts of days in one year (i.e., 365), two years (i.e., 730), or twenty years (i.e., 7300). This might be a consequence of directly providing the unit of "days" in the question asked, indicating the influence of the unit system on estimations. As Durtschi, Hillson, and Pacini proposed with their observation of natural data (2004), deviation from Benford's proportions is common when tailored for specific purposes. In general, digit-1 was the most frequent number produced, and a strong peak was found at digit-5 when generating weight estimates (M = .166, SD = .141) and group-size estimates (M = .136, SD = .118). The Benford-Newcomb distribution (BND) accounted for a substantial portion of the variance in these estimates: 74.4% in weight, 66.8% in lifespan, and 79.1% in group-size. Despite some minor fluctuations, these findings suggested a good fit to Benford's distribution. Thus, it failed to show that changing the variable estimated could dramatically shift the Benford bias.

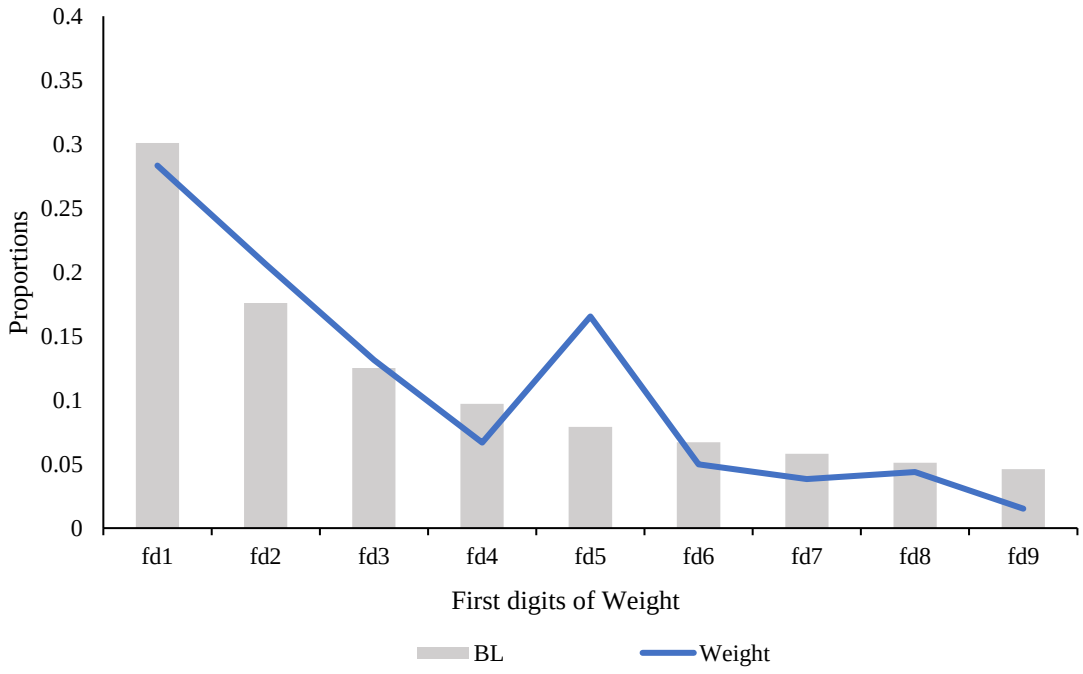


Figure 5. The first-digit proportions of the estimates in response to the weight of animals questioned in the Animal Estimation, compared against Benford's proportions in Exp.1.

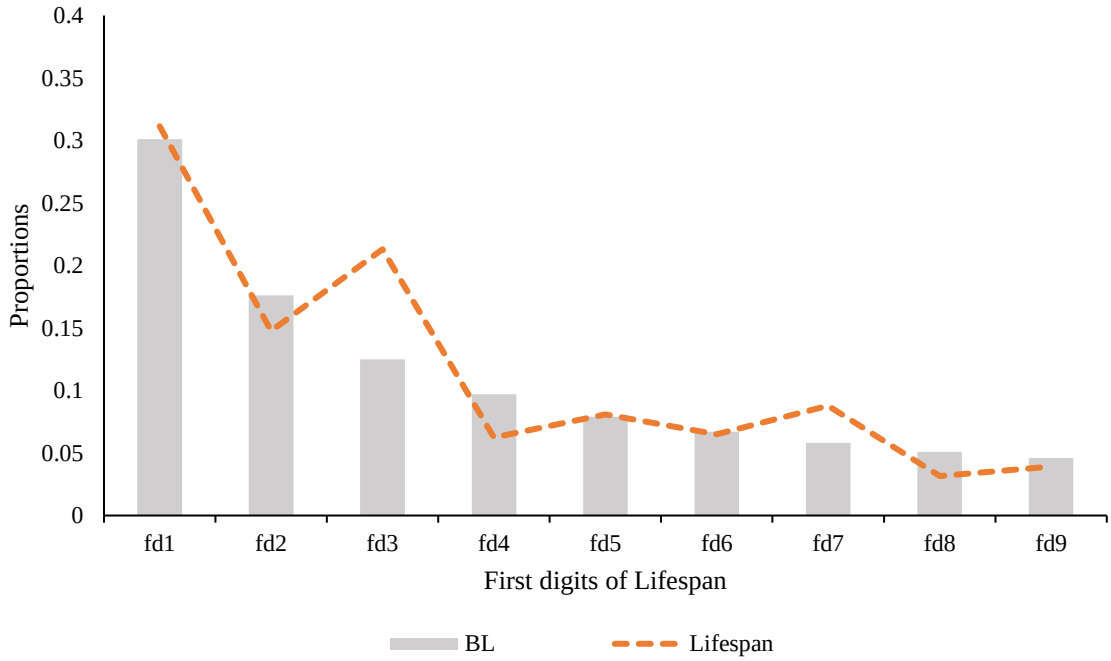


Figure 6. The first-digit proportions of the estimates in response to the lifespan of animals questioned in Animal Estimation, compared against Benford's proportions in Exp.1.

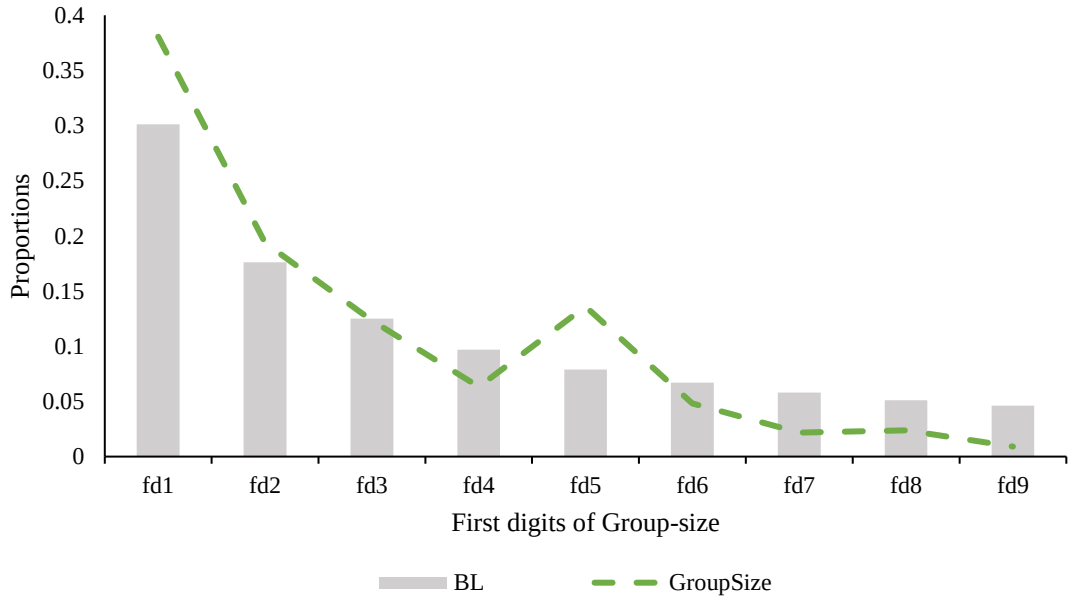


Figure 7. The first-digit proportions of the estimates in response to the group-size of animals questioned in Animal Estimation, compared against Benford's proportions in Exp.1.

Distributions of full-number. In assessing the Benford bias across different variables, I found no significant variation in based on the type of variable questioned. This led to an analysis of whether the estimates mirrored the actual distributions of the variables inquired about. As each individual generated three responses (i.e., weight, lifespan, and group-size) for nine different animals, a total of twenty-seven distributions were analysed in the normality test. Unfortunately, none of those patterns was normally distributed, as reported by the Shaprio-Wilk test (all $p < .001$). All distributions were extremely positively skewed, with skewness ranging from 1.655 to 13.086 for the weight, 6.254 to 13.528 for the lifespan, and 9.408 to 13.526 for the group-size questions. Notably, the weight distributions sometimes exhibited multiple spikes. However, the skewness of the full-number of weights, lifespan and group-size was an effect of aggregation across individuals; thus, we shall not be able to definitely conclude that at the individual level, people lacked the ability to follow the normal distribution where the weight was supposed to be. Moreover, practically, the analysis at the individual level could not be measured based on the current design, as each person only generated one estimate for each question for one specific animal.

Other factors. Further analysis was conducted to explore potential differences in the patterns attributable to the specific animals involved. This analysis confirmed that the general trend of a monotonic decline in leading digits was consistent across all nine animals, though some experienced minor fluctuations at certain digits. Additionally, factors such as gender (male versus female) and language (English versus Chinese) were investigated for their influence on Benford bias, but no significant effects were reported by the repeated measures ANOVA (both $p > .05$). The task did not include an analysis of the accuracy of estimates since the actual average weight, lifespan, and group-size for animals can vary widely due to several factors, such as habitat conditions and breeding season, making it impractical to compare against a single true value directly.

Results for Dots Displays

First-digit distributions in Dots Displays. The proportions of the first digits produced in response to the Dots Displays over two rounds were presented separately in Figure 8. Digit-1 remained the most frequently generated number when the patterns appeared much closer to the uniform(true) distribution. The test of within-subject contrasts weighted by the proportions of BL reported a significant interaction between the first digits and two rounds, indicating that the two patterns were different from each other, $F(1, 182) = 36.433$, $p < .001$, $\eta^2 = .167$. The BND explained 22.3% of the variances of the answers in Round 1, showing a poor fit, and even became weaker in Round 2, as only 6.2% can be explained. A small peak at digit-5 was only found in the first round.

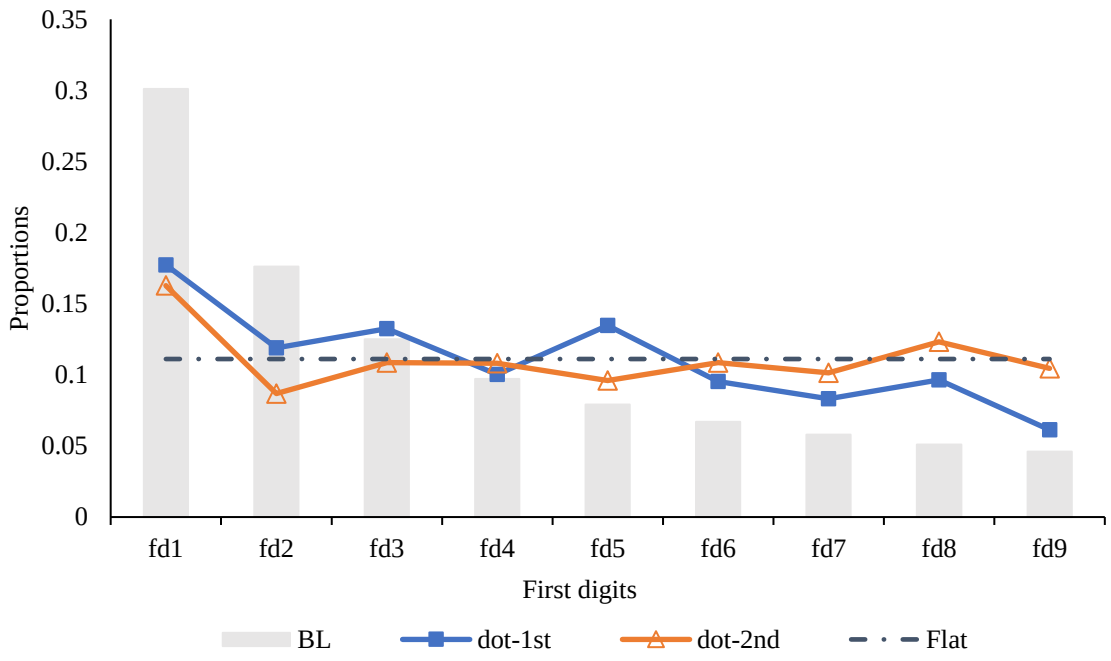


Figure 8. The first-digit proportions of the estimates in response to the Dots Displays in rounds 1 and 2, compared against Benford's law and flat(true) distribution in Exp.1.

Benford bias emerged from the first-item responses. Surprisingly, the first-digit patterns in the two rounds did not exhibit a reasonable Benford bias when accumulating the responses across all the items. This unexpected result was likely due to the immediate application of feedback that may have prompted a quick adjustment in the participants' responses. To investigate this, the analysis was restructured according to the *order* of items, focusing on whether the Benford bias could re-emerge from the responses to the very first item where feedback had not yet influenced the participants. Individuals experienced nine trials in random order within each round and only produced one estimate for each trial; hence, this analysis adopted the chi-squared test to examine the degree of fit. To present a clear pattern of the transition in each round, one graph was plotted for the first four trials, while a separate graph displays the pattern from trials 5 to 9. As shown in Figure 9, the proportions of the first digits accumulated from the first trial in Round 1 substantially followed the BND, $\chi^2(8) = 13.482$, $p = .096$. However, the patterns from trial 2 began to show a deviation ($p < .05$), progressively approaching the uniform distribution in later trials, as shown in Figure 10. In Round 2 (see Figures 11 and 12), regardless of the item order, all the frequency patterns failed to show a monotonic decline, with most patterns uniformly distributed based on the chi-squared test ($p > .05$).

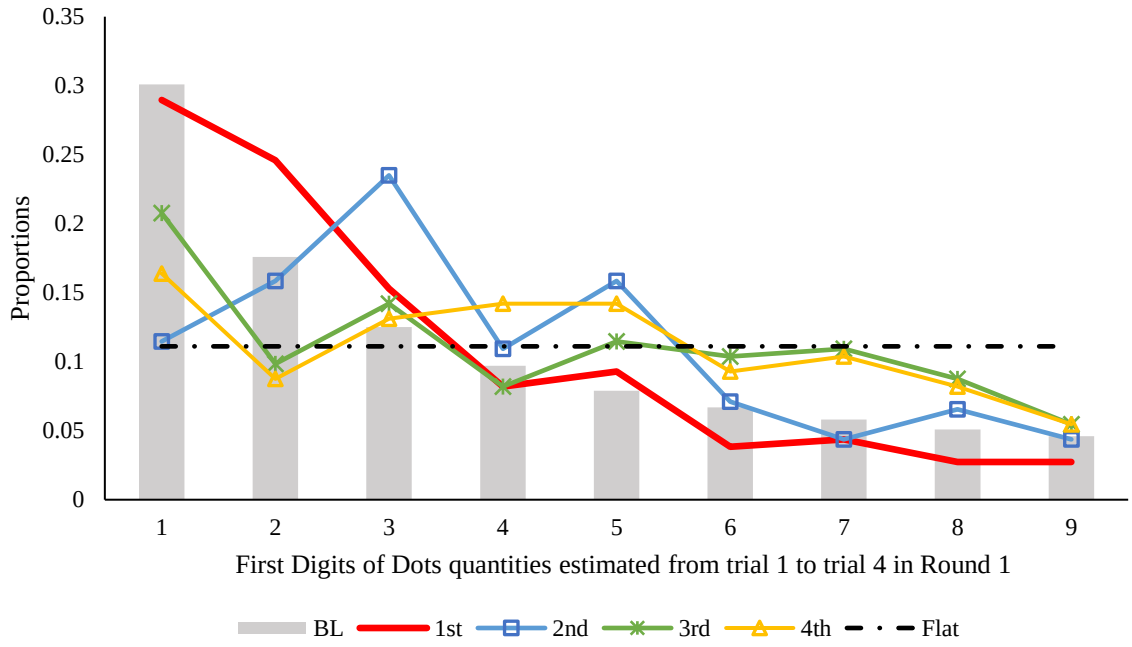


Figure 9. The first-digit proportions of the estimates to the Dots Displays in Round 1 were arranged by the order of the items tested from trial 1 to trial 4, compared against Benford's law and flat(true) distribution in Exp.1. The first trial was graphed with a thick line for better contrast.

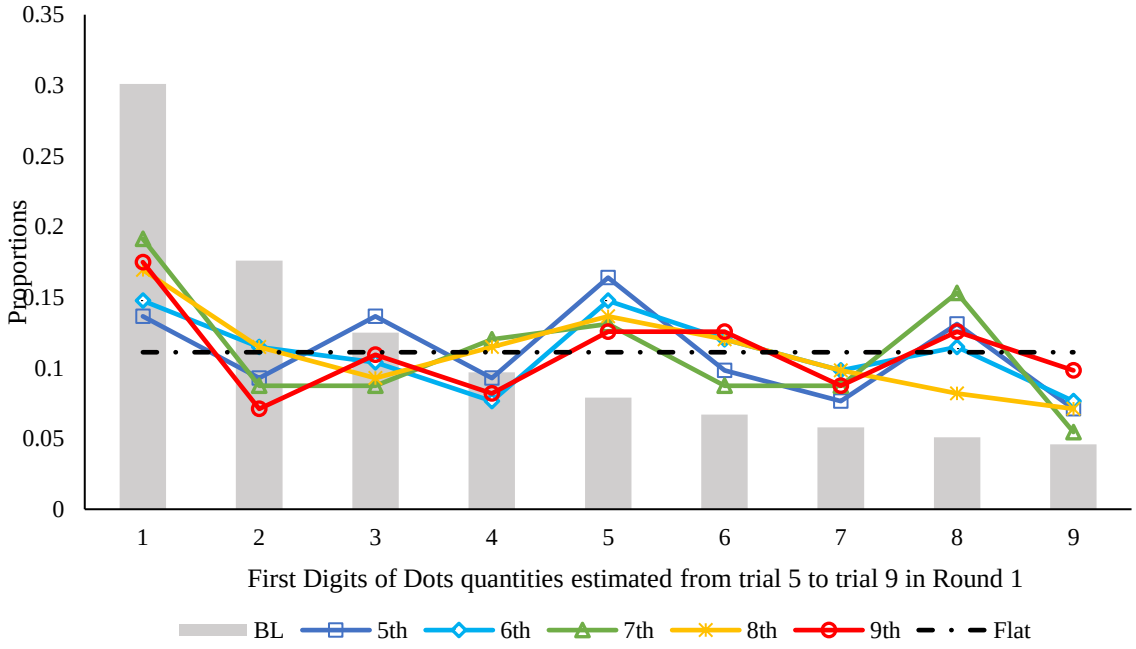


Figure 10. The first-digit proportions of the estimates to the Dots Display in Round 1, arranged by the order of the items tested from trial 5 to trial 9, compared against Benford's law and flat(true) distribution in Exp.1.

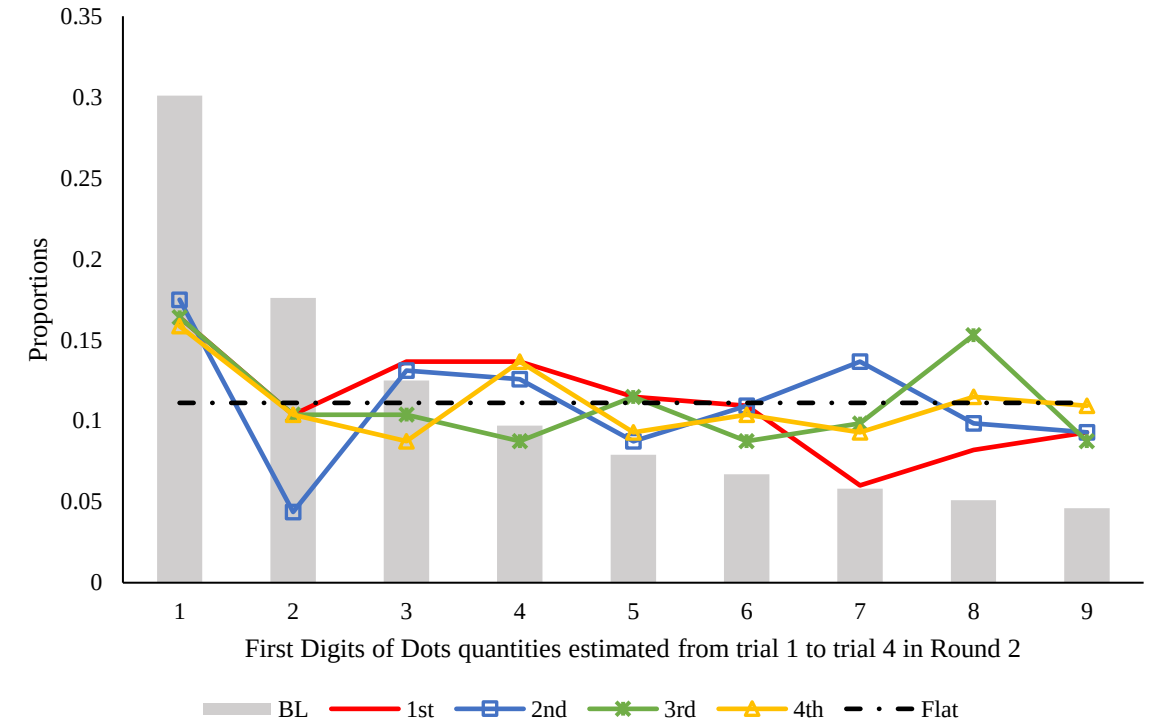


Figure 11. The first-digit proportions of the estimates to the Dots Displays in Round 2 were arranged by the order of the items tested from trial 1 to trial 4, compared against Benford's law and flat(true) distribution in Exp.1.

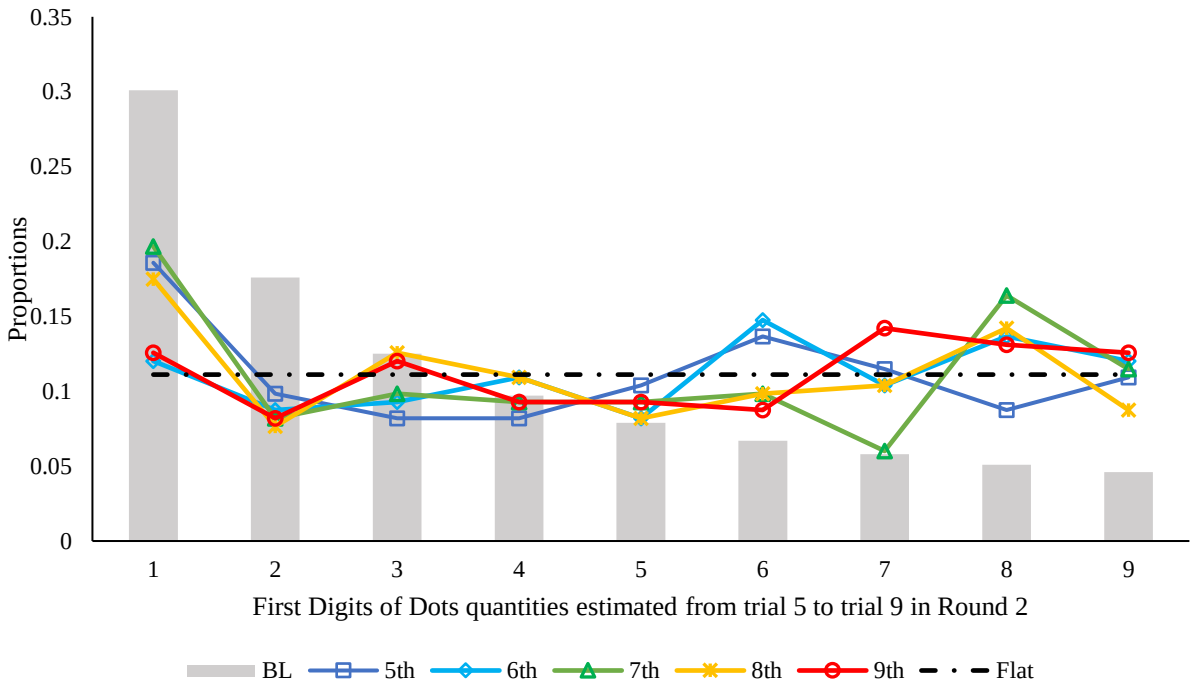


Figure 12. The first-digit proportions of the estimates to the Dots Display in Round 2, arranged by the order of the items tested from trial 5 to trial 9, compared against Benford's law and flat(true) distribution in Exp.1.

Accuracy analysis of full-number. Experiencing feedback seemed to strongly suppress the Benford bias, thereby evening out the first-digit distribution. To understand such changes, it was essential to see if the vague feedback influenced the estimation performance. Perceiving large numerosities is generally supported by the Approximate Number System (ANS) and the mapping account, which is evidenced by a tendency towards underestimation in performance. Thus, I first calculated and presented the descriptive data of the mean estimates of nine pictures for rounds 1 and 2 in Figure 13. Overall, the estimation of the items with 250, 350, 450, 550, 650, 750, and 850 dots was accurate under the influence of feedback, with no significant deviation from true values across both rounds based on the t-test (all $p > .05$). But a substantial underestimation was reported in both rounds when showing larger numerosities like 950 and 1050 ($p < .05$), despite supplying the feedback.

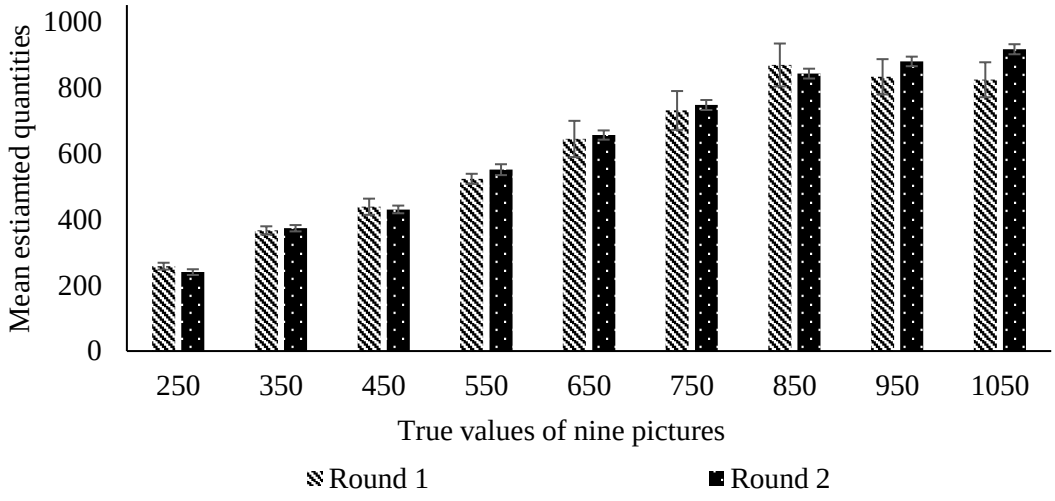


Figure 13. The mean estimated responses to nine pictures in Dots Displays across two rounds in Exp.1.

The first part presented the descriptive data; however, to examine the learning effect on estimation performance between two rounds, two types of accuracy analysis were conducted regarding the size of errors. First, I evaluated the accuracy of the full-number of estimated values, measured by the absolute differences (*AD*s) between the estimates generated and the true values, with lower *AD*s indicating better performance. Non-parametric

tests found that *ADs* in Round 2 were significantly smaller than those in Round 1 (all $p < .05$), except for the item with a true value of 250 dots, as shown in Figure 14.

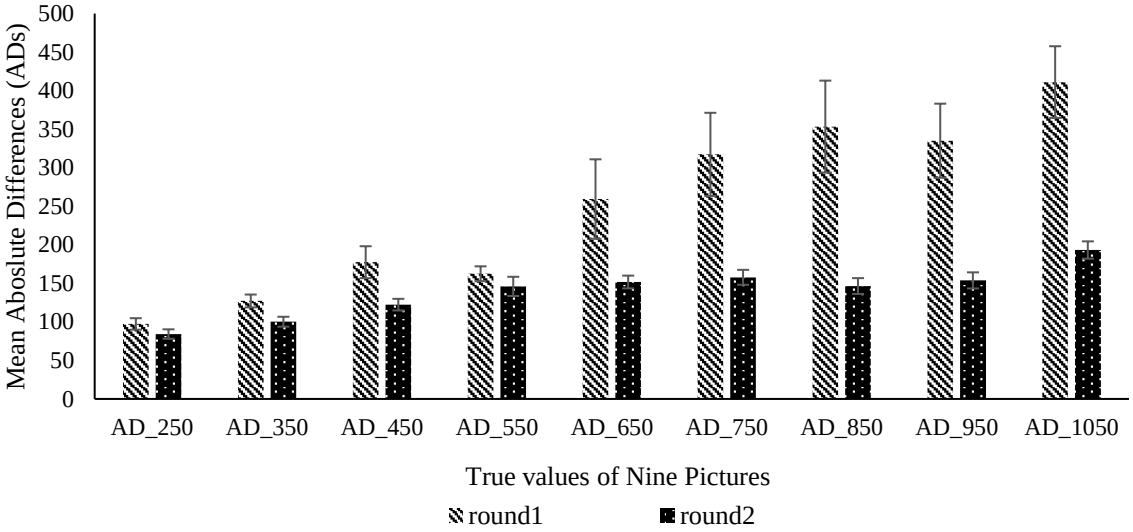


Figure 14. The mean absolute differences (*ADs*) between the estimates and true values for nine Dots Displays over two rounds in Exp.1.

Accuracy analysis of first-digit estimates. In addition, I assessed the accuracy rate regarding the *first-digit* of the estimated values in each round. This was calculated by aggregating the number of correct first digits estimated and dividing by nine, the total number of test items in each round. There was a significant improvement in the first-digit accuracy rate from Round 1 ($M = .214$, $SD = .139$) to Round 2 ($M = .289$, $SD = .176$), $t(182) = 4.998$, $p < .001$. A positive association was also detected between the full-number accuracy measured in Round 1 and Round 2 ($r = .954$, $p < .05$), as well as for the first-digit accuracy ($r = .209$, $p < .05$). In general, providing vague feedback appeared to consistently result in better estimation performance, although a control group without feedback was not supplied in this study.

Confidence and other factors. On average, participants exhibited stronger confidence in their answers in Round 2 ($M = 2.802$, $SD = .725$) than those in Round 1 **(***M* = 2.740, $SD = .702$), $t(182) = 2.131$, $p = .034$. The confidence ratings of these two rounds were

positively correlated with each other, $r = .849$, $p < .005$. Interstingly, higher confidence ratings were associated with higher *ADs* (worse estimation performance) in Round 1, $r = .168$, $p < .05$.

The repeated measure ANOVA found a difference between English and Chinese speakers regarding the first-digit pattern of the estimates generated in Round 2, $F(8,1296) = 2.145$, $p = .029$, $\eta^2 = .013$. However, given this very weak effect size, it only accounted for a small proportion of the variances. Upon visual examination of their distributions, apart from some fluctuations at certain digits, the general pattern was similar between the two groups. No gender differences were found in this experiment.

Discussion of Experiment 1

The Animal Estimation task aimed to examine whether estimating Gaussian variables such as weight would deviate from Benford's law, but the contrast analysis overall demonstrated consistent Benford bias regardless of the variable questioned. The interaction effect between the first-digit pattern and the variables indicated that people might produce specific biases tied to specific domains of questions during estimation. For instance, there were noticeable spikes at digits three and seven in responses to lifespan-related questions, while the elevation of digit-5 was reported in weight and group-size estimates. Despite these variations, the overall monotonic decline characteristic of the Benford bias was distinctly maintained. Contrary to the distribution hypothesis, I did not find evidence that responding to a variable with a Gaussian distribution suppressed the Benford bias. Following this, we also checked the distributions of full-number estimates to see whether people were indeed producing data following its actual pattern. Unlike the results typically observed from optimal statistical inferences in everyday prediction (e.g., Griffith & Tenenbaum, 2006; Lewandowsky, Griffiths & Kalish, 2009), most responses to the weight, lifespan, and group-

size questions were all positively skewed with a large variance. The similar shape of distributions might result from a lack of sufficient knowledge about how each dataset should be spread, which questions the effectiveness of the learning process during the exposure stage. The following attempt could benefit from a manipulation check to verify adequate comprehension of the distribution materials initially presented. Additionally, several peaks were sometimes detected in the distribution of weight estimates, suggesting that the data were likely generated from multiple distributions, possibly due to differences in free parameters. Thus, it seems essential to update the design in the next experiment by providing specific parameters as a reference point (e.g., the average weight).

Consistent with previous studies (e.g., Burns & Krygier, 2014; Chi, 2020), the Benford bias emerged when generating the initial estimate of the dots displays without any additional information. As expected, providing feedback on the estimation performance, even if vague, disrupted the monotonic decline of the first-digit pattern, leading to a distribution that more closely resembled the uniform distribution of a true scenario. Surprisingly, this change was rapid and robust, even captured in the responses from the first round. By the second round, the distribution of the first digits was nearly uniform, indicating a significant impact of feedback. Such changes could be potentially attributed to the improvement of accuracy as having feedback substantially improved estimation performance in both the *full-number* answer and the *first-digit* of an answer. However, a simple disruption of the Benford bias could be alternatively explained as a consequence of the changes in the cognitive process of how estimates were generated (Burns, 2009; Chi & Burns, 2022). The following attempts may benefit from including a control block without feedback to serve as a baseline, which could support the observation of a smoother transition caused by feedback. Moreover, in this experimental design, the correct distribution pattern for the first digits was intentionally set to be uniform. My next attempt shall consider experimenting with different actual proportions

for the first digits. Such variations could provide further insights into the nature of the disruption and help to clarify and distinguish between potential explanations.

The substantial influence introduced by feedback on the Benford bias pointed out that people were indeed sensitive to this extra information and capable of reacting to it by updating the existing behaviour pattern, which is a promising finding. However, the failure to capture the effect by estimating variables following different underlying distributions highlighted that certain constraints could limit the acquisition or employment of the statistical information.

CHAPTER 5: EXPERIMENT 2

The revised design of the second experiment incorporated critical enhancements to address the limitations observed in the first experiment. Two major updates were made for the Animal Estimation task. First, a quiz module consisting of six multiple-choice items was implemented prior to the estimation questions to assess participants' basic understanding of three exemplar distributions. In my first attempt, it was argued that participants' failure to follow the estimated variable distribution could result from insufficient knowledge gained during the exposure stage. Thus, to facilitate effective learning, a quiz session was supplied. Participants had to re-visit the materials and repeat the quiz if any of their answers were incorrect. This process continued until they answered all questions correctly, with a maximum limit of five attempts. The quiz was structured to match a baseline level of comprehension, predicated on the assumption that recognising basic concepts, such as the fact that weight is normally distributed and age is typically skewed, would be sufficient for students to provide accurate responses.

In addition, two types of data, namely average (e.g., the mean weight of the Megabat) or juvenile information (e.g., the current age of a young sea otter) of an animal, were provided as reference points in the estimation questions. In the initial experiment, I noted that the distribution of participants' estimates sometimes displayed multiple spikes on a larger scale, indicative of behaviour influenced by free parameters. Introducing reference points was expected to help regulate this large variation. Statistical measures of central tendency, such as means and medians, were considered as they are very commonly introduced in statistical contexts and curriculums. Nevertheless, understanding the concept of median data has been reported to be more challenging for students (Bakker, 2004; Cobb, McClain & Gravemeijer, 2003; Konold & Higgins, 2003; Zawojewski & Shaughnessy, 2000). Since estimating a reasonable number was commonly associated with the proper activation of executive

functioning (Strauss, Sherman & Spreen, 2006), conceptual understanding played a more important role in facilitating this process than computational skills, given the current testing purposes. Therefore, showing the average data of a variable estimated seemed to be a better option than the median data. Meanwhile, the position of the average data in the exemplar distributions was also tested in the quiz when introducing how each variable is supposed to be distributed, making it a suitable reference point. However, a potential concern with using means as a reference point was the possibility of individuals producing a tight range by making minor adjustments from the average value. To counter this, four images depicting animals of various sizes and age groups were prepared for each specific animal, with one randomly assigned to each participant. This approach aimed to vary the magnitude of responses and eliminate potential picture effects.

An alternative reference point, the juvenile data, was proposed to facilitate an additional comparison with the commonly used average information. As previously noted, a concern with employing average data as a reference point was the potential for participants to replicate the provided mean in their responses. However, presenting juvenile data of a starting point value was expected to encourage deviations from the references based on assumptions about the nature of growth. The use of a starting point as a reference was also inspired by the study design from Griffith and Tenenbaum (2006), which demonstrated people’s sensitivity to the underlying distribution of a variable. In their test, participants were asked to predict the data of everyday events drawn from different distributions based on an existing point, such as predicting a man’s lifespan based on his age of 65. With these updates for the Animal Estimation task, I still expected to see that Benford bias would only emerge from estimates following the power law but not from those Gaussian variables if participants are able to acquire the regularities of those underlying distributions.

In the Dots Displays from the first experiment, a disrupted Benford bias was immediately detected after several trials given the feedback information in the first round. Thus, this did not allow us to observe a baseline condition with the progress of transitioning. Therefore, this new experiment was revised by adding an additional block to the original two rounds. The first block served as a control condition without feedback, creating a baseline pattern for comparison with the subsequent two rounds showing feedback. Previously, the first digits of true values were carefully selected to follow a uniform distribution across nine leading digits. This, however, posed a challenge in distinguishing between two potential causes of the suppressed Benford bias observed: was it due to learning the true first-digit distribution, or was it a fundamental change in the cognitive process of numerical estimation? To clarify this and expand my observations, the correct first-digit pattern was altered in three distinct categories: a uniform distribution (uniform condition), a pattern following the logarithm nature of Benford law (Benford-like condition), or a pattern exhibiting a preference toward larger first-digit contrary to Benford law (anti-Benford condition). Through consistent interaction with the feedback, if participants are responsive to the actual first-digit distribution, then the initial Benford bias should be disrupted and subsequently shift towards the true distributions, especially in cases where the true patterns disagree with Benford's law. Conversely, if the Benford bias merely changed in a similar manner, irrespective of the actual first-digit pattern, it is more plausible that this is due to alterations in the cognitive process of number estimation, such as changes in strategies.

In general, Experiment 2 included the two tasks used in the first experiment. The task of Animal Estimation was a 3 (variables: weight vs. lifespan vs. group-size) × 2 (references: average vs. juvenile) mixed design. The task of Dots Displays also had a 3 (true distribution: uniform vs. Benford-like vs. anti-Benford) × 2 (feedback: control vs. feedback) mixed design.

Methods of Experiment 2

Participants

A total of 242 responses of the first-year psychology students were received from SONA, and 236 were included in the analysis after data screening. The average age was 20.17 (*SD* = 2.335). Most participants identified as female (65.7%), while males accounted for about 33.1%, and the remaining 1.27% identified as other gender categories. The two most common native languages were English (46.6%) and Chinese (38.1%).

Procedure and Materials

Each participant completed 27 estimation items in both the Animal Estimation and Dots Displays tasks. Although these two tasks were not directly related to each other, the order was randomised.

Animal Estimation. This task followed the structure of Experiment 1, asking participants to estimate three variables (weight, lifespan, and group-size) of animals presented in pictures. However, in this case, participants were presented with a question containing a reference point related to either the average information (see Appendix I) or the juvenile data (see Appendix J). In the average reference group, four distinct images were prepared for each animal, each depicting the animal at different sizes and ages. Among these, one image (labelled as A, B, C, or D) was randomly assigned to a participant for estimation (see Figure 15). Those in the juvenile reference group, however, were asked to predict the animal's mature weight, total lifespan, and final group-size after gathering, based on the starting point given for each variable. An example is shown in Figure 16. Across the task, participants were presented with nine species: sea otters, Atlantic herrings, cape buffalo, red-winged blackbirds, bottlenose dolphins, plains garter snakes, Atlantic horseshoe crabs, Megabats, and greater flamingos. These wild animals were carefully selected so that the first

digits of those reference data for each variable were uniformly distributed to ensure each leading digit occurred once. This also minimised the influence of the anchoring effect.

Similar to the procedures in Experiment 1, participants were initially given information about three example distributions of human weight, lifespan, and group-size. This was followed by a manipulation check, which comprised six quiz questions to ensure a basic level of understanding (see Appendix K). The first three questions required participants to match each variable with the appropriate distribution shape from a list of three options graphed. The subsequent three questions tasked participants with selecting a reasonable average value positioned from three options for each variable, as displayed on their respective distribution graphs. If any responses were incorrect, participants were instructed to review the learning materials and retake the quiz. This could be done up to five times to ensure comprehension before moving forward.

Here is a series of pictures of a Sea Otter.
Please generate the estimates to Picture A only:

If the average weight of the Sea Otter is 25,000 g, please estimate the weight for this Sea Otter. _____ (g)

If the average lifespan of the Sea Otter is 4,500 days, please estimate the lifespan for this Sea Otter. _____(days)

If the average group size of the Sea Otters in the wild is 250, please estimate the group size for this Sea Otter. _____

Figure 15. This is an example of questions including the average information regarding the sea otters' weight, lifespan, and group-size in the task of animal estimation, Exp.2. However, a potential concern with using means as a reference point was the possibility of individuals producing a limited range by making minor adjustments from the average value. To counter this, four images depicting animals of various sizes and age groups were prepared for each specific animal, with one randomly assigned to each participant. This approach aimed to vary the magnitude of responses and eliminate potential picture effects. In this case, picture A was allocated to make an estimate. The units were kept consistent across items; that is, gram(s) was used for the questions asking about weight, and day(s) was used for the questions about lifespan.

It is a picture of a juvenile Megabat.
Please give an estimate to the questions below:

If the weight of this young Megabat is 750 g, please estimate its weight once it matures._____ (g)

If this young Megabat has been living for 4,500 days, please estimate its total lifespan. _____ (days)

If 8,500 Megabats were observed at a point of time during their gathering in the wild, please estimate the final group size. _____

Figure 16. This is an example of questions including the juvenile information regarding a young megabat's weight, lifespan, and group-size in the task of animal estimation in Exp.2. The units were kept consistent across items; that is, gram(s) were used for the questions asking for the weight, and day(s) was used for the questions asking for the lifespan.

Dots Displays. Similar to Experiment 1, the same set of nine dots pictures was repeated in random order over three rounds. Feedback on the estimated values was provided in the second and third rounds, with the first round acting as a baseline test without feedback. The display time setup and the method of providing feedback remained consistent with the first experiment. Participants were randomly assigned to one of three groups, each with a distinct set of true values for the dots displays:

1) The Uniform Group (see Appendix D): This group encountered nine dot images with true values randomly selected from a set including 250, 350, 450, 550, 650, 750, 850, 950, and 1050. This ensured each leading digit appeared once, providing an equal chance of occurrence;

2) The Benford-like Group (see Appendix L): In this group, the dot images represented true values of 220, 280, 350, 450, 550, 650, 1030, 1060, and 1090 with a frequency pattern where smaller leading digits appeared more often than larger ones. That is, the leading digit -1 occurs about 33.3%, followed by digit-2 at 22.2%, closely matching Benford's Law.

3) The Anti-Benford Group (see Appendix M): This set of dot images were arranged to have a majority of true values with larger leading digits, contradicting the typical pattern of BL. The true values ranged in a random sequence from 450, 550, 650, 750, 820, 880, 930, 960, to 990. The leading digit-9 appears approximately 33.3% of the time, followed by digit-8 at 22.2%, and so on, which is the complete inverse of the distribution predicted by Benford's law.

The questionnaires, general instructions, catch pages and confidence ratings. Materials were the same as those developed for Experiment 1, with minor textual modifications to accommodate the revised design of Experiment 2.

Results of Experiment 2

Results for Dots Displays

Among 236 students, eighty-one received the dots displays that followed the uniform distribution, seventy-seven were assigned to the Benford-like group, and seventy-eight experienced the anti-Benford condition.

I found three-way interaction between the first digits, rounds and distribution groups, as indicated by the within-contrast analysis weighted by the proportions of BL, $F(2,232) = 10.169$, $p < .001$, $\eta^2 = .081$. This suggested the impact of distribution groups on the first-digit pattern across rounds. My analysis began with examining the first-digit pattern in the initial round, and then proceeded to discuss it in the following rounds for the uniform, Benford-like, and anti-Benford groups, respectively.

Benford bias emerged in Round 1. The proportions of the leading digits of the estimates generated for the three groups of true distributions in Round 1 were plotted

separately in Figure 17. It was important to note that no feedback was supplied in the first round. Based on the contrast analysis, the Benford-Newcomb distribution (BND) explained 65.8% of the variances in the group with a uniform distribution and 68.9% of the variances in the Benford-like condition. For the anti-Benford group, 37.8% of the variances were explained by the BND. This amount of variances explained was small, primarily due to an unexpectedly low frequency of digit-1 generated; nonetheless, a general preference over the smaller leading digits remained consistent with the notion of Benford bias. I considered this a fluke, especially since subsequent experiments yielded a good fit. Overall, in the absence of feedback, the first-digit distribution in Round 1 demonstrated a reasonable Benford bias, independent of the true value distributions presented to participants. A small elevation at digit-5 was only observed in the Benford-like group.

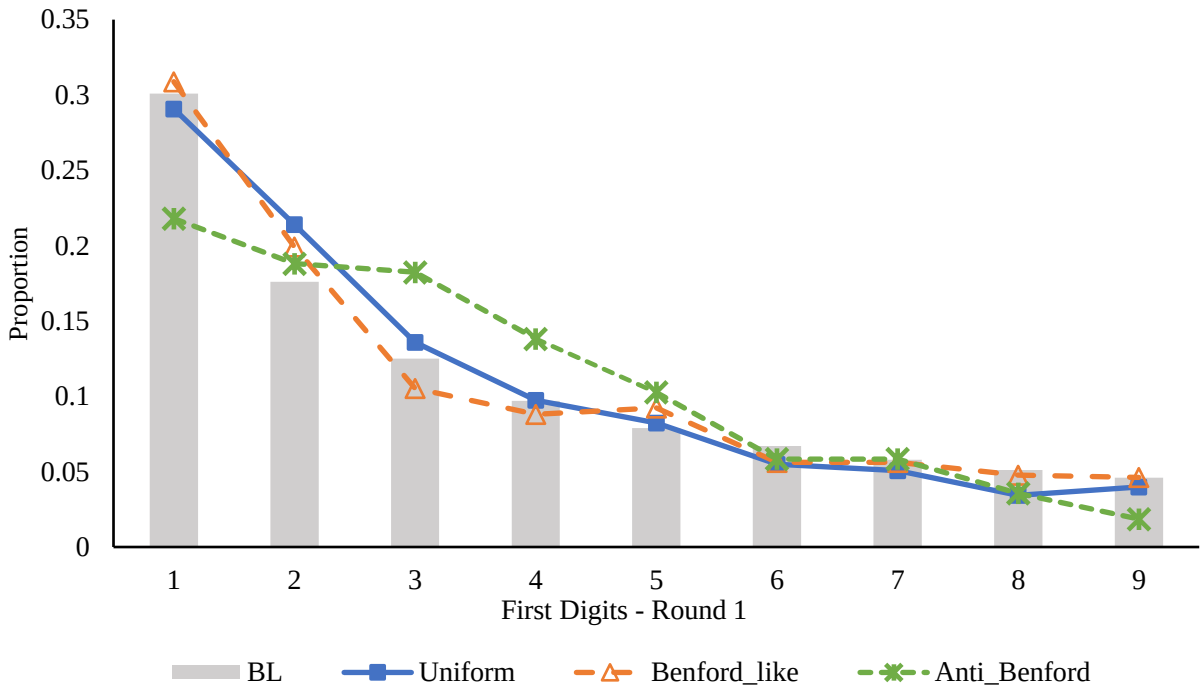


Figure 17. The first-digit proportions of the estimates to the Dots Display in Round 1 when no feedback was supplied for the uniform, Benford-like and Anti-Benford groups, compared against Benford's law in Exp.2.

The suppressed Benford bias in the uniform condition. A robust Benford bias was replicated in Round 1, as expected. This prompted a subsequent analysis to assess the influence of feedback provided in Rounds 2 and 3, contrasting it with the baseline dataset. The first-digit proportions of three rounds for each true distribution group were plotted in separate graphs. For the uniform group where the first-digit pattern of nine true values was evenly distributed, the within-subject contrast analysis weighted by the proportions of BL reported an interaction between the leading digits and three rounds, $F(1,80) = 60.405$, $p < .001$, $\eta^2 = .43$, As illustrated in Figure 18, the frequencies of smaller first digits became noticeably suppressed in response to feedback in the latter two rounds. The contrast analysis indicated a substantial decrease in the explanatory power of BND, from 65.8% in Round 1 to 28.1% in Round 2 and further down to 14.2% in Round 3, demonstrating a progressive impact of feedback.

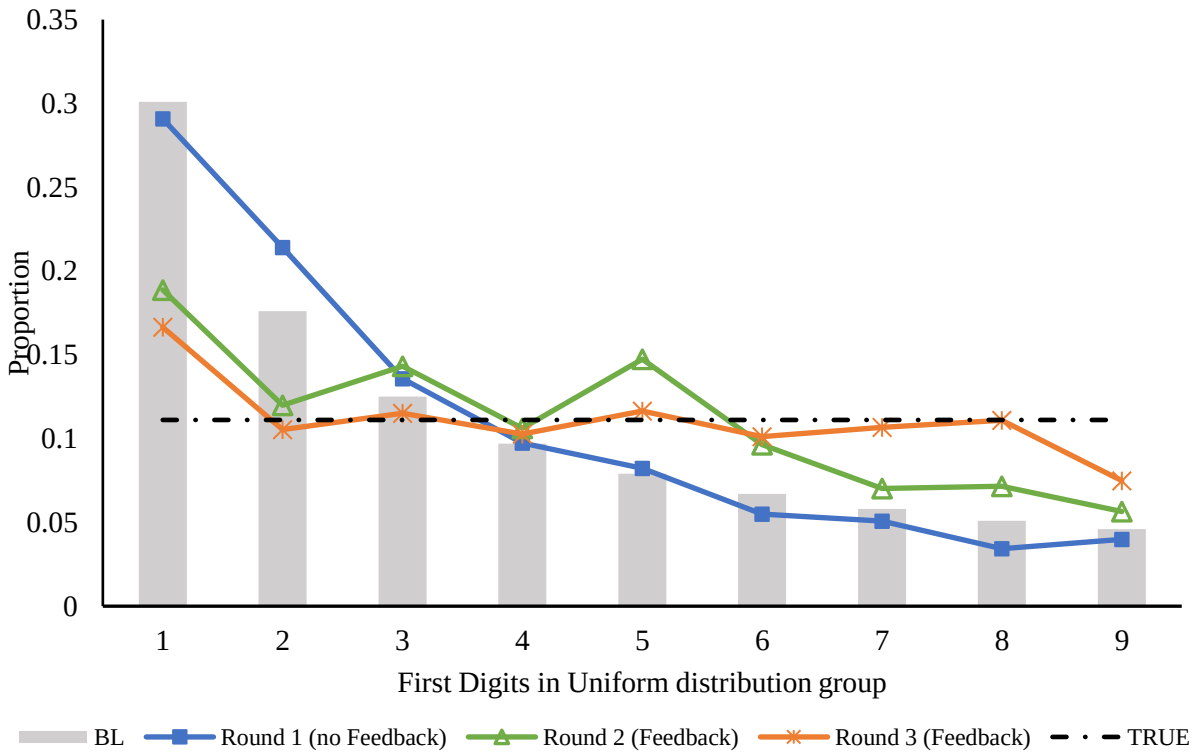


Figure 18. The first-digit proportions of the estimates to the Dots Display in three rounds for the Uniform group, compared against Benford's law and true distribution in Exp.2.

The shifted Benford bias in the anti-Benford condition. Similarly, in the anti-Benford group, where the larger leading digits occurred more frequently as true values, the initial Benford bias was also disrupted after receiving feedback. This was supported by a significant interaction (leading digits × three rounds) reported by the test of within-subject contrasts weighted by the proportions of BL, $F(1,76) = 45.422$, $p < .001$, $\eta^2 = .374$. Notably, there was a pronounced increase in the frequency of larger leading digits such as 7, 8, and 9 by Round 3 based on the pairwise comparisons (all $p < .05$). Thus, the first-digit pattern in the last round not only deviated from the monotonic decline but also visually demonstrated an upward potential. This observation encouraged us to conduct another contrast analysis weighted by the true distribution (rather than Benford's proportions), which was graphed with a dashed line in Figure 19. It reported that this true first-digit distribution could account for 27.6% of the variances of the estimates in the last round. Though not strong, its capacity to explain the variances in Round 3 was markedly superior to the mere 4.0% accounted by Benford's proportions.

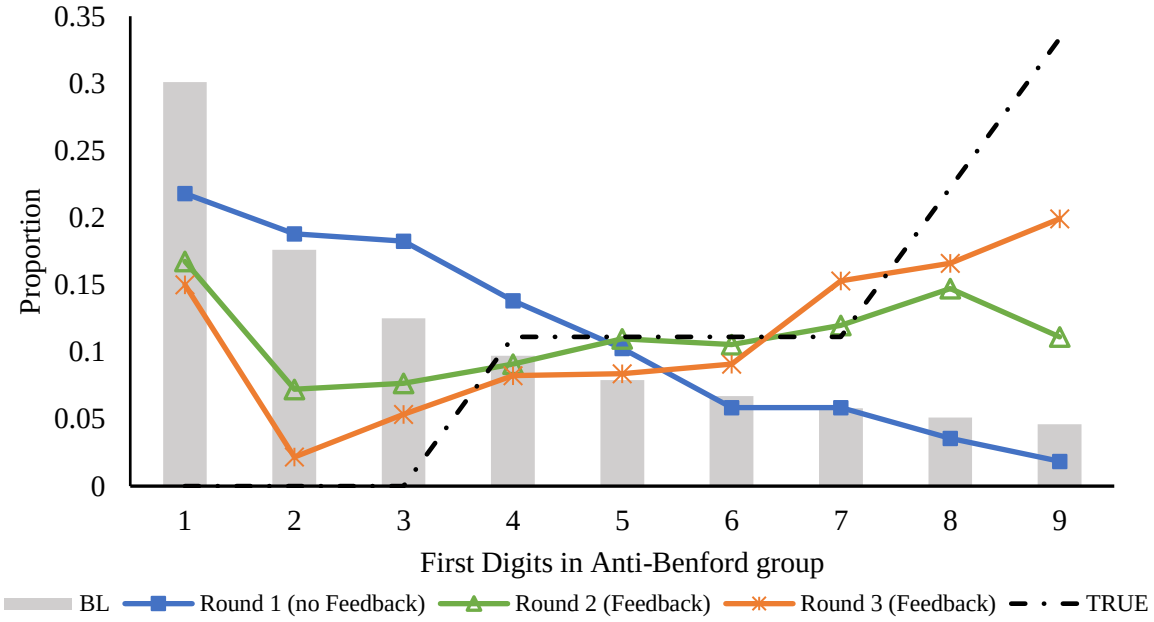


Figure 19. The first-digit proportions of the estimates to the Dots Displays in three rounds for the Anti-Benford group, compared to Benford's law and its true distribution in Exp.2.

The robust Benford bias in the Benford-like condition. In the Benford-like group, where the first-digit pattern of nine true values agrees with BL, the proportions of leading digits of the estimates generated over three rounds were plotted in Figure 20. Contrast analysis indicated that the variances of estimates in three rounds were all strongly explained by Benford's proportions (Round 1: η^2 = .689; Round 2: η^2 = .612; Round 3: η^2 = .634), demonstrating a strong agreement with the behaviour of BL, which corresponded to the actual distribution presented to the participants. Although the test of within-subject contrast analysis weighted by the proportions of BL found an interaction between the first digits and three rounds, $F(1,76)$ =5.608, p =.020, η^2 = .069, the effect size was not strong, and the monotonic decline was still preserved despite minor fluctuations at certain digits. Thus, providing feedback in the subsequent rounds did not, in itself, substantially alter the initial Benford bias. Such results established a good contrast with those changed patterns displayed in the uniform and anti-Benford condition, demonstrating that there was not inherently a change over three rounds, or that feedback itself caused a change.

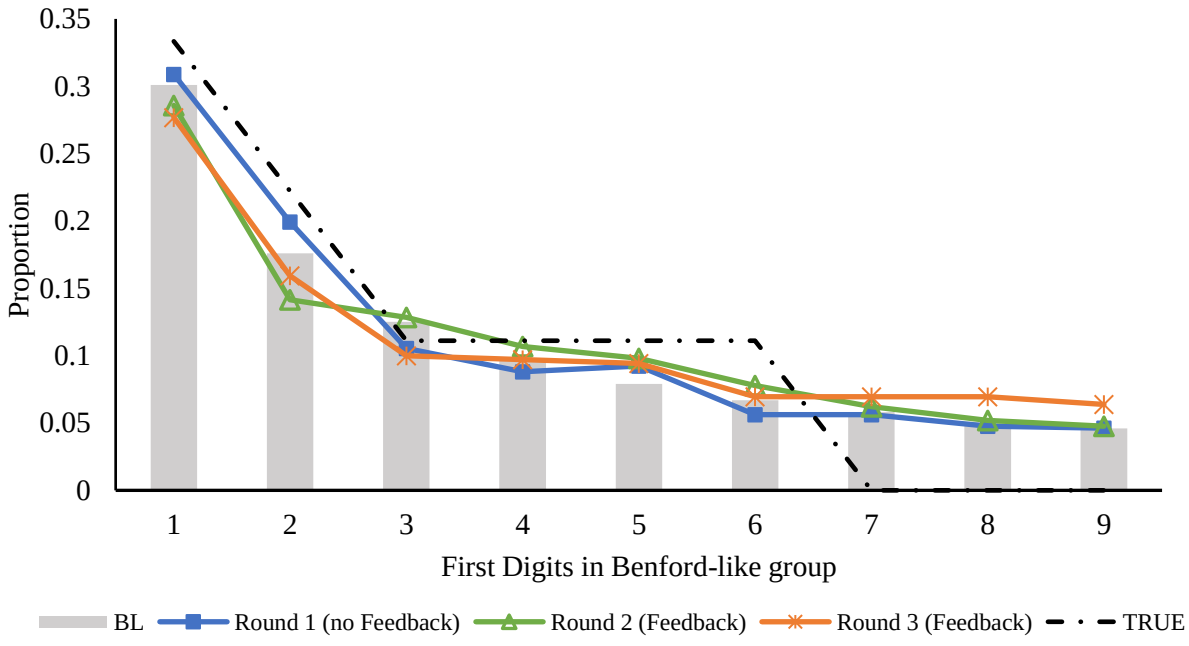


Figure 20. The first-digit proportions of the estimates to the Dots Displays in three rounds for the Benford-like group, compared to Benford's law and true distribution in Exp.2.

Accuracy analysis of full-number. Perceiving large numerosities is generally supported by the Approximate Number System (ANS), which is evidenced by a tendency towards underestimation in performance. Thus, I first calculated and presented the descriptive data of the mean estimates of nine pictures generated across different distribution groups in Round 1, where feedback was absent, as shown in Table 2. Notably, a consistent underestimation was observed in the Benford-like group with the presentation of 220, 280, 350, 1030, 1060, and 1090 dots, as well as in all test items for the anti-Benford group (450, 550, 650, 750, 820, 880, 930, 960, and 990), with all differences being statistically significant ($p < .05$). Surprisingly, all the mean estimates in the uniform group closely matched the actual quantities in the first round.

Table 2.

Means of the estimates generated in Round 1 without feedback for Uniform, Benford-like and Anti-Benford groups of Dots Displays in Exp.2.

	Uniform *(N=81)*			*Benford-like* *(N=77)*			*Anti-Benford.* *(N=78)*	
True values	*M*	*SD*	**True values**	*M*	*SD*	**True values**	*M*	*SD*
250	246.96	592.71	**220**	166.14*	236.59	**450**	273.00*	188.83
350	314.34	700.21	**280**	159.60*	171.88	**550**	361.47*	468.63
450	357.98	852.93	**350**	256.46*	329.73	**650**	393.35*	296.57
550	531.24	1283.91	**450**	347.75	691.46	**750**	512.01*	615.82
650	571.89	1537.42	**550**	428.92	559.94	**820**	520.53*	608.85
750	813.52	1855.02	**650**	515.39	772.48	**880**	558.81*	678.53
850	715.99	1566.63	**1030**	756.14*	1173.28	**930**	502.32*	321.24
950	750.66	1572.07	**1060**	731.74*	1257.88	**960**	565.90*	518.41
1050	835.64	1590.01	**1090**	773.22*	1255.21	**990**	612.73*	818.49

Note: * < .05, compared to the true value.

The first part presented the descriptive data of mean estimates in Round 1; however, to examine the learning effect on estimation performance across three rounds, two types of accuracy analysis were computed with respect to the size of errors. First, as in Experiment 1, I evaluated the accuracy of the *full-number* estimated by aggregating absolute differences (*ADs*) across nine items for each round. A repeated-measure ANOVA did not find the interaction between rounds and distribution type, $F(4,466) = .678$, $p = .607$, but reported the main effect of rounds, $F(2,466) = 46.108$, $p < .001$, $\eta^2 = .165$. This indicated that estimation performance differed among the three rounds, regardless of the pattern of the true values. Notably, marked improvement was observed from Round 1 to Round 2, but was substantially reduced from Round 2 to Round 3 (see Figure 21). Such a pattern suggested a diminishing learning rate with respect to the performance on the *full-number* estimated under the consistent influence of feedback.

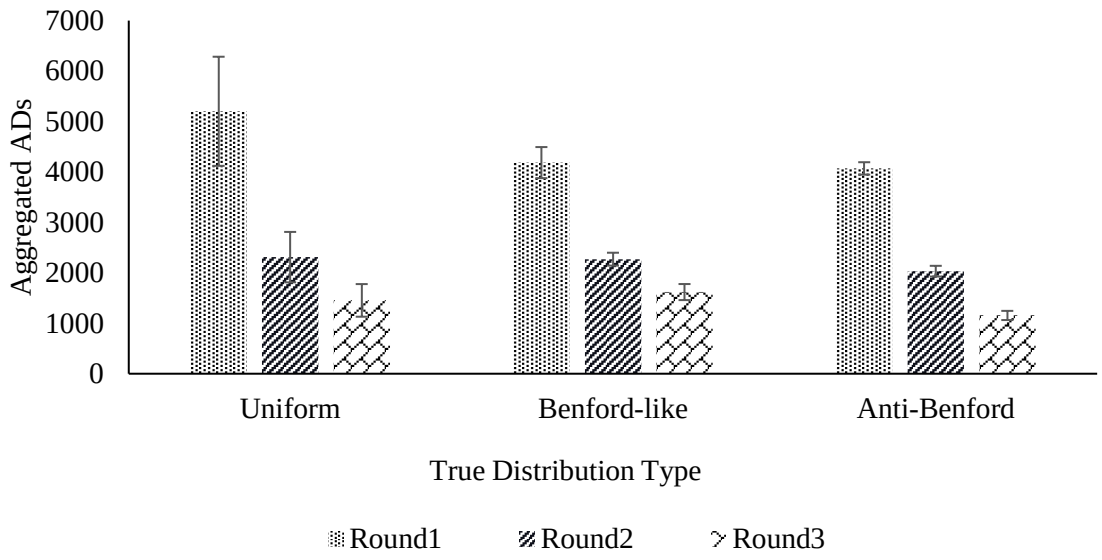


Figure 21. The mean of absolute differences (*ADs*) between the estimates and true values aggregated from nine dots items across three rounds of Dots Displays in Exp.2.

Accuracy analysis for first-digit estimates. In addition, I assessed the *first-digit* accuracy of the estimates, examining how often the estimated first digits matched the actual first digits across nine items in each round. The repeated measures ANOVA found the main effect of repeats on the first-digit accuracy rate, $F(2,466) = 123.35$, $p < .001$, $\eta^2 = .346$), indicating the influence of feedback. It also reported an effect of the true distribution type, $F(2,233) = 10.761$, $p < .001$, $\eta^2 = .085$, suggesting that the Benford-like group constantly achieved the best accuracy rate of first-digit estimated across three rounds. As displayed in Figure 22, pairwise comparisons revealed steady improvement across all three rounds (all $p < .05$), irrespective of the distribution type. This round-by-round improvement in the first-digit accuracy corresponded well with the consistent shift observed in the first-digit distribution.

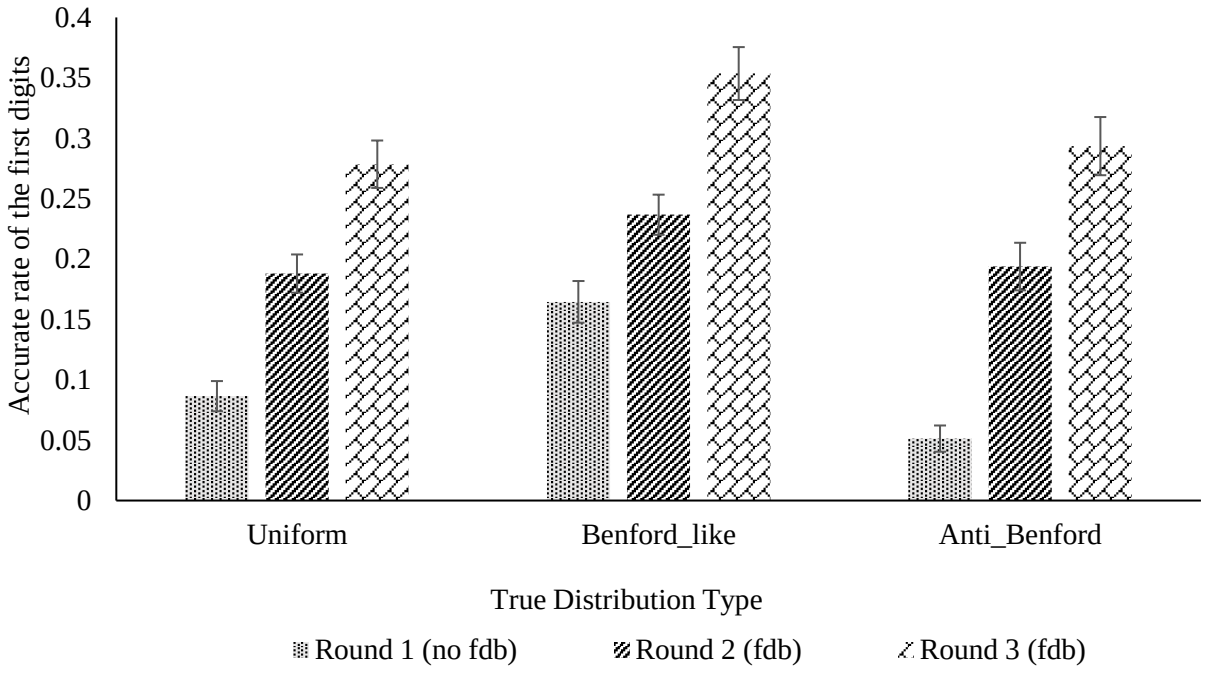


Figure 22. The first-digit accuracy rate for three true distribution groups across three rounds of Dots Displays in Exp.2.

Confidence and other analysis. The average confidence rating in each round is listed below in Table 3. The repeated measures ANOVA failed to report an interaction between

rounds and distribution types, $F(4,464) = 1.605$, $p = .107$, nor the effect of rounds, $F(2,464) = .156$, $p = .856$. Neither the presence of feedback nor the variation in true value patterns altered the confidence levels in the students' responses. In general, the accuracy was not associated with the confidence, except for a positive correlation between the confidence and first-digit accuracy rate reported in Round 3 within the uniform group, $r = .23$, $p < .05$. No differences were found in the first-digit pattern produced between English and Chinese speakers, or between females and males.

Table 3.

Means of confidence ratings on a five-point scale regarding participants' answers in three rounds of Dots Displays in Exp.2.

	Confidence rating	
Rounds	*M*	*SD*
Round 1	2.483	.683
Round 2	2.528	.793
Round 3	2.493	1.657

Results summary of Dots Displays. In line with the expectation, the Benford bias has been disrupted under the influence of feedback information when the true values deviate from the pattern of BL. Moreover, there was evidence suggesting that the first-digit pattern progressively shifts towards the true distribution, though more repeated measures are required to demonstrate a clearer trend. Such changes due to feedback are generally supported by improved performance, in particular, the enhanced accuracy detected at the first-digit place of the estimates generated.

Results for Animal Estimation

Two hundred eleven students remained in the analysis pool after excluding twenty-four students who failed the quiz modules and one incomplete dataset. 106 students were assigned to the questions with average information, while 105 students undertook the estimation task based on the juvenile data. In addition, during data cleaning, some invalid responses were selectively removed, such as when the value of the total lifespan of an animal was less than the starting point provided. In particular, the implausible answers were mostly detected for the juvenile group. Thus, the dataset for each individual might not always include nine answers for each variable asked.

A three-way interaction was found between the first digits, variables asked and reference type, as indicated by the contrast analysis weighted by the proportions of BL, $F(1,169) = 7.462$, $p = .007$, $\eta^2 = .042$. Additionally, a main effect of reference type on the first-digit distribution was reported, $F(1,169) = 198.728$, $p < .001$, $\eta^2 = .540$, along with a main effect of the variable asked, though with a small effect size, $F(1,169) = 5.491$, $p = .020$, $\eta^2 = .031$. Therefore, the following sections separately discussed the first-digit pattern observed for average and juvenile references.

First-digit patterns when given juvenile references. Two types of reference data were offered in this task. In the group where participants were given juvenile information, the first-digit distribution of the estimates for three variables asked all visually displayed a reasonable monotonic decline (See Figure 23). The test of within-subject contrasts weighted by the proportions of BL indicated that the variances of estimates in three rounds were all reasonably explained by Benford's proportions (Weight: $\eta^2 = .767$; Lifespan: $\eta^2 = .734$; Group-size 3: $\eta^2 = .584$). It also reported a significant interaction between the variable questioned and the first digits, $F(1,64) = 4.482$, $p = .038$, $\eta^2 = .065$. Some fluctuations were

observed at digit-1, digit-2, and digit-8. Digit-5 was consistently elevated, irrespective of the question type. Nevertheless, a strong Benford bias was consistently present across all three variables. Contrary to expectations, Benford bias was not disrupted as a result of introducing a Gaussian variable.

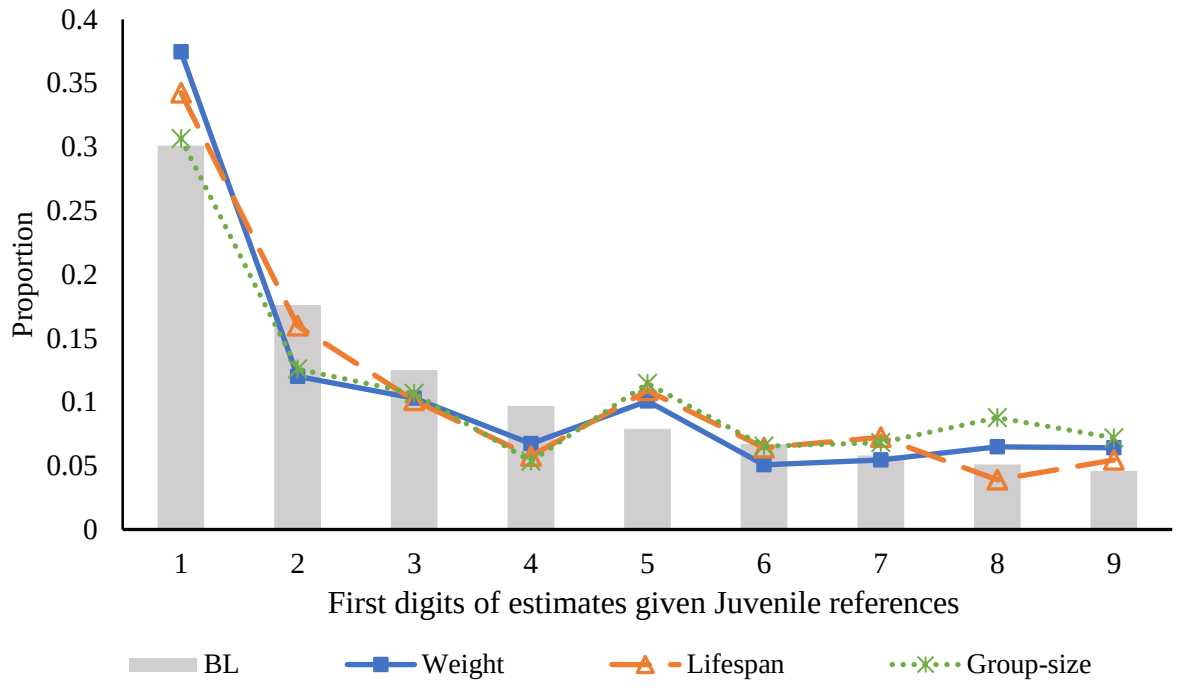


Figure 23. The first-digit proportions of the estimates in response to three types of variables questioned with Juvenile references in Animal Estimation, compared to Benford's law in Exp.2.

First-digit patterns when given average references. On the contrary, in the group where average information was offered, the first-digit distribution largely deviated from a monotonic decline (see Figure 24). The contrast analysis showed that the BND accounted for only a small portion of the variances for the three variables (weight: η^2 = .081, lifespan: η^2 = .297, group-size: η^2 = .180). The first-digit patterns were not significantly different from each other reported by the within-subject contrast analysis weighted by the proportions of BL, $F(1,105) = .203$, p =.653. Unlike the juvenile group, it lacked clear evidence to support the presence of a Benford bias when showing average data during estimation.

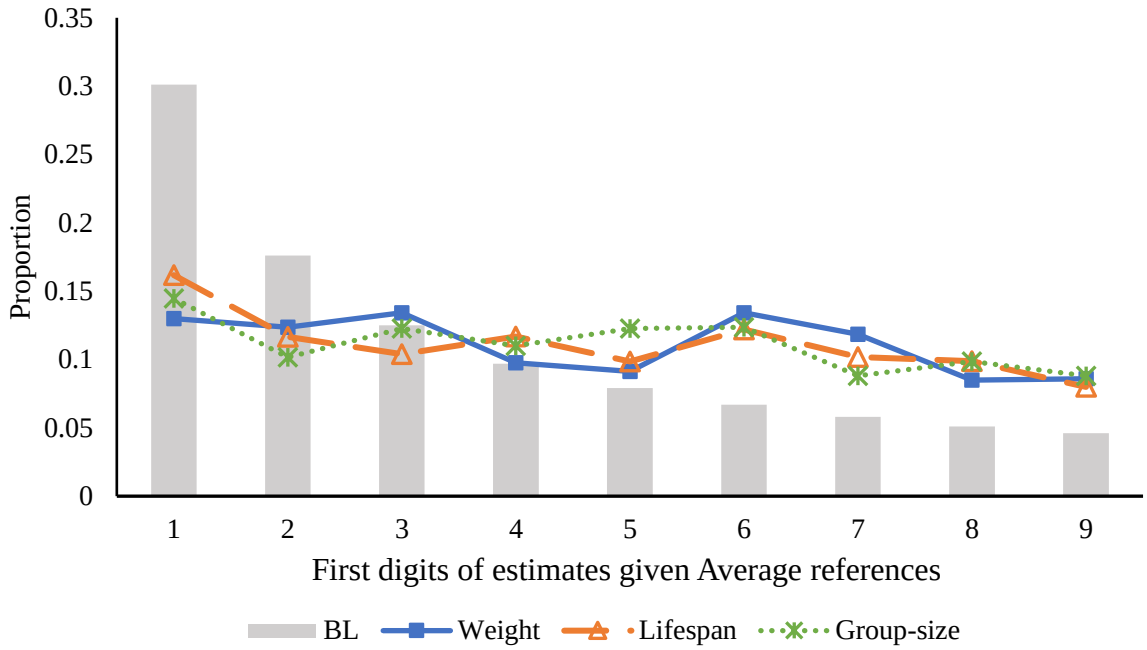


Figure 24. The first-digit proportions of the estimates in response to three types of variables questioned with Average references in Animal Estimation, compared to Benford's law in Exp.2.

Distributions of full-number. Further examination of the distributions of the full-number estimated may shed additional light on the above findings. In the case where participants provided estimates based on juvenile references, none of the distributions for the estimated answers conformed to normality. All twenty-seven distributions (3 variables × 9 animals) exhibited strong positive skewness based on the normality test (see Appendix N). In contrast, normal distributions were more frequently detected in the group provided with average references (see Appendix O). Five specific patterns of estimates, the group-size of herrings, the lifespan of blackbirds, the weight of dolphins, the lifespan of snakes, and both the age and group-size of crabs, were found to be closely normally distributed. Their skewness values fell within the -1 to 1 range, meeting the criteria for approximate normal distributions defined by Hair et al. (2022). Some estimates were even found to be negatively skewed, such as the estimates for the group-size of Garter snakes. It seems that the type of

variable being estimated did not systematically affect the underlying distribution of the numerical responses; however, the nature of the reference data provided did have an impact.

I also assessed the influence of reference data by measuring the differences between the mean estimated values and reference values provided. The group that received the average reference (see Table 4) showed that most estimated weights were not statistically different from the average data in the question. Even though the lifespan and group-size items were not supposed to be normally distributed, the mean estimates were still centred around the average data supplied in the question. The data suggested that approximately 55.4% of the weight estimates, 69.9% of the lifespan estimates, and 65.9% of the group-size estimates were generated within 10% differences of the average reference value provided in the questions. This implies that people are more likely to make minor adjustments when the average value is known. However, the group with juvenile references produced estimates that largely deviated from the starting point values offered in the questions, demonstrating a completely different estimation pattern (see Table 5). The data showed that, in the juvenile group, about 56.3% of the weight estimates, 64.7% of the lifespan estimates, and 47.8% of the group-size estimates were more than double the starting point value. More specifically, over a third of the estimates were a direct result of multiplying the starting value (e.g., 750) or the first digit of the starting value (e.g., 700) provided in the question by an integer, such as 2, 3, 4 or 10.

Other factors. An influence of gender on the first-digit pattern was not generally observed. The repeated measures ANOVA reported that the first-digit distribution of the weight estimates given the average data was different between English and Chinese speakers, $F(8, 728) = 2.142$, $p = .030$, $\eta^2 = .023$. However, this weak effect size suggested that this

Table 4.

Means estimates to nine items with reference value in the *Average group of Animal Estimation in Exp.2.*

	Weight			*Lifespan*			*Group-size*	
Reference	*M*	*SD*	**Reference**	*M*	*SD*	**Reference**	*M*	*SD*
25,000	27698.1*	12976.7	**4,500**	4441.1	1552.2	**250**	277.7	466.9
850	806.2*	271.1	**5,500**	5019.9*	1157.7	**4,500,000,000**	3641375496.0*	1678728584.0
750,000	818186.8	923728.5	**9,500**	9252.2	2506.3	**75**	71.6*	15.2
55	52.4	19.5	**850**	803.21*	174.1	**650,000**	646228.0*	601053.0
650,000	644964.8	136982.8	**15,000**	14763.4	3371	**550**	528.8*	79.2
150	160.4	116	**3,500**	3268.6*	940.8	**950**	866.2*	201.9
4,500	4405.2	1881.5	**7,500**	7086.8*	1229.1	**3,500**	3275.5*	682.0
950	923.8	170.1	**6,500**	6487.7	1531.4	**15,000**	14698.1	9065.4
3,500	3356.8	1074.7	**20,500**	20598.9	5731.3	**850**	802.0*	129.2

Note. * $p < .05$, compared to the average reference.

Table 5.

Means estimates to nine items with reference value in the *Juvenile group of Animal Estimation in Exp.2.*

	Weight			*Lifespan*			*Group-size*	
Reference	*M*	*SD*	**Reference**	*M*	SD	**Reference**	*M*	*SD*
9,500	28358.8*	35004.5	**2,500**	12914.4*	13733.1	**95**	26209.6*	183516.5
650	2421.6*	6967.1	**3,500**	9794.5*	10584.9	**2,500,000,000**	4117462963.0*	2155394905.0
550,000	1982662.9*	2668258.0	**7,500**	17799.4*	8798.7	**55**	1350.5*	8744.9
35	412.3*	1315.6	**650**	3223.7*	6956.2	**450,000**	89095711.0*	578548966.0
450,000	1199405.0*	1329097.0	**8,500**	20132.5*	10091.2	**350**	25989.1*	134572.3
85	1189.1*	4243.3	**1,500**	6680.8*	8840.9	**750**	14582.4*	91682.9
2,500	5950.0*	5827.1	**5,500**	15947.4*	13503.6	**1,500**	1421617.8*	11073791.7
750	2160.9*	2217.4	**4,500**	12036.3*	10643.3	**8,500**	191765.3*	1060566.9
1,500	8259.9*	12906.3	**9,500**	20088.4*	10545.2	**650**	47531*	196591.8

Note. * $p < .05$, compared to the juvenile reference.

interaction effect explained only a minor proportion of the variances. The language effect was not found in other conditions. Within the average reference group, participants were instructed to estimate based on one randomly selected image out of four (labelled with A, B, C, or D), and the analysis showed that the first-digit pattern did not vary significantly across these labels. Therefore, no picture effect was reported on the first-digit pattern in this case.

Results summary of Animal Estimation. Similar to Experiment 1, it failed to show that estimating different variables, such as weight and group-size, resulted in varied first-digit distributions. Regardless of the variable questioned, the Benford bias remained fairly robust. However, the provision of different types of references did lead to changes in the first-digit pattern of the estimates. Notably, the Benford bias was substantially suppressed when the average reference was provided, compared to the supply of a starting point value.

Discussion of Experiment 2

Consistent with the results in Experiment 1, once again, the second attempt failed to show a significant difference in the first-digit pattern when generating different variables. The estimated values exhibited similar shapes of distributions regardless of the items questioned. This contrasts with previous studies (e.g., Griffith & Tenenbaum, 2006; Lewandowsky, Griffiths & Kalish, 2009), which suggested that people's predictions were influenced by accurate prior probabilities. The results indicated that even when participants were explicitly informed about the underlying distributions, they might not apply those statistical relationships precisely when solving unknown questions. For example, most estimated weights in the sample became much deviated from a typical bell curve when they were supposed to be a Gaussian variable. Thus, the evidence collected so far challenged the reliability of the Distribution hypothesis that emphasised the sensitivity to the distributions of the estimated variables.

However, by examining the first-digit patterns between the groups receiving average and juvenile information, it was found that the type of reference point introduced different estimation behaviours. The group provided with juvenile parameters predominantly followed Benford bias, while the group relying on average data displayed a deviation. Further analysis of the distribution of estimates provided plausible justifications for these differences. I found that the responses generated based on the juvenile information were mostly positively skewed, resembling the power law distribution. In contrast, normal and negatively skewed distributions frequently emerged when using average references. As demonstrated in mathematics, Benford's law is a product of values generated from distributions that follow a power law (Berger & Hill, 2015; 2020). Therefore, it was not surprising to observe a suppressed Benford bias when the underlying pattern of the estimated answers deviates substantially from the power law. Such changes were also elaborated through the analysis of outputs compared to the reference values. While most participants made minor adjustments away from the average data, the estimates in the juvenile group substantially deviated from the starting point. Thus, the use of a particular type of reference data appeared able to disrupt the Benford bias by changing the cognitive processing involved in number generation.

While the evidence regarding people's sensitivity to underlying distributions from the animal estimation task was weak, promising results emerged from the Dots Display tasks when feedback information was provided. With the updated design in this second attempt, the feedback effect was clearly evident, as seen by the comparison of the first-digit pattern with feedback against the monotonic decline obtained in the initial block without feedback. As anticipated, the initial Benford bias remained robust in the Benford-like group, where the true values followed a logarithmic distribution. However, such a bias was immediately disrupted in the groups (i.e., uniform and anti-Benford) where the true values disobeyed Benford's proportions once feedback was introduced. Importantly, these changes extend beyond just

evening out the frequencies. Particularly in the anti-Benford condition, where the true values were heavily loaded on the *larger* leading digits, the proportions of the first digits began to exhibit an upward trend in the last round. This offered a positive indicator implying that the first-digit pattern tended to shift towards the true first-digit distribution. Of course, more repeated measures are desired for a follow-up on such observations.

To understand the continuous shift of the first-digit pattern with the feedback information, it was essential to investigate the accuracy, as it is reasonable to expect that feedback, even vague, could enhance task performance. Two types of accuracy analyses were conducted: 1) the accuracy of the *full-number* answered, and 2) the accuracy of the *first-digit* of the estimated answers. Results revealed that irrespective of the actual distribution of true values, participants' estimates in Round 2 were significantly better than their initial attempts under the influence of feedback, but the improvement between Rounds 2 and 3 was not always reported. Yet, a prominent enhancement in the accuracy of the leading digits estimated was consistently detected across all three rounds. This suggested consistent and steady improvement to the first-digit estimated as opposed to the full-number estimated. Hence, we argued that the ability to acquire the true first-digit distribution could provide a more reliable explanation for the continuous shift of Benford bias.

The results from the Dots Display tasks in this experiment demonstrated that estimation behaviour was sensitive to additional information, potentially highlighting the possibility that people are capable of learning the true first-digit distribution through interacting with feedback. However, a follow-up experiment with additional repeated measures is necessary to determine how enduring these changes are. If the observed shifts simply even out the probabilities *without* a clear movement towards the true pattern, then these deviations could alternatively be explained as a disruption in the cognitive process of how a number was generated, such as the change of strategies in number estimation.

Moreover, it was essential to include a between-subject control group without any feedback across all blocks in the next attempt to clarify that such a change in performance was not simply due to repetition. These led to the updates for the next experiment.

The findings of the animal estimation task, again, failed to support the distribution hypothesis, as there was insufficient evidence to demonstrate that people's estimation could consistently follow the underlying distribution of a variable, even when explicit knowledge of such distributions was provided. Nonetheless, this continuous failure highlighted a gap between knowing and applying, which has not been thoroughly investigated. As Fiser and Aslin (2002) discussed in their studies on implicit learning, the mechanism of automatic acquisition and application of statistical relationships requires further research to fully understand its limitations. We may not always be able to find a phenomenon like "probability matching," exhibiting a strong agreement between people's responses and the underlying frequencies. Hence, exploring the circumstances under which it may or may not emerge might become more informative than the phenomenon per se.

CHAPTER 6: EXPERIMENT 3

The attempts with Animal Estimation tasks failed to provide reliable evidence to support the Distribution hypothesis. However, promising results emerged from estimating the quantities of dots displayed. Therefore, Experiment 3 predominantly focused on the tests with Dots Displays, aiming to further clarify people's sensitivity to the first-digit distributions. The past experiments showed that the Benford bias could be disrupted under the influence of feedback, specifically when the actual first-digit values deviated from the norms of Benford's law, as seen in the uniform and anti-Benford groups. Hence, these two conditions were selected for further investigation in this experiment.

The result in Experiment 2 suggested that the estimation behaviour was sensitive to the feedback information given a rapid shift of Benford bias during Rounds 2 and 3. Yet, determining the final trend of these changes was challenging simply based on two rounds of feedback. As discovered by the empirical studies (e.g., Gebhart, Aslin & Newport, 2009; Weiss Gerfen & Mitchel, 2009), the initial statistical regularities acquired were very enduring and demanded constant exposure to the new pattern to trigger an update. In light of this, Experiment 3 included an additional round to assess whether more repetitions would reveal a more distinct and salient shift pattern, especially in the anti-Benford group, where larger leading digits are more common. To manage participant fatigue and maintain the task within the proposed 15-minute protocol, the test was confined to four rounds. Furthermore, I also introduced a between-subject control group where participants received no feedback in any repeats. This allowed us to contrast the feedback group to clarify if the disrupted pattern was simply due to extensive repetition and familiarity. Based on existing observations, I expected to see that the first-digit distribution of the estimates would progressively converge with the actual distribution under the consistent influence of feedback, assuming people can acquire the first-digit pattern given such sensitivity.

Methods of Experiment 3

This 2x2 repeated-measures design involved two between-subject variables: 1) the true distribution, alternating between uniform and anti-Benford conditions, and 2) the presence or absence of feedback.

Participants

A total of 337 participants from were recruited, and 327 responses were included in the analysis after data screening. Of these, 199 students were recruited from the university course Analytical Thinking (course code: ATHK1001), and 128 were from the SONA pool. The average age was 19.67 (*SD* = 2.28). The gender distribution was 57.8% female, 41.0% male, and 1.22% identifying as others. Chinese (43.4%) and English (41.3%) were the two most commonly spoken languages.

Procedure and Materials

Dots Displays. The general procedure remained the same as Experiment 2, but with the addition of a fourth round as another repetition. Participants experienced the same set of nine dot pictures in random order four times. They were randomly assigned to either the Feedback or Control group. In the initial round, no feedback was supplied. From the second round onward, the Feedback group received information about the accuracy of their estimates, while the Control group received no feedback throughout the tasks. Participants were also assigned to one of two distribution groups, where the true first-digit values conformed to either the uniform or anti-Benford distributions. The procedure for setting up the display time and feedback information was consistent with previous ones.

Questionnaires, general instructions, catch pages and confidence ratings. These used the identical materials developed in Experiment 2 with minor text changes to fit the updated design in Experiment 3.

Results of Experiment 3

Ninety-one students were allocated to the uniform distribution group with feedback, while seventy-seven received no feedback. In the anti-Benford condition, eighty-one students answered the questions based on the feedback, and seventy-eight in the control condition.

The contrast analysis weighted by the proportions of BL found a three-way interaction between the first digits, rounds and feedback groups, $F(1, 306) = 53.583$, $p < .001$, $\eta^2 = .149$. This suggested that the first-digit distribution across four rounds in the feedback group differed from that in the control condition. Additionally, the effect of distribution type was observed on the first-digit pattern across four rounds, albeit with a small effect size, $F(1, 306) = 4.047$, $p = .045$, $\eta^2 = .014$. Thus, I will present the analysis separately for the feedback and control conditions, and within each, I will discuss the pattern for the uniform and anti-Benford groups, respectively.

Benford Bias Emerged in Round 1

Consistent with the findings in Experiment 2, the first-digit distribution generated in Round 1 all demonstrated a reasonable Benford bias when no feedback information was provided (see Figure 25). The within-subject contrast analysis weighted by the proportions of BL failed to report an interaction between the first digits, distribution type and feedback condition, $F(1, 320) = .361$, $p = .548$, suggesting that the first-digit patterns in Round 1 were similar across conditions. The Benford-Newcomb distribution (BND) accounted for 59.4% of the variances in the feedback group and 64.0% in the control group when the true values followed the uniform distribution. For the anti-Benford group, the BND explained 46.2% and 42.8% of the variances for the feedback and control conditions, respectively. It is important to note that even if the participants were assigned to the feedback group, the first round never

provided them with any feedback information. A spike at digit-5 was not observed in this case. It was important to note that, unlike the observation in Experiment 2, the first-digit pattern obtained from the anti-Benford condition reported a good fit to Benford's distribution in this case.

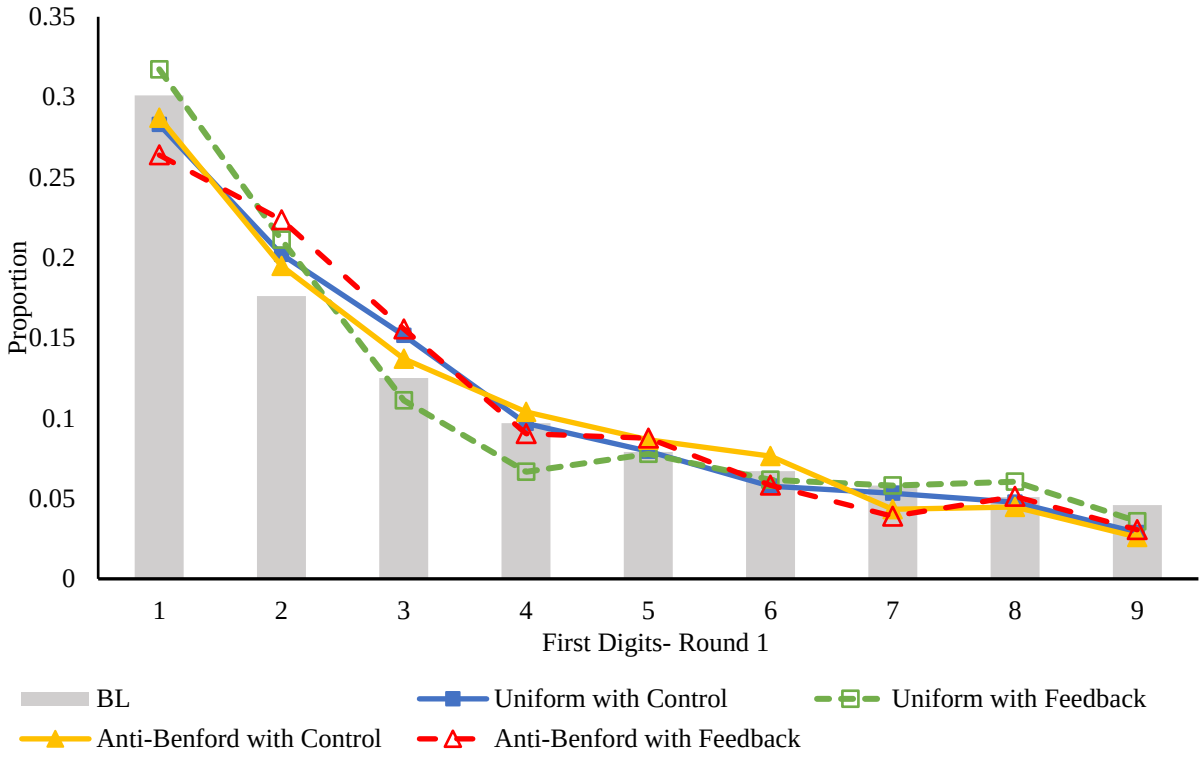


Figure 25. The first-digit proportions of the estimates to the Dots Display in Round 1 when no feedback was ever supplied irrespective of the feedback or control condition for both uniform and anti-Benford groups, compared against Benford's law in Exp.3.

Disrupted Benford Bias When Given Feedback

To assess the impact of feedback, the first-digit distribution produced from Round 1 to Round 4 was graphed separately for each distribution type (see Figures 26 and 27). The contrast analysis weighted by the proportions of BL reported a three-way interaction between the first digits, rounds and distribution type within the feedback condition, $F(1, 161) = 5.836$, $p = .017$, $\eta^2 = .035$, showing the impact of distribution type but with a small effect size. As

expected, the first digit pattern deviated from BL after providing feedback in the uniform and anti-Benford groups given a significant interaction effect between the first digits and four rounds reported by the within-subject contrast analysis (Uniform: $F(1, 87) = 52.819$, $p < .001$, $\eta^2 = .378$; Anti-Benford: $F(1, 74) = 76.061$, $p < .001$, $\eta^2 = .507$). This suggested that the first-digit patterns were different in each round. The explanatory power of BND progressively declined from Round 2 to Round 4 in the uniform distribution condition (Round 2: $\eta^2 = .424$; Round 3: $\eta^2 = .173$, Round 4: $\eta^2 = .103$). As shown in Figure 25, the first-digit pattern becomes closer to the true distribution, a flat line, despite digit-1 still spiking. Likewise, in the anti-Benford condition, Benford bias also shifted from Round 2 ($\eta^2 = .097$) onwards, given the feedback, and it progressively moved towards the true proportions where higher leading digits occurred more frequently (see Figure 27). A contrast analysis weighted by the anti-Benford distribution (rather than the BL) reported that it could explain 27.4% of the variances in Round 3 and 52.0% in Round 4, indicating an increasing conformity to the true distribution over time.

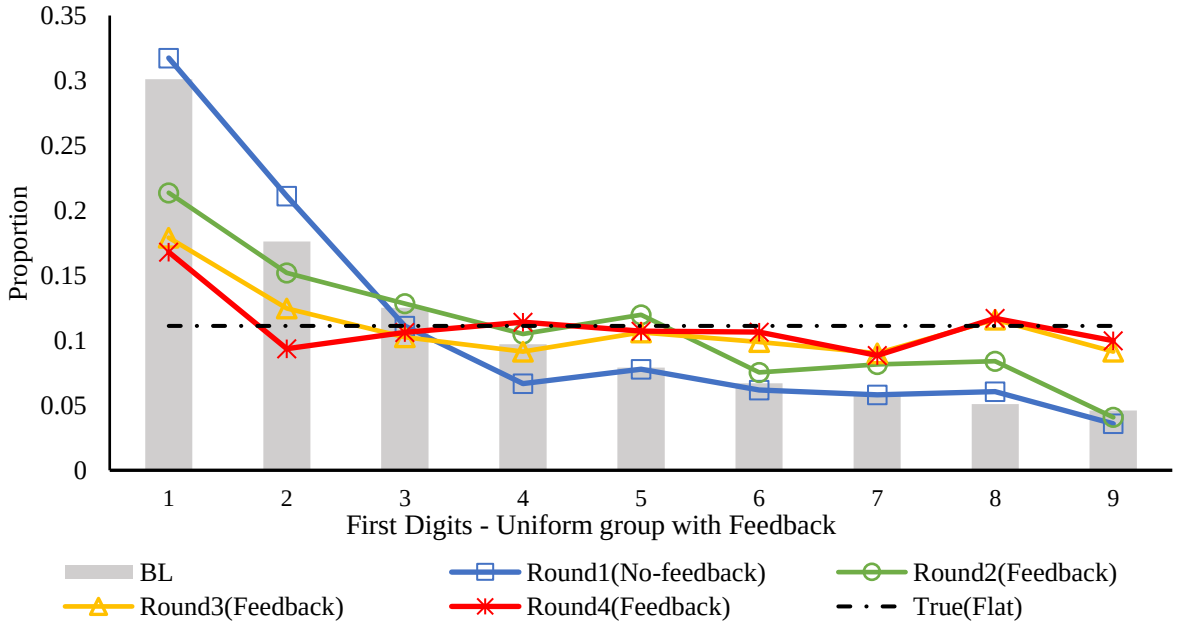


Figure 26. The first-digit proportions of the estimates to the Dots Displays across four rounds in the Uniform condition with feedback from Round 2 to 4, compared against Benford's law and its true distribution in Exp.3.

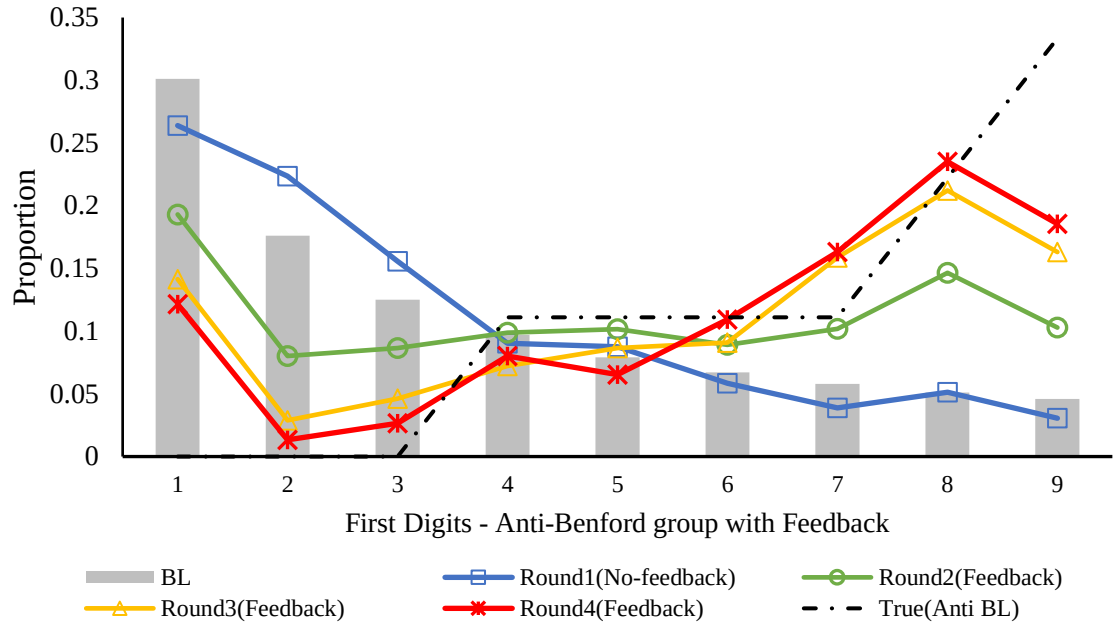


Figure 27. The first-digit proportions of the estimates to the Dots Displays across four rounds in the anti-Benford condition with feedback from rounds 2 to 4, compared against Benford's law and its true distribution in Exp.3.

Robust Benford Bias Without Feedback

The examination of the estimates from the control group additionally demonstrated a clear contrast, showing that Benford bias remained relatively stable across all four rounds when no feedback was provided, irrespective of the actual distribution. Unlike the feedback condition, Figure 28 and Figure 29 illustrate a consistent monotonic decline of the first-digit distributions through rounds 1 to 4 for both uniform and anti-Benford groups, respectively. Again, the elevation at digit-5 in the control condition was not observed in this case. While the proportions of some digits did fluctuate without feedback inputs, the overall pattern from these repeated tests did not show significant statistical variance, as indicated by the within-subject contrast analysis weighted by the proportions of BL (uniform: $F(1, 73) = 1.801$, $p = .184$; anti-Benford: $F(1, 72) = 1.515$, $p = .222$). The effect of distribution was not found in

this control condition due to a non-significant interaction between first digits, rounds and distribution groups based on the contrast analysis, $F(1, 145) = .119$, $p = .731$.

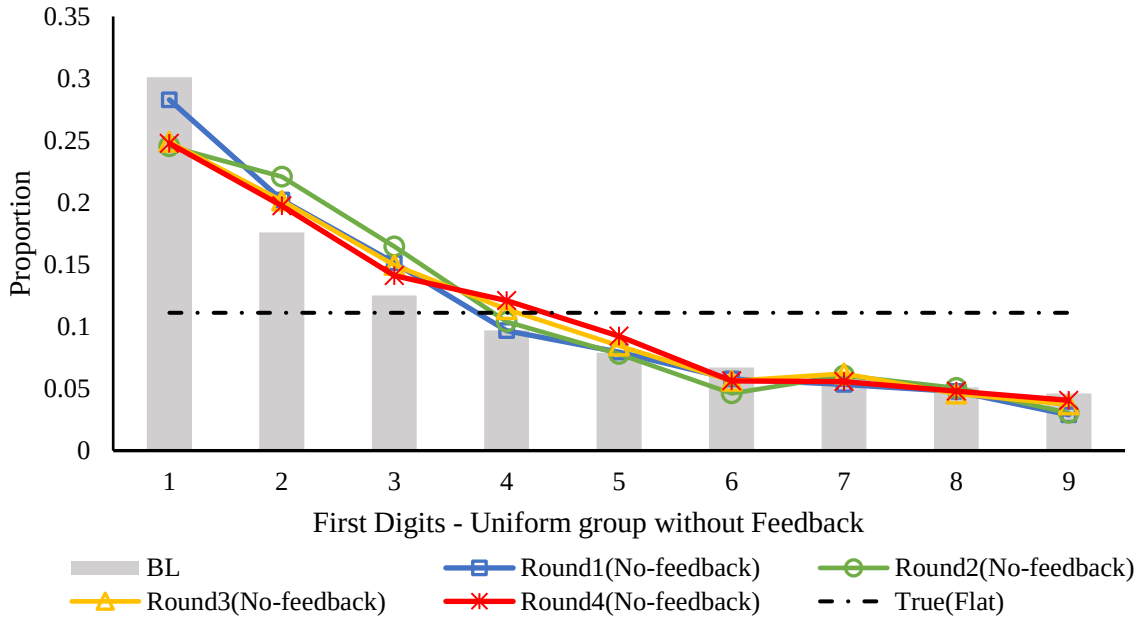


Figure 28. The first-digit proportions of the estimates to the Dots Displays across four rounds in the Uniform group without feedback, compared against Benford's law and its true distribution in Exp.3.

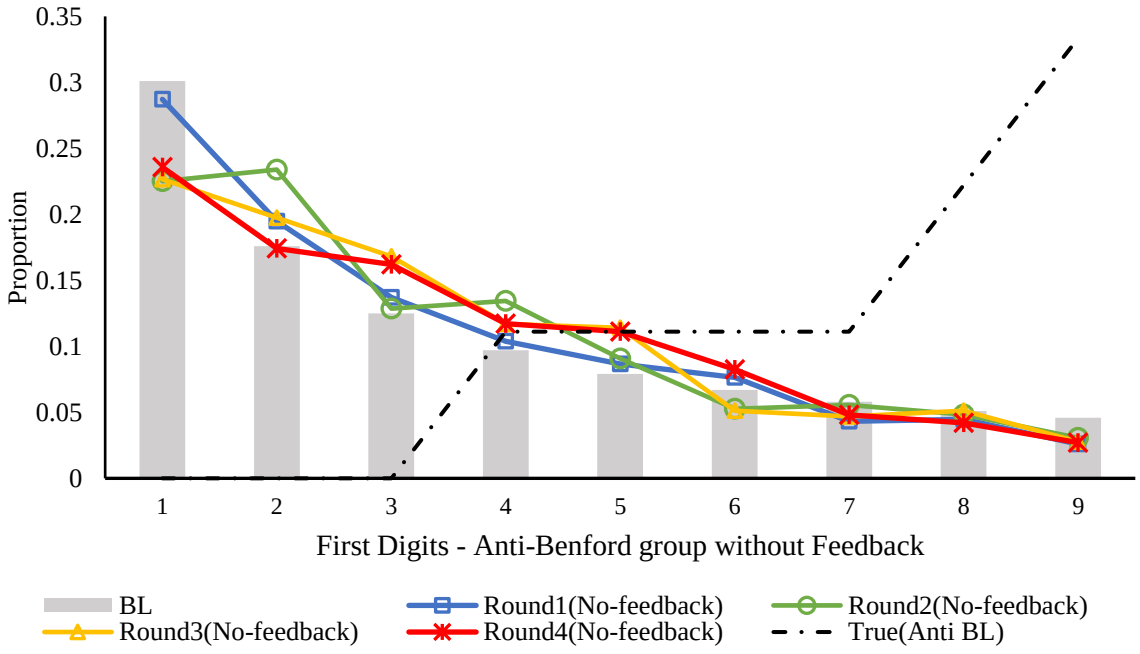


Figure 29. The first-digit proportions of the estimates to the Dots Display across four rounds in the anti-Benford group without feedback, compared against Benford's law and its true distribution in Exp.3.

Accuracy Analysis of Full-number

Table 6 listed the descriptive data regarding the mean estimates generated for nine test items in Round 1 for uniform and anti-Benford groups, respectively. The feedback was consistently absent in Round 1; thus, the Feedback and Control groups were not differentiated in this analysis. An underestimation was always observed for both distribution groups based on the t-test when comparing the mean against its true value.

Table 6.

Means estimates of quantities of nine dots pictures in Round 1 without feedback for uniform and Anti-Benford groups of Dots Displays in Exp.3.

	Uniform (N=168)			*Anti-Benford (N=159)*	
True values	*M*	*SD*	**True values**	*M*	*SD*
250	164.725*	275.809	**450**	298.611*	523.560
350	215.729*	356.636	**550**	404.490*	906.639
450	252.713*	255.111	**650**	429.044*	812.369
550	333.054*	639.611	**750**	551.253*	1035.543
650	381.359*	412.300	**820**	570.000*	985.984
750	453.371*	691.829	**880**	601.449*	1369.211
850	536.575*	869.773	**930**	652.449*	1313.551
950	539.071*	938.289	**960**	595.510*	1164.565
1050	601.401*	779.227	**990**	699.554*	1466.852

Note: * < .05, compared to the true value.

To further examine the effect of learning on estimation performance, I computed the aggregated absolute differences (*ADs*) across nine items for each round to measure the size of errors. The repeated measures ANOVA reported the interaction between the four rounds and

feedback conditions, $F(3, 918) = 34.061$, $p < .001$, $\eta^2 = .128$, as well as for the distribution conditions but with a small effect size, $F(3, 918) = 4.227$, $p = .006$, $\eta^2 = .014$. Similar to the findings in Experiment 2, the aggregated *AD*s varied under the influence of feedback, as evident by the main effect of rounds reported from the repeated-measures ANOVA for both distribution groups (uniform: $F(3, 261) = 85.289$, $p < .001$, $\eta^2 = .495$; anti-Benford: $F(3, 222) = 40.569$, $p < .001$, $\eta^2 = .354$). Pairwise comparisons revealed progressively reduced *AD*s across the first three rounds (all $p < .05$), showing steady improvement. However, this learning effect decayed in the final round, as the differences between the last two rounds were nonsignificant for both distribution groups (both $p > .05$), suggesting that feedback did not further enhance the accuracy of the full-number estimated from Round 3 to Round 4. When no feedback was ever provided, neither the uniform, $F(3, 219) = .147$, $p = .931$, nor the anti-Benford group, $F(3, 216) = 1.617$, $p = .186$, reported the round effect from the repeated measures ANOVA, as display in Figure 30.

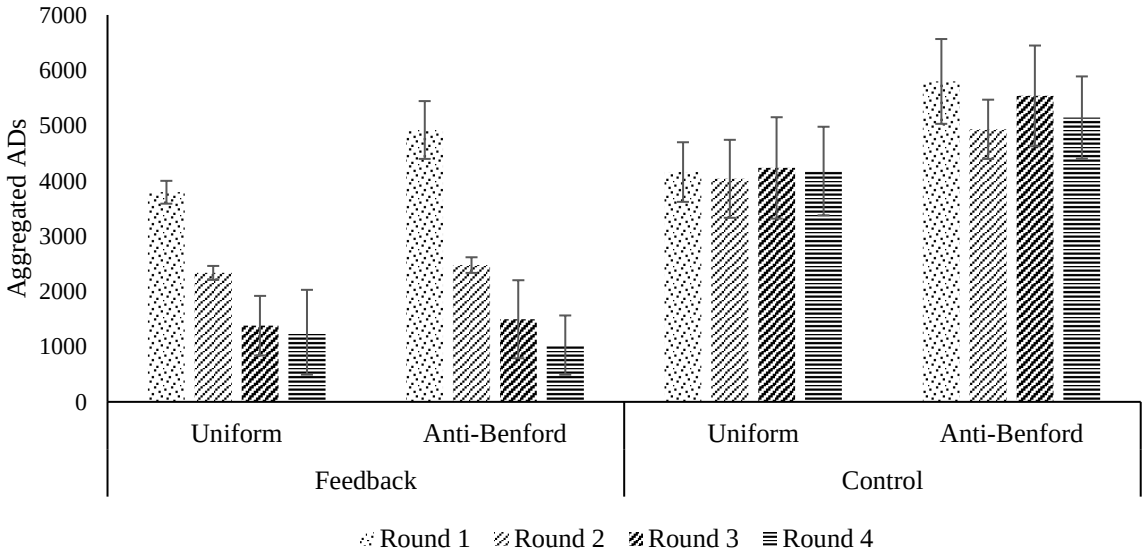


Figure 30. The mean of absolute differences (*AD*s) between the estimates and true values aggregated from nine dots items over four rounds in Exp.3.

Accuracy Analysis for First-Digit Estimates

I also evaluated the accuracy rate concerning the first digits of estimated quantities. The repeated measures ANOVA reported an interaction between the rounds and feedback conditions, $F(3, 903) = 70.910$, $p < .001$, $\eta^2 = .191$, showing the general effect of the feedback on such performance. As shown in Figure 31, in the condition where the feedback was never supplied, the first-digit accuracy did not change across four rounds (uniform: $F(3, 216) = .968$, $p = .409$; anti-Benford: $F(3, 216) = 1.571$, $p = .196$). But when people consistently interacted with the feedback information, the first-digit accuracy varied across rounds, as evident by the main effect of rounds for both distribution groups according to the repeated measures ANOVA (uniform: $F(3, 258) = 46.710$, $p < .001$, $\eta^2 = .352$; anti-Benford: $F(3, 213) = 60.733$, $p < .001$, $\eta^2 = .461$). The pairwise comparison further revealed that the first-digit accuracy rate was enhanced round by round (all $p < .05$), except for the difference between Round 3 and Round 4 in the uniform group ($p > .05$). Nevertheless, this corresponded with the observation regarding the shift of Benford bias with feedback that only a minor change on the first-digit distribution was detected between the last two rounds in the uniform group but a substantial shift in the anti-Benford condition, as indicated by the

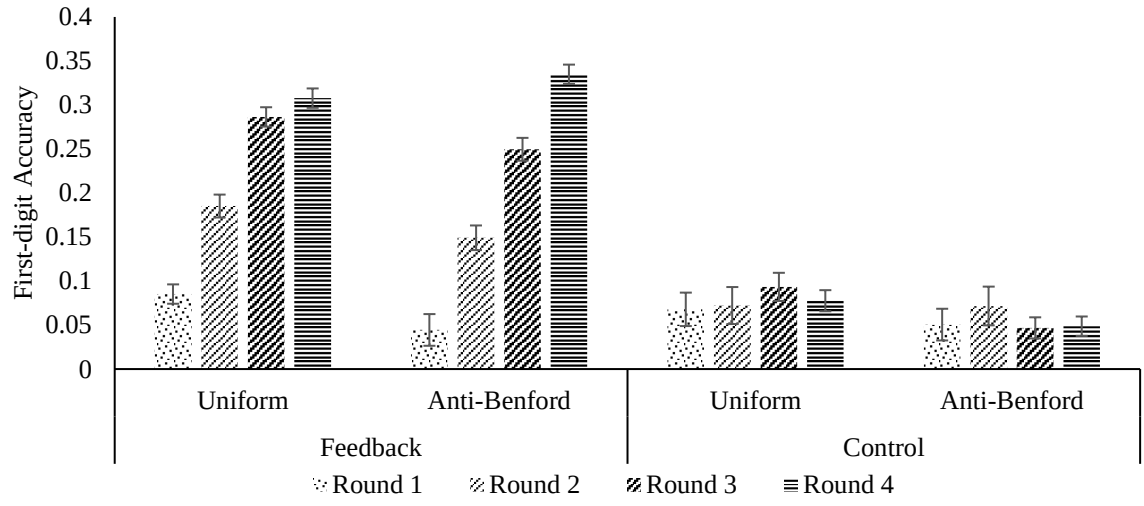


Figure 31. The first-digit accuracy rate in four conditions across four rounds of Dots Displays in Exp.3.

contrast analysis with the eta-square. Compared to the analysis result for the estimation performance on the full-number generated, it seems that the pattern of the first-digit accuracy well captured and echoed the changes of the Benford bias in each distribution group.

Confidence and Other Analysis

As shown in Figure 32, the repeated measures ANOVA reported an interaction between the four rounds and feedback groups for the confidence ratings, $F(3,924) = 6.648$, $p < .001$, $\eta^2 = .021$. Pairwise comparisons indicated that the confidence in the last round was higher than that in the first round under the influence of feedback, irrespective of the true distribution type (both $p < .05$). However, given the small effect size, the confidence was not strongly enhanced. The control group did not report any changes in the confidence ratings from the repeated measures ANOVA ($p > .05$). The influence of the distribution group was not found on the confidence level, $F(3,924) = 2.054$, $p = .105$. The correlational analysis did not report any association between the estimation performance and the confidence. No gender or language differences were detected regarding the first-digit pattern in this case.

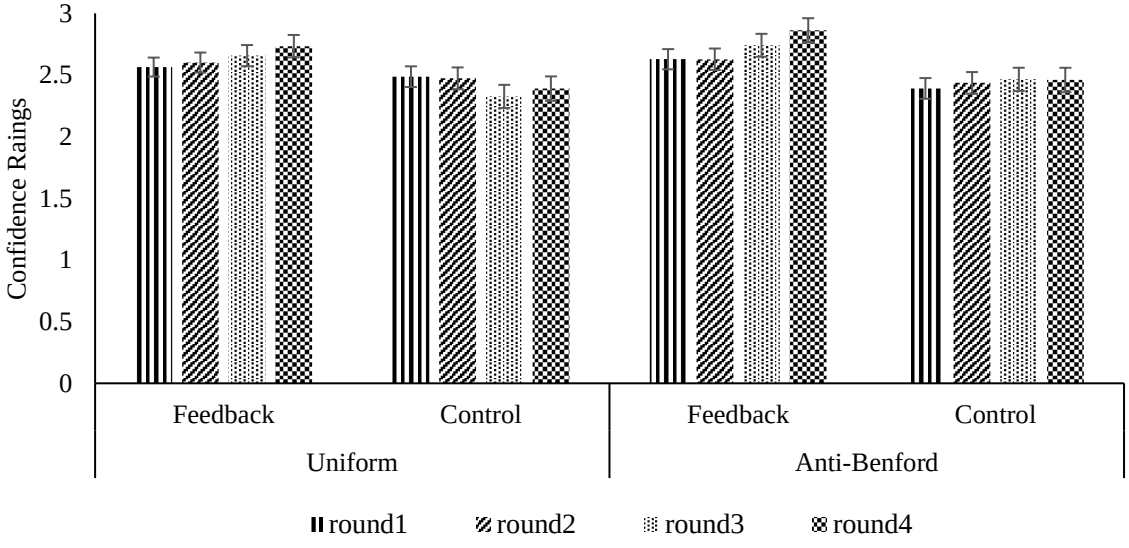


Figure 32. Mean confidence ratings on a five-point scale over four rounds of Dots Displays in Exp. 3.

Discussion of Experiment 3

Consistent with Experiment 2, Benford bias emerged in the initial round when estimating the quantities of dots without additional information. This bias was quickly disrupted once feedback was introduced, particularly when the actual first-digit distribution diverged from Benford's law. Remarkably, this disruption not only diminished the prevalence of smaller first digits but also, with more repetitions, shifted noticeably towards the true proportions. Particularly, in the anti-Benford condition, the initial monotonic decline transitioned to an inclining pattern in the final round, where larger leading digits became more common. To analyse this shift, I examined the estimation performance. It discovered that, consistent with previous experiments, changes in first-digit accuracy more precisely captured the shifting pattern of Benford bias, compared to the estimation of full-number. Therefore, the first digits of the estimates appear more sensitive to feedback information and should provide a more reliable basis for explaining the observed shift in Benford bias.

The results from the Dots Display task implied that estimation behaviour quickly adapted to new information, even if it was vague and subtle. Moreover, such changes became more sensitive to the first-digit proportions, suggesting that people could acquire the regularities of the first-digit pattern by interacting with feedback. This provided promising evidence to support one of the speculations that Benford bias is a consequence of learning the first-digit distribution. However, such findings remained inconclusive due to the limited magnitude of true values tested with a single domain (dots). That is, the observed changes might also be associated with the actual size of the answer, mostly constrained to three digits in this test. Thus, to further clarify the impact of the magnitudes of true values, my next attempt would vary the magnitude of the answers. Additionally, the visual estimation tasks used so far have been limited to items from the same domain, such as estimating the number

of dots or jellybeans in a jar (e.g., Chi & Burns, 2022). Unlike number generation tasks involving trivia questions across various contexts, this limitation might affect the generalisability of the findings. Thus, to extend my observation and test the scope of people's sensitivity to first-digit regularities, I plan to introduce visual stimuli from multiple domains and quantities in the subsequent experiment to see if this learning behaviour persists across varied contexts and magnitudes.

CHAPTER 7: PRELIMINARY STUDY FOR EXPERIMENT 4

Building on the limitations identified in Experiment 3, my final investigations sought to address these by varying the magnitude of true values and incorporating multiple domains for estimation targets. The significance of including various magnitudes also stems from the limitations observed when Benford's law failed to emerge in datasets restricted to a narrow range in naturally occurring settings (Durtschi, Hillson & Pacini, 2004). Therefore, this revised design enabled us to broaden the observations in examining the feedback effect on Benford bias. The ability to apply the extracted underlying structure to novel situations has been well-documented in studies of statistical learning, serving as evidence of knowledge transfer (e.g., Reeder, Newport & Aslin, 2013; Orbán, Fiser, Aslin & Lengye, 2008). This design intends to offer ample opportunities for learners to acquire the rules governing the proportions of the first digits across magnitudes. In other words, I was testing whether the feedback for a target response in one domain could enhance the estimation performance in another, irrespective of the magnitude of true values. Moreover, if performance related to the first digits of estimated responses can be improved through interaction with vague feedback in a more open and unrestricted condition, this would not only help to clarify its association with answer size and context but also expand the scope of testing people's ability to follow the first-digit distribution. Thus, such investigation could provide additional support for the generalisation of statistical learning.

Another crucial update to the testing materials involved the dynamic presentation of visual elements, contrasting with the static pictures used in previous tasks. Past literature has shown that cognitive representation of numerosity can be influenced by the perception of depth (e.g., Aida, Kusano, Shimono & Tam, 2015) and movement (e.g., Au & Watanabe, 2013). However, the previous test using dynamic images of jellybeans (Chi & Burns, 2022) consistently exhibited Benford bias in the number generation tasks. Therefore, presenting

visual elements in motion shall not be expected to significantly affect the occurrence of Benford bias in the number estimation.

The proposed design of the final set of experiments marked my first attempt to extend visual estimation tasks to various domains with multiple magnitude values while investigating Benford bias. However, I never tested whether Benford bias also manifests in estimates involving multi-magnitude values from diverse domains. Thus, I believed that running a pilot experiment in the control condition was essential before introducing the feedback. Given the empirical evidence that consistent Benford bias was reported through generating the numerical responses to various trivia or general knowledge questions with multi-digit true values (e.g., Burns, 2009), I also expected to see Benford bias in a paradigm using visual stimuli when the true values vary significantly.

This preliminary test was a 2 × 2 mixed design. The within-subject variable manipulated was whether the true values of nine estimated items were selected from single magnitude (i.e., four-digit length) or multi-magnitude (i.e., varied between three to five-digit length) answers. The presentation of targets from the same or different domains was a between-subject variable for estimates involving single-magnitude answers, but this became a within-subject variable when estimating the items with multi-magnitude answers. This design was intended to create a compact test that could be completed within 15 minutes while gathering more data using fewer participants over a brief recruitment period.

Methods of Preliminary Study for Experiment 4

Participants

A total of 113 undergraduate students were recruited, and 112 were included in the analysis after data screening, with an average age of 20.38 (*SD*=3.28). Of these participants, 64.6% identified as female, 33.6% as male, and 0.9% as other gender categories. The most

common native languages among the participants were English (46.0%) and Chinese (33.6%).

Procedures and Materials

Video Estimation. Participants were asked to estimate the number of targets displayed in a short video clip, with the average display time ranging from 10 to 13 seconds. The video clips used in the task were formatted in mp4, ensuring compatibility with commonly used internet browsers such as Google Chrome, Firefox, and Internet Explorer. Except for the dots displays, which were presented in motion at 2D display, the other scenes were rendered in 3D motioning involving depth. These 3D scenes displayed the targets from at least two different angles, providing both insider and outsider views of the entire population. Participants were required to click a button to initiate the display, which began only after the clip was fully loaded.

This task consisted of three randomised blocks, each containing nine estimation items in random order, resulting in a total of twenty-seven items. One block, known as the single-magnitude condition, included nine clips with true quantities of targets exclusively from four-digit lengths: 1500, 2500, 3500, 4500, 5500, 6500, 7500, 8500, and 9500. In this block, participants encountered items either from the same domain or different domains, assigned randomly. The different domain condition was presented with clips from nine different targets, each randomly paired with a distinct true value. These scenarios included diverse subjects such as Audiences, Birds, Cheering Men, Crows, Dancers, Moving Dots, Protestors, Skeletons, and Whales. Conversely, the same domain condition consistently presented clips from only one of those scenarios above. An example is illustrated in Figure 33. Appendix P shows a partial list of video clips created in the item pool and provides links to them so that they can be viewed.

The remaining two blocks comprised multi-magnitude conditions. Each block contained nine clips with true values spanning multiple digit lengths. Specifically, three clips contained targets in the hundreds (e.g., 450), another three in the thousands (e.g., 5500), and the last set of three clips in the tens of thousands (e.g., 15000). One of these two blocks presented items from the same domain randomly selected from the nine scenarios, while the other varied the domains of nine clips.

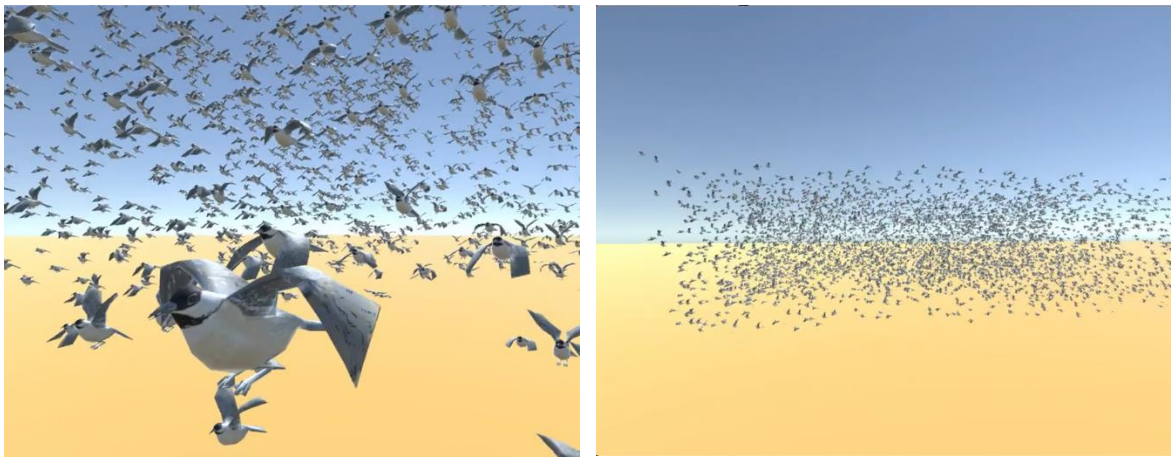


Figure 33. This is an example of two screenshots of the video clip showing 6,500 birds in a flock. This video lasted for ten seconds in a dimension of 932 × 666. The left screenshot presents an insider view of observing the birds inside of the flock. The motion of this view displayed is self-centred and autobiographical in 360 degrees. The screenshot on the right captures the side view of observing the birds flock at a distance. The motion of this view displays the overall size of the bird flock in a 360-degree rotation.

Questionnaires, general instructions, catch pages and confidence ratings. These used the identical materials developed in Experiment 3 with minor changes to the texts to fit the updated design in this pilot test. A significant change from previous experiments was showing twenty-seven video clips with dynamic items instead of static pictures. This change was anticipated to require considerable data usage, a point that was explicitly noted on the warning page at the start of the test. Given that the clips featured numerous similar objects moving rapidly, a warning page fully disclosed this aspect to participants (see Appendix Q). To assess participants' experience and the validity of their responses, a feedback page was

included at the end of the task, providing an opportunity for participants to leave comments and report any issues they encountered (see Appendix R).

Results of Preliminary Study for Experiment 4

The experiment conducted in this chapter was considered a pilot test to examine if the Benford bias can also be observed when showing quantities across magnitudes among different targets. Thus, to recruit more data within a short period of test time, the design of this experiment was not ideally structured. Specifically, the domain variable was manipulated as a within-subject factor in the multi-magnitude condition but as a between-subject factor in the single-magnitude group. Thus, this made it challenging to perform an analysis in terms of a general interaction effect across all. So, a specific analysis was conducted within each magnitude group.

The contrast analysis weighted by the proportions of BL indicated that presenting the same or different domains did not significantly influence the degree of Benford bias, regardless of the magnitude provided (single-magnitude: $F(1,110) = .411$, $p = .523$; multi-magnitude: $F(1,111) = .029$, $p = .864$). Thus, we focused on testing the effect of magnitude on the first-digit pattern, as illustrated in Figure 34. The within-subject contrast weighted by Benford's proportions reported an interaction between the nine digits and the two types of magnitude, $F(1,111) = 8.935$, $p = .003$, $\eta^2 = .074$. The contrast analysis further showed that the Benford-Newcomb distribution (BND) accounted for 77.8% of the variances in the multi-magnitude condition and 55.4% variance in the single-magnitude condition. This result indicates a more pronounced Benford bias in estimates pertaining to quantities that cover multiple-digit lengths, as opposed to those concerning single-magnitude quantities limited to four-digit numbers. Notably, digit-5 was significantly elevated in both magnitude conditions.

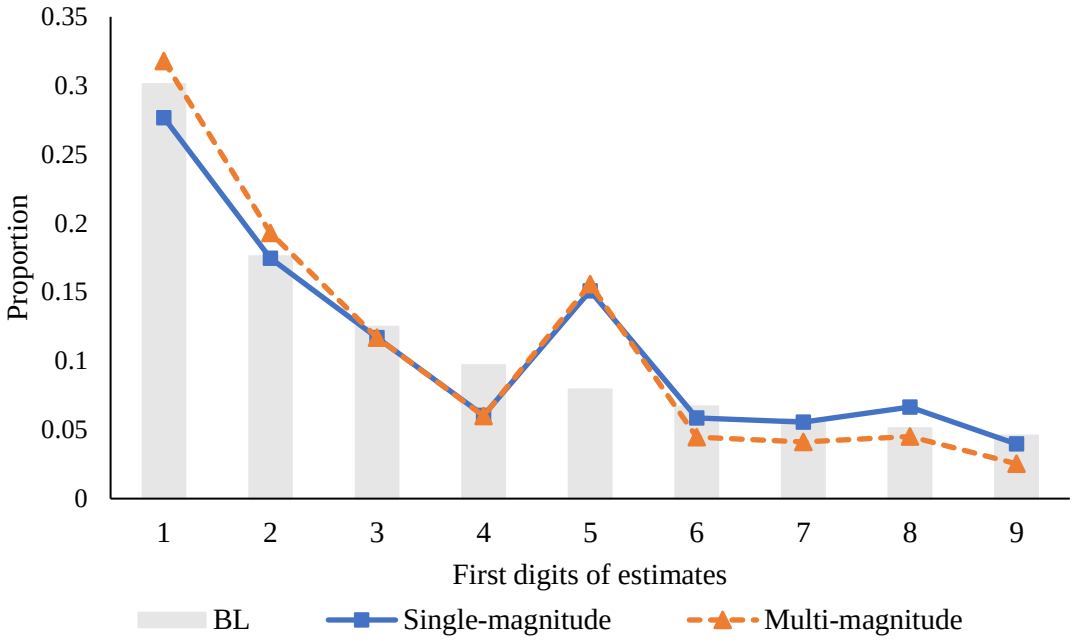


Figure 34. The first-digit proportions of the estimates to Video Estimation tasks in single-magnitude and multi-magnitude conditions, compared to Benford's law in the preliminary study for Exp.4.

Discussion of Preliminary Study for Experiment 4

The finding of this test was the first attempt to demonstrate that Benford bias could be extended to the visual estimation involving multi-magnitude values from various domains. Notably, the estimation of multi-magnitude values yielded a stronger Benford bias than those that had four-digit true values only. Previous research has indicated that datasets in nature confined to a narrow range may disrupt BL (Durtschi, Hillson & Pacini, 2004). Introducing quantities with varied magnitudes naturally broadens the estimation range, which is expected to increase adherence to Benford's distribution. It was also interesting to see the spikes at digit-5, which was absent in estimating the static dots stimuli before. Perhaps this might be an indication of a more difficult task.

However, the influence of the domain was not observed, possibly due to the aggregation effect. In the same domain condition, each individual encountered the same type of target within each block, but the specific domain varied among individuals. For instance,

one participant might consistently estimate quantities of birds, while another would always see skeletons. Consequently, the final pattern observed was based on aggregated estimates across nine domains. A more streamlined approach would involve presenting one domain of targets across all participants in the same domain condition, enabling more straightforward comparisons against the group involving multiple domains.

This preliminary test aimed to investigate whether the Benford bias persists when both the magnitude and the domain of the estimation target varied. The results provided promising support for this hypothesis. Consequently, we are encouraged to proceed to the main testing phase, where we plan to explore the impact of feedback on this newly developed test.

CHAPTER 8: EXPERIMENT 4A

The first three experiments have demonstrated the influence of feedback on the frequencies of the first-digit pattern, showing continuous improvement in the estimation accuracy with respect to the first digits generated. However, its association with the size of answers remained unclear, as the true answers were predominantly selected from values with three-digit lengths. Based on this limitation, my preliminary study primarily tested whether people could also exhibit Benford bias when estimating quantities across magnitudes and domains of targets. It revealed that Benford bias was a robust phenomenon, independent of the magnitude of the values. Thus, the final attempts aimed to examine whether the feedback learning regarding the first-digit distribution could still be observed when estimating multi-magnitude values, aiming to clarify its association with the size of answers.

Despite the preliminary study not demonstrating the domain effect on the first-digit pattern, I included as a manipulated variable in the following test. This decision was informed by the prevalent concept of domain adaptation, often observed and generally regarded as a case of learning transfer. This theory posits that knowledge acquired in one domain can be transferred to a different yet similar target domain (Redko, Morvant, Habrard, Sebban & Bennani, 2019). Thus, we were interested in examining whether the principle of domain adaptation applies to the number estimation across varying domains under the influence of feedback. Although this task only required estimating the quantities of visual targets, introducing different domains of elements within each round could lead to domain-specific noise (Kouw & Loog, 2019). Additional factors, such as similarity, may also influence the extent to which knowledge transfer occurs between different domains. Therefore, it was reasonable to expect that immediate feedback effects might not be as pronounced when estimating values in multiple contexts, compared to a single context situation like the Dots Display tasks. Considering this, comparing results between same-

domain and multi-domain conditions could offer a reasonable explanation in case I fail to observe the expected learning effect on the first-digit pattern. The design of Experiment 4a was a 2 (Feedback vs. Control) × 2 (Different vs. Same domain) between-subject structure with repeated measures. The magnitudes of the presented quantities always varied.

Methods of Experiment 4a

Participants

228 undergraduate students were recruited from SONA, and 226 were included in the final analysis after data screening, with an average age of 19.91 (*SD* = 3.27). Of these, 69.5% identified as female, 29.2% as male, and 1.33% belonged to other gender categories. The two most common first languages among the participants were English (54.9%) and Chinese (31.0%).

Procedure and Materials

Video Estimation. The overall structure was maintained as in Experiment 3, but with the key change of replacing visual elements with video clips. It was a repeated measure design comprising four rounds, each consisting of nine video clips. The clips in the last three rounds repeated those presented in the first round in random order. Participants were asked to estimate the number of targets in each clip. Half were given nine clips from the same domain, while the other half received clips from nine different domains, following the procedures in the preliminary study. Additionally, participants were also randomly allocated to either the feedback condition, where vague feedback was given from Round 2 onwards, or the control group, where no feedback was provided at all. The feedback setup was consistent with the one used in the Dots Displays task. Each of the nine clips presented different true values across magnitudes in random order: 15000, 25000, 35000, 4500, 5500, 6500, 750, 850, and 950, ensuring that each leading digit occurred once. This set of true values deviated from the

conventional design, where smaller leading digits typically correspond to smaller true values. Instead, I paired smaller values with larger leading digits and larger numbers with smaller first digits. The nine items were sorted into three magnitude groups based on their actual quantities: three items with three-digit values, three with four-digit values, and another three with five-digit values. The nine clips were selected from an item pool consisting of a total of eighty-one (nine domains × nine true values) clips (see Appendix P), using the materials developed in the preliminary study.

Questionnaires, general instructions, catch pages and confidence ratings. These components utilised the same materials as those in the preliminary study, with minor textual modifications to align with the updated design in this test.

Results of Experiment 4a

The interaction of two between-subject variables (Feedback or Control × Same or Different Domain) created four cells of conditions. 56 students were allocated to each condition except for the control group with single-domain targets, where the responses of 58 students were analysed.

The within-subject contrast analysis weighted by the proportion of BL failed to report a three-way interaction between the first digits, four rounds and feedback conditions, $F(1, 222) = .124$, $p = .986$. The domain effect was not observed either, $F(1, 222) = 1.181$, $p = .278$. But it found an interaction between the first digits and rounds, $F(1, 222) = 21.505$, $p < .001$, $\eta^2 = .088$, showing that the first-digit distributions were different across rounds. My analysis commenced by investigating the first-digit pattern in the initial round, and we subsequently delved into discussing whether the first-digit pattern evolved in the feedback and control groups during subsequent rounds.

Benford Bias Emerged in Round 1

The first-digit proportions of the estimates in Round 1 were graphed for four different experimental conditions in Figure 35. It is important to note that no feedback was ever provided in Round 1, even in the feedback group. The within-subject contrast weighted by Benford's proportions revealed no significant interaction effect between the first digits, domain groups and feedback conditions, $F(1, 222) = .012$, $p = .881$, indicating that the first-digit pattern in Round 1 was not significantly different across conditions. The BND explained 67.6% of variances in the different domains with feedback, 76.0% in the different domains without feedback, 66.8% in the same domain with feedback and 70.7% in the same domain without feedback. This suggests that all four conditions, irrespective of feedback or domain type, exhibited a strong Benford bias when estimating multi-magnitude values in the initial stage. Notably, a spike at digit-5 was consistently observed.

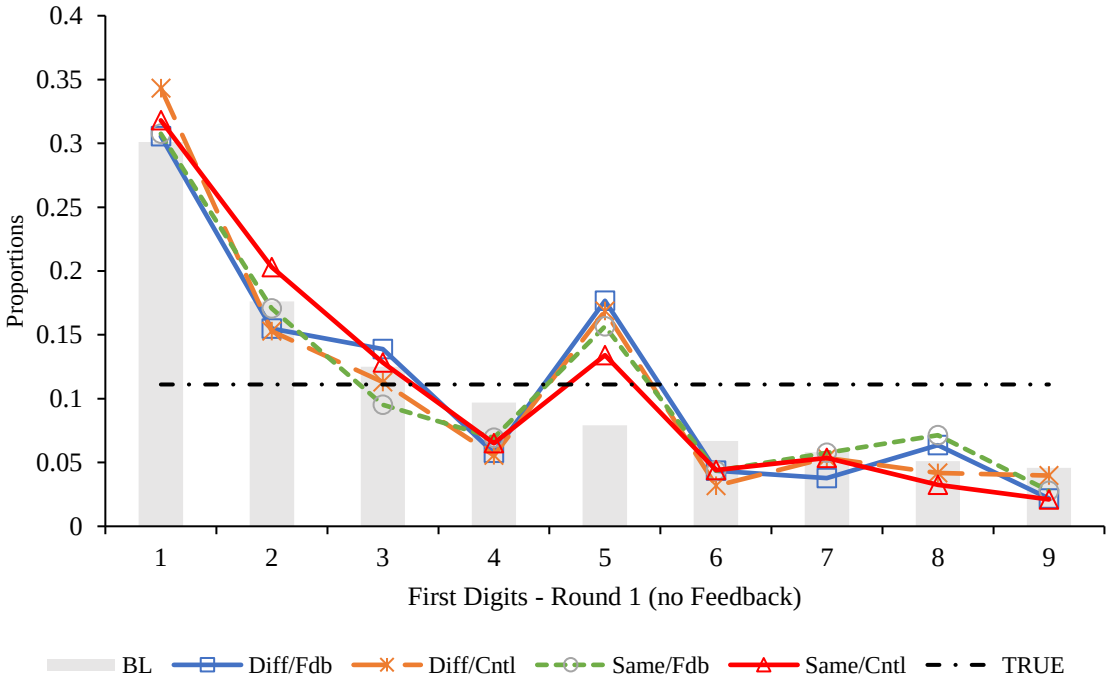


Figure 35. The first-digit proportions of the estimates to the video clips in Round 1where no feedback was ever supplied, separately displayed for the group in the same or different domain, and feedback or control condition, compared to Benford's law in Exp.4a.

First-digit Patterns in Feedback Group

While the general feedback effect was not reported in the test of the three-way interaction, it was observed that the first-digit distribution varied across rounds. To assess the influence of repetitions on Benford bias, the first-digit patterns generated over four rounds were plotted in Figure 36. Thicker lines were used to represent the initial and final rounds for enhanced visual contrast. The within-subject contrast weighted by proportions of BL found a difference in the first-digit patterns between Round 1 and Round 4 in the feedback condition, $F(1,110) = 10.956$, $p = .001$, $\eta^2 = .091$. That is, a marked decrease in the explanatory power of the BND for the variances in estimates, from 67.2% in Round 1 to 48.5% in Round 4, indicating a disruption of the initial Benford bias. Pairwise comparisons revealed a notable suppression in the proportions of smaller leading digits in the final round, such as 1 and 2, while the occurrence of larger first digits (i.e., 6, 7, 8, and 9) increased (all $p < .05$). Moreover, the elevation of digit-5 noticeably decreased ($p < .05$). The observed differences between Round 1 and 4 provided some positive evidence to support that the final pattern displayed a potential moving towards the true distribution under the influence of consistent feedback information.

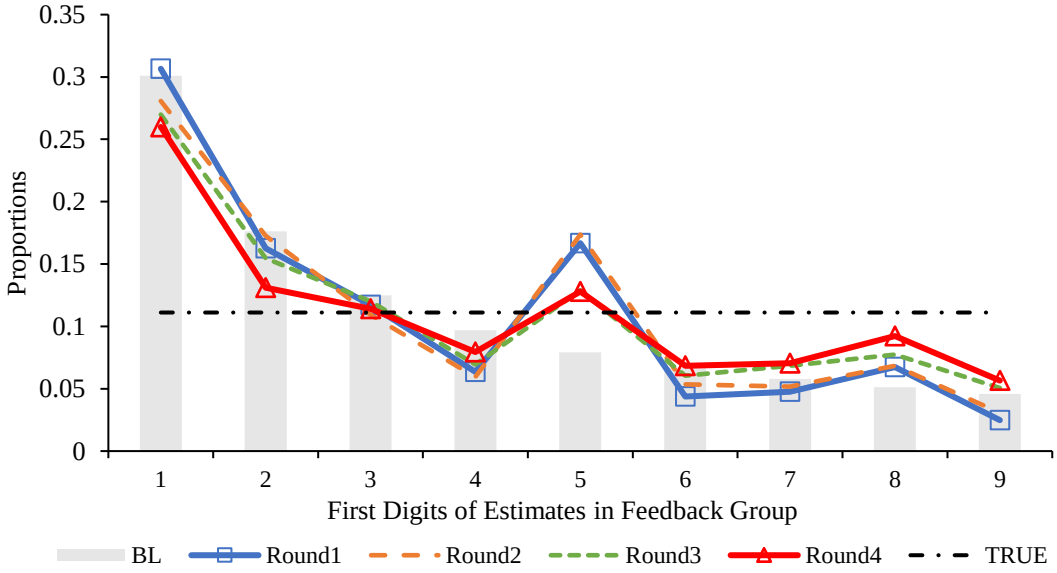


Figure 36. The first-digit proportions of the estimates to the video clips across four rounds in the Feedback condition, compared to Benford's law and the true distribution in Exp.4a.

First-digit Patterns in Control Group

To contrast the shifted pattern obtained in the feedback group, I also analysed the first-digit distribution in the control condition where no feedback was ever provided. Figure 37 displays the proportions of the leading digits produced over four rounds with thick lines in the first and last rounds for better contrast. The within-subject contrast analysis weighted by Benford's proportions also found an interaction between Round 1 and Round 4, $F(1,112) = 10.672$, $p = .001$, $\eta^2 = .087$. The BND accounted for 73.1% variances in Round 1 and reduced to 56.3% in Round 4. Nevertheless, significant changes upon pairwise comparisons were only noted for digits 1 and 8 (both $p < .05$). Hence, in contrast to the feedback condition, the differences observed in the condition group did not offer substantial evidence to suggest that the changes in the final round indicated a trend approaching a uniform distribution.

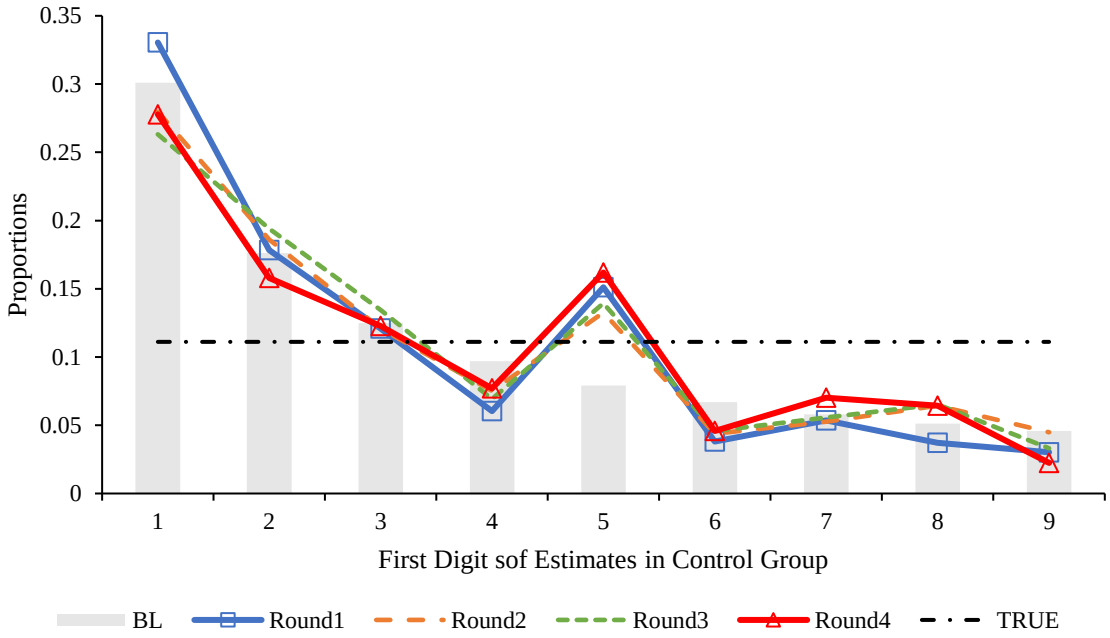


Figure 37. The first-digit proportions of the estimates to the video clips across four rounds in the Control condition without feedback, compared to Benford's law and the true distribution in Exp.4a.

Accuracy Analysis of Full-number

The estimates in the video estimation tasks were strongly skewed. Hence, the medians were also reported to indicate central tendency, as shown in Table 7. All medians deviated from the true values, exhibiting underestimation performance. With the exception of the true value of 15,000, the magnitude suggested by the median corresponded to the size of its respective true value. For example, when the target quantities were in the range of five-digit numbers, the median consistently matched this digit length. The variability of responses grows with the rising magnitude of true values.

Table 7.

Means and medians of estimates to the items with nine true values in Round 1 of Video Estimation in Exp.4a.

True values	Estimates in Round 1		
	Mean	*Median*	*SD*
15000	4494843.022	6000	66515324.885
25000	5116571.137	10000	66809890.005
35000	9688288.248	10000	93971717.751
4500	61182.814	2000	669003.183
5500	72344.066	3000	678727.239
6500	972616.996	3000	9403456.576
750	5996.819	500	47320.726
850	4619.155	500	34833.077
950	4843.230	500	34614.332

To examine the learning effect, I analysed the accuracy of the full-number estimated in terms of the size of errors. However, unlike previous displays of hundreds of dots, the quantities presented in the video clips varied significantly across different scales. To standardise these on a single scale, I converted the absolute differences (*ADs*) between the estimated numbers and the true quantities for each test item into standardised z-scores. Since differences in the first-digit distribution were observed between the first and last rounds, we compared the z-scores of the absolute differences in these two rounds across nine items. This comparison was conducted to determine whether the observed changes can be attributed to improved performance in the estimation of full-number. The repeated measures ANOVA failed to report the interaction between the nine z-scores of *AD*s, rounds and feedback conditions, $F(8,1760) = .290$, $p = .970$. Thus, the accuracy of the full-number estimates in the final round did not generally exhibit enhancement compared to the initial round. The absence of improvement failed to supply evidence supporting the impact of feedback on full-number estimation performance.

Accuracy Analysis of First-Digit Estimates

I also assessed the accuracy of the first digits generated to see if such performance was changed. The repeated-measure ANOVA reported an interaction effect between the rounds and feedback conditions, $F(3,666) = 15.469$, $p < .001$, $\eta^2 = .065$, but not with the domain groups, $F(3,666) = 2.61$, $p = .050$. The pairwise comparison further showed that the first-digit accuracy rates were significantly enhanced in the last two rounds compared to Round 1, regardless of the domain group (all $p < .05$), as illustrated in Figure 38. Conversely, no such improvement was detected in the group without feedback (all $p > .05$).

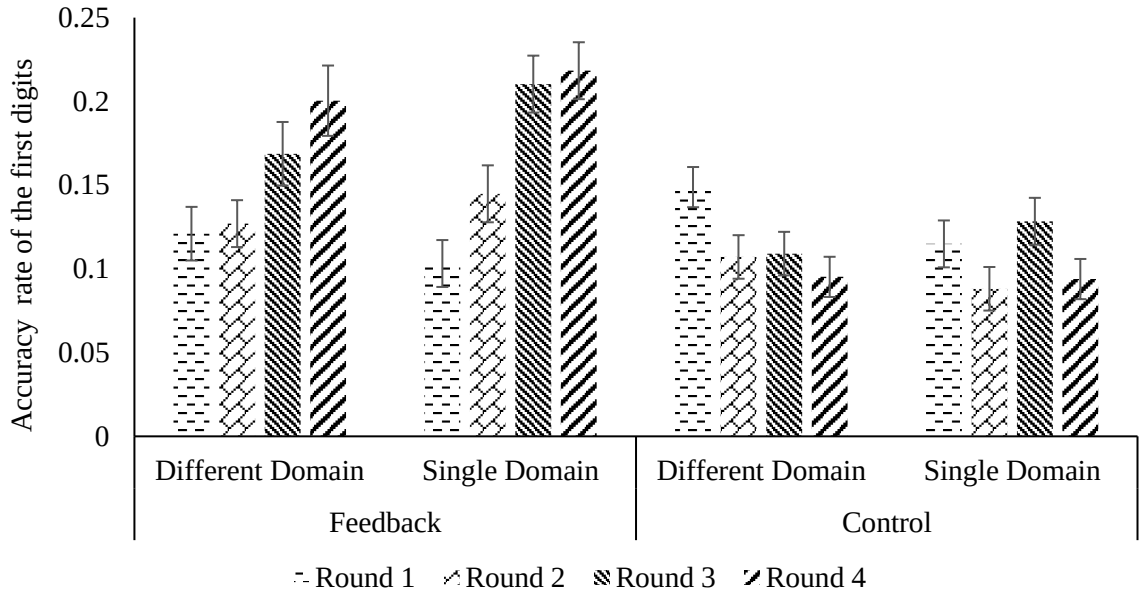


Figure 38. The first-digit accuracy rate of the estimates generated to video clips given feedback or control, and single or different domain conditions in Exp.4a.

Analysis of Confidence

The repeated measures ANOVA did not find the effect of domain or feedback on the confidence ratings (both $p > .05$). The mean confidence for the responses in each round was graphed in Figure 39. Accuracy was not found to correlate with confidence.

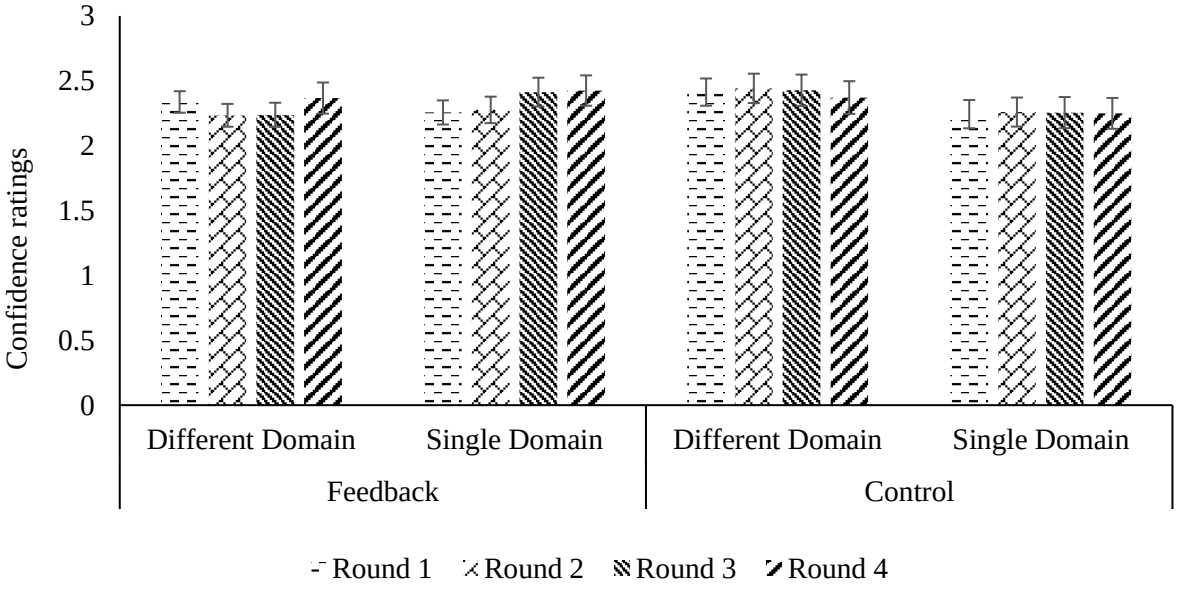


Figure 39. Means of confidence ratings of answers to video clips on a five-point scale across four rounds in Exp.4a.

Discussion of Experiment 4a

Similar to the results from the preliminary test, people initially exhibited Benford bias when estimating the numbers of visual elements with multi-magnitude quantities. Such a phenomenon can even be extended to dynamic targets from multiple domains within a single session. Although the analysis did not report a general effect of feedback, largely due to the changes also detected in the control condition, I still observed some positive evidence to support the impact of feedback on the Benford bias as a result of learning. The data in the feedback condition showed that the smaller leading digits (i.e., 1 and 2) were less frequently produced, while the larger ones (i.e., 6, 7, 8 and 9) became more common after three repetitions, displaying a tendency to approach the true pattern that was uniformly distributed. This was further evident by the improved performance on the first-digit accuracy. Even the control condition (i.e., no feedback group) also reported a different first-digit pattern in Round 4; I argued that such changes did not provide sufficient evidence to show that the shift was approaching the uniform (true) distribution simply based on the changes at digit-1 and digit-8. In addition, unlike the feedback group, the estimation performance was not found to be enhanced in the control condition when no feedback was ever supplied.

Nevertheless, I recognised that the suppressed Benford bias in the video estimation under the influence of feedback was not as rapid and strong as those reported in the Dots Displays with static pictures. Unlike presenting hundreds of dots, I did not obtain an immediate shift in the second round with video estimation tasks, which potentially reflected the difficulty level imposed by varying the magnitude of values. Thus, along with all the past findings, it was evident that the use of vague feedback shall be able to disrupt Benford bias when the true values disobey it, and this effect could be generalised to visual estimation with multi-digit true values across multiple domains of targets, at least with four repetitions.

However, the process of learning was largely delayed when the size of quantities varied substantially.

To understand the changes in Benford bias under the influence of feedback, I also assessed the estimation performance. Unfortunately, there was no observed improvement in the accuracy of full-number estimates from Round 1 to Round 4. Nevertheless, considering only three rounds of feedback were used in the current design, we may not immediately dismiss the possibility of changes with more repetitions in future studies as past research suggested that altering initial regularities may require consistent exposure over time (e.g., Gebhart, Aslin & Newport, 2009; Weiss, Gerfen & Mitchel, 2009). In contrast, participants' estimations of the first digits significantly improved in the later rounds, irrespective of the assigned domain. While the performance in estimating the first digits did not show immediate improvement in the second round as observed in the Dots Displays experiment, it still exhibited a relatively higher level of sensitivity and a quicker response mechanism compared to the full-number estimation performance. Therefore, the enhanced performance in estimating first digits offers some support for the learning effect and is considered a more reliable explanation for the disrupted bias in the feedback condition.

The existing exploration of visual estimation with multi-digit numbers across different domains demonstrated that changes in the first-digit pattern are independent of the magnitude or context of the target values. This finding additionally extends the scope regarding people's sensitivity to the proportions of first digits and adaptability of learning with minimum intervention (i.e., vague feedback). The influence of domain was generally not found in either the first-digit pattern or the estimation performance. This finding offers valuable insights into the framework of domain adaptation, a concept within the broader theories of transfer learning.

In the current design, each student repeatedly went through the same set of nine clips over four rounds, though the order was randomised. This means that once the clips were selected from the pool of eighty-one (nine domains × nine true values) in the first round, the true value remained consistently paired with that particular domain in subsequent blocks. Comments provided by the students indicated a few noticed the repetition of items in the test, even though this was not explicitly mentioned in the instructions. These remarks led us to recognise the need to enhance the experimental design by introducing greater variability and breaking the repetitive pairing pattern between the true values and a particular domain. To address this, I decided to conduct a follow-up study as a second part of this experiment. In this upcoming trial, the nine true values will be randomly paired with nine different domains in each round. This approach shall allow participants to respond to four distinct sets of clips in total, thus eliminating the constraints of repeated association and increasing the variability for different domain conditions.

CHAPTER 9: EXPERIMENT 4B

The final phase of this project aimed to present participants with the ultimate challenge by subjecting them to the most demanding version of the test. Previously, the evidence of a learning effect was observed when presenting quantities across varying magnitudes and domains under the influence of feedback. However, the repetitive nature of the stimuli made it unclear whether participants were learning the stimuli themselves or the quantities associated with them. In the past tests, each participant repeatedly encountered the same set of nine items across four rounds. Even in the different domain condition, a specific true value consistently corresponded to a particular domain in subsequent sessions. For example, participants were continuously presented with quantities like 25,000 birds, 5,500 whales, and 950 skeletons from rounds 1 to 4.

This revised design disrupted the stable association between the true values and their corresponding domains. In each round, the nine domains of targets were randomly reassigned with the nine true values. For instance, a participant might encounter 15,000 crows in the first round, 15,000 cheering men in the second, 15,000 dancers in the third, and 15,000 members of the audience in the last round. While the underlying pattern of the nine true values remained constant, the targets were completely randomised. This randomisation introduced a more demanding task, rendering it impractical for participants to rely on their memory, such as recalling that 1,000 birds was a “close answer” to enhance their performance in subsequent rounds, since 1,000 would be assigned to a different domain target. My final attempt was to assess whether participants could still capture and respond to the true first-digit pattern under the influence of feedback in this dynamic and unpredictable pairing environment.

The design of Experiment 4b had one between-subjects variable, with participants either receiving feedback or being placed in a control group without feedback. This setup included repeated measures for both groups.

Methods of Experiment 4b

Participants

147 undergraduate students were recruited from SONA, but a sample of 142 was included in the final analysis after data screening, with an average age of 19.74 (*SD* = 1.65). Most participants were female, constituting 64.0% of the total sample, while males accounted for 35.0%. There was one participant falling into the other category. More than half of them were English speakers (61.3%).

Procedure and Materials

New Video Estimation. Similar to Experiment 4a, participants were required to estimate the number of targets displayed in nine video clips over four rounds. Within each round, participants were randomly presented with nine clips, each having a true value randomly paired with a different domain-specific target: 15,000 [target 1], 25,000 [target 2], 35,000 [target 3], 4,500 [target 4], 5,500 [target 5], 6,500 [target 6], 750 [target 7], 850 [target 8], and 950 [target 9]. Nine domains were randomised for targets 1-9 in each round. The nine different domain groups included the targets of Audience, Birds, Cheering men, Crows, Dancers, Moving dots, Protestors, Skeletons or Whales. Nine items were selected from a pool of eighty-one clips in total (nine domain × nine true values) used in Experiment 4a (see Appendix P). Participants were randomly assigned to either the feedback or control group. In the feedback group, participants were given vague feedback (i.e., “too high”, “too low”, or “very close”) on their estimation performance starting from the second round, following a similar process used in Experiment 4a. No feedback information was provided in the control condition.

Questionnaires, general instructions, catch pages and confidence ratings. These used the identical materials in Experiment 4a.

Results of Experiment 4b

Seventy-two students were allocated to the feedback condition, while seventy received the control group without feedback.

The within-subject contrast analysis weighted by the proportions of BL failed to report a three-way interaction between the first digits, four rounds and feedback conditions, $F(1, 140) = 2.555$, $p = .112$. But it found an interaction between the first digits and rounds, $F(1, 140) = 15.122$, $p < .001$, $\eta^2 = .097$, showing that the first-digit distributions were different across four rounds. Thus, my analysis specifically examined how the first-digit pattern differed from Round 1 to Round 4 with respect to the degree of fit to BL for the feedback and control group, respectively.

First-digit Pattern in the Feedback Condition

The proportions of the first digits in the estimates generated across four rounds under the influence of feedback were plotted in Figure 40. The within-subject contrast analysis weighted by Benford's proportions reported an interaction between the nine leading digits and four rounds in the feedback condition, $F(1,71) = 13.114$, $p = .001$, $\eta^2 = .156$, suggesting variations in the degree of fit to BL. In the first round, the BND explained 73.9% of the variances in the estimates. This percentage dropped to 58.2% in round 2, 58.6% in round 3, and 57.6% in round 4 when feedback was introduced. Pairwise comparisons further revealed that in the last round, digit-1 was produced less frequently, whereas digits 3, 4, and 6 became more common compared to their occurrence in round 1 (all $p < .05$). It appeared that providing feedback led to some fluctuations at certain digits in the last round, but these was not sufficient to demonstrate that Benford bias began to shift towards the flat distribution. According to the eta-squared reported from the contrast analysis, the patterns in all rounds still showed a reasonable monotonic decline. The spikes at digit-5 remained across all rounds.

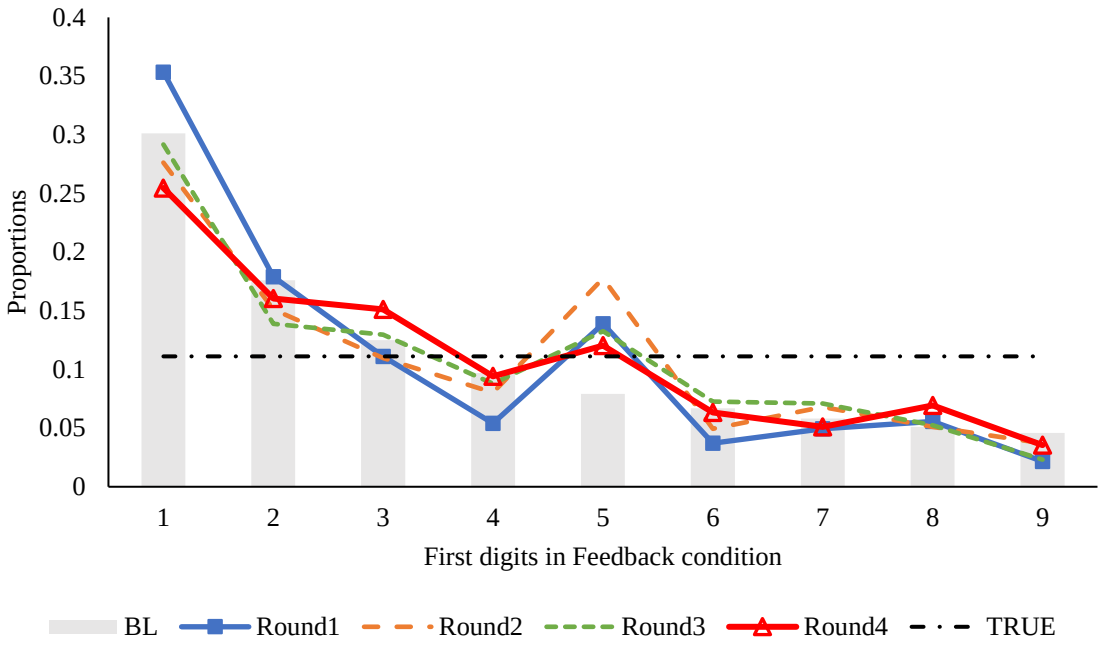


Figure 40. The first-digit distribution of the estimates from the New Video Estimation task across four rounds given the Feedback condition, compared to Benford's law and the true distribution in Exp.4b.

First-digit Pattern in the Control Condition

In contrast to the feedback condition, the first-digit patterns of the estimates produced in the control group, where no feedback was ever supplied, were graphed separately for four rounds in Figure 41. It was clear that the monotonic decline remained relatively stable from round 1 to round 4 within the control condition, as supported by a non-significant interaction between the nine leading digits and four rounds in the within-subject contrast analysis weighted by Benford's proportions, $F(1, 69) = 3.112$, $p = .082$. The BND accounted for 77.3% of the variances in the estimates in round 1, 76.2% in round 2, 67.7% in round 3, and 61.2% in round 4, all showing good agreement. As anticipated, the occurrence of digit-5 remained elevated in this case.

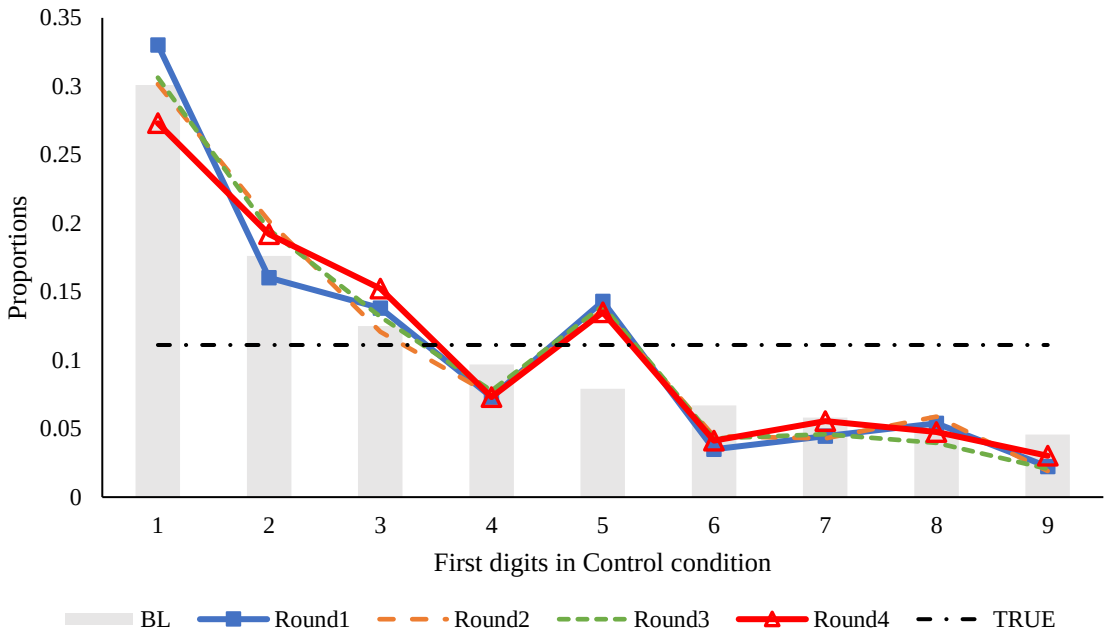


Figure 41. The first-digit distribution of the estimates from the New Video Estimation task across four rounds given the control condition, compared to Benford's law and the true distribution in Exp.4b.

Accuracy Analysis of Full-number

The full-number estimates from the video estimation tasks showed a strong skewness; therefore, medians were also reported to indicate central tendency, as displayed in Table 8. All medians deviated from the true values, exhibiting a pattern of systematic underestimation. The magnitude indicated by the median closely aligned with the size of its corresponding true value. For example, when the target quantity was a three-digit number, the median typically fell within the hundred range.

To examine whether there was a learning effect on the estimation performance, following the past methods, the absolute differences (*ADs*) were calculated for each of the nine items as a measure of the size of errors. Again, to standardise the *ADs* on the same scale, they were converted to standard z-scores. The repeated measures ANOVA did not report a three-way interaction between the z-scores, rounds and feedback condition, $F(24, 1704) =$

1.006, p = .454. Thus, the general impact of feedback was not observed on the accuracy of full-number.

Table 8.

Means and medians of estimates to the items with nine true values in Round 1 of New Video Estimation in Exp.4b.

True values	Estimates in Round 1		
	Mean	*Median*	*SD*
15000	1681889.324	12500	19492003.300
25000	395635.514	21287	2046501.520
35000	78870.511	27000	192465.990
4500	13615.092	3500	71407.374
5500	25412.352	4500	124542.766
6500	4245807.430	5048	50348679.530
750	7048.133	450	75416.684
850	244697.644	500	2900695.792
950	2336.888	650	11145.276

Accuracy Analysis of First-Digit Estimates

In terms of the first-digit accuracy, the repeated measures ANOVA did not show any influence of feedback either, as shown in Figure 42. The interaction between the first-digit accuracy rate across four rounds and the two feedback conditions was not statistically different, $F(3, 420) = .696$, $p = .555$. Unlike the tests in the previous experiments, this

attempt did not find that the estimation performance on the first digits was enhanced in the feedback condition.

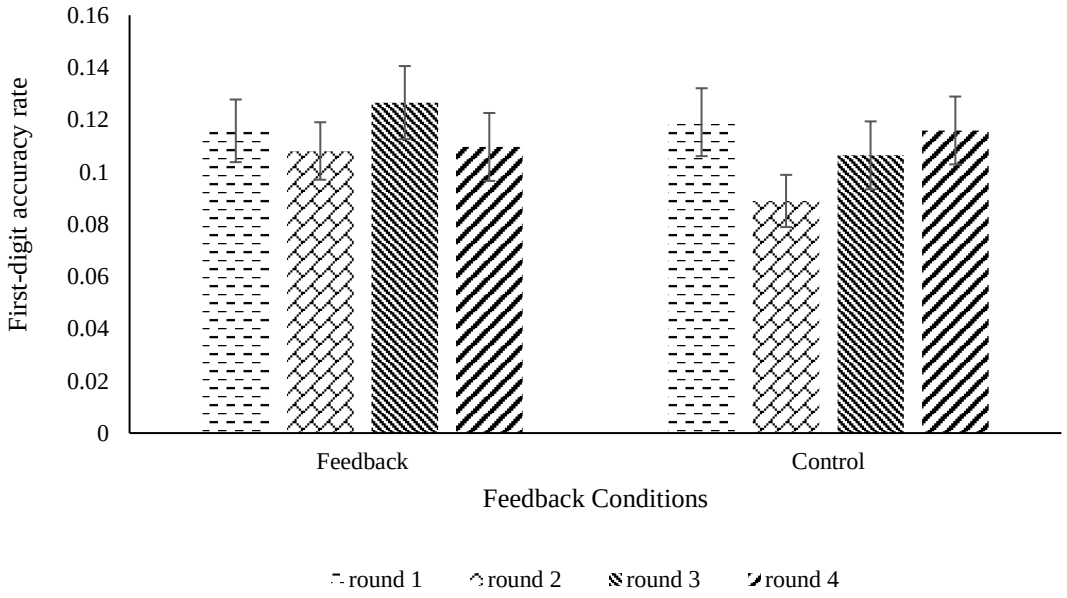


Figure 42. The first-digit accuracy rate of the estimates generated from the New Video Estimation task in the feedback and control conditions in Exp.4b.

Analysis of Confidence

The average confidence ratings across four rounds were graphed separately for both the control and feedback conditions in Figure 43. Receiving feedback in later rounds did not significantly change participants' confidence in their answers based on the repeated measures ANOVA. Those who received feedback exhibited similar levels of confidence in their estimates as those in the control group ($p > .05$). There was no observed correlation between confidence and estimation accuracy. Additionally, the confidence ratings in this test were not significantly different from those in Experiment 4a.

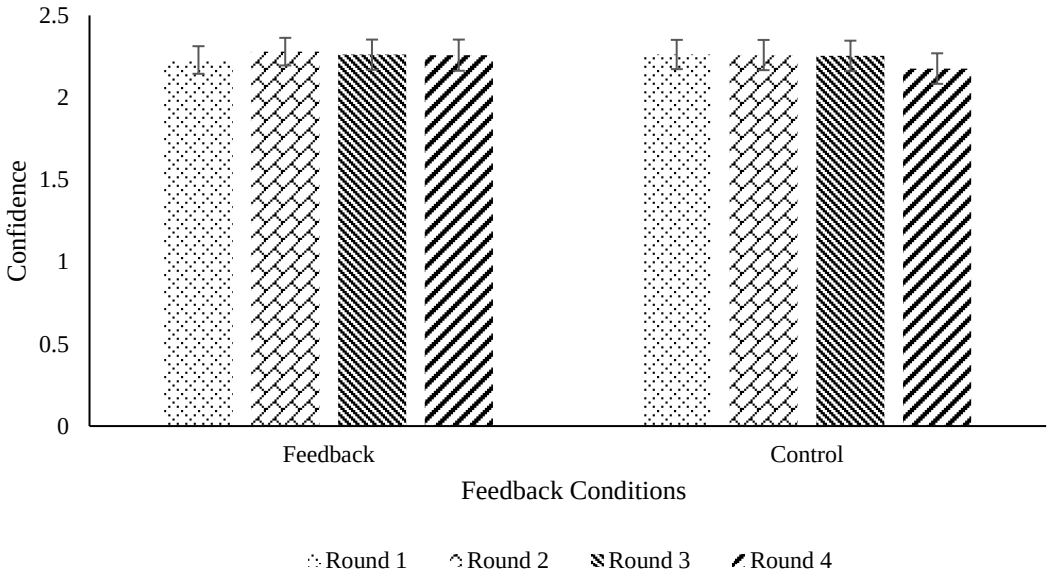


Figure 43. Means of confidence ratings of answers to new video clips on a five-point scale across four rounds under different feedback conditions in Exp.4b.

Discussion of Experiment 4b

The last stage of this project revealed that participants showed a consistent Benford bias in their initial attempts at estimating quantities of various magnitudes using different targets in the video clips. Despite feedback in the later rounds seemingly making some fluctuation at certain digits, it did not lead to a pronounced shift towards a uniform distribution, contrasting with findings from Experiment 4a. This was further reinforced by the accuracy analysis as almost no substantial influence of feedback was found on the estimation performance. Of course, I did not expect immediate adaptation to the true first-digit distribution after merely showing three rounds of feedback, especially considering the potential learning delays caused by the random pairing of targets with true values, which might require more time and exposure due to less familiarity (Gebhart, Aslin & Newport, 2009; Weiss, Gerfen & Mitchel, 2009). This contrasted sharply with the rapid impact of feedback observed in earlier tests. A single change, disrupting the stable pairing between true

values and domains, albeit unrelated to the task request in relation to quantity estimation, was enough to significantly constrain the ability to acquire the underlying regularities.

To further explore whether the absence of the feedback effect in this test was due to the more challenging test structure, I briefly compared the estimation performance regarding first-digit accuracy and confidence ratings between Experiments 4a and 4b. The data and graphs of these two experiments can be referred to in the result section of this chapter and Chapter 8. Surprisingly, participants did not perform worse or exhibit less confidence in this more demanding test than the previous one. While the absence of a learning effect from feedback in the final test might initially appear disappointing, it actually presented a valuable opportunity to enrich the conceptual framework of statistical learning by expanding the test scope and exploring the limitations in the process of number estimation.

CHAPTER 10: GENERAL DISCUSSION

An Overview of Findings Across Four Experiments

The present project aimed to examine the presence of Benford bias emerged from generating meaningful estimates for decision-making by testing the potential hypothesises under the influence of statistical learning in number estimation. Previously, Benford bias was only detected in situations where people were asked to generate unknown numbers, indicating such a phenomenon is exclusively linked to the cognitive process of number estimation. As outlined in Chapter 3, the stronger preference for the smaller leading digits in estimates that approximate BL could be a product of sensitivity to the statistical regularities. This could be due to either the generation of variables following a power law or the acquisition of first-digit distribution. However, these speculations were not mutually exclusive, as we did not expect that a single explanation could fully account for this phenomenon. Each hypothesis offers a testable direction for exploring potential solutions.

I conducted four experiments to assess and validate the reliability and explanatory power of these speculations, using a classic learning paradigm for testing. Particularly, I was interested in whether the established pattern of the first-digit (i.e., Benford bias) could be disrupted or whether it can be altered at all through interaction with the statistical stimuli. The experiments yielded diverse results in how estimation behaviour interacts with statistical information. A summary of the key findings will be outlined based on what has been tested and what remains unanswered.

Absence of Evidence for the Distribution Hypothesis

In this project, the Animal Estimation task was designed to specifically test the Distribution hypothesis. This hypothesis posits that Benford bias should emerge when producing estimates for variables that follow a power law but not for the Gaussian variables

due to sensitivity to different distributional data. Unlike previous studies that presented trivia questions directly (e.g., Burns, 2009; Chi & Burns, 2022), my approach involved exposing participants to exemplars of underlying distributions to facilitate learning about probabilistic relationships in the first stage. As expected, people consistently exhibited the Benford bias when generating the estimates to the group-size related questions for the social animals in the wild. However, when estimating Gaussian variables like weight or lifespan, the first-digit pattern did not deviate from a monotonic decline despite some domain-specific biases, such as spikes at digits 3 and 7 in lifespan estimates. Moreover, the estimated values produced for three types of questions (group-size, weight, lifespan) were similarly distributed, all showing strong positive skewness. This potentially implied that participants may not possess an innate ability to discriminate between different types of distributional data. Overall, the students, in the first attempt, failed to exhibit different estimation behaviours when generating estimates for different variables questioned.

In the second attempt, to enhance the effectiveness of learning the distributional information, students were required to pass a quiz test on the material before proceeding to the estimation tasks. To address the problem of insufficient parameters, I introduced a reference point, either in the form of an average or a starting value, as an anchor. This approach was inspired by the test paradigm used in the studies that reported the role of optimal statistical inferences in cognitive judgment. Despite these updates, the Benford bias was lost though when the average reference point was given, but it did not matter what the underlying distribution was. The distributional patterns of the estimated responses to questions on three different variables showed no significant variation from one another. Such results failed to follow the previous research (e.g., Griffith & Tenenbaum, 2006; Lewandowsky, Griffiths, & Kalish, 2009), which posited that predictions are shaped by accurate prior probabilities. These findings suggested a different narrative: even when

participants were fully informed about the underlying distributions, the application of this statistical knowledge to unfamiliar questions did not align precisely with the expected distributions. For instance, the weight estimates in my study significantly deviated from the expected Gaussian distribution, at least at the aggregated level. Therefore, the outcomes of these two attempts in the animal estimation cast doubt on the reliability of the Distribution hypothesis.

Indications of Changes in Cognitive Processes

The animal estimation task, while not providing conclusive evidence about people's sensitivity to the underlying distributions, did reveal differences in estimation behaviour based on the reference point used. When comparing first-digit patterns between groups given either average or juvenile data, distinct reactions were reported. The group using juvenile data consistently showed Benford bias, whereas those with average data demonstrated noticeable deviations. The analysis of the distributions of numerical answers further sheds light on this phenomenon. Estimates based on juvenile information tended to exhibit a positive skew, aligning more closely with a logarithmic distribution, as opposed to the normal or negative skew observed with average data. Moreover, the mean estimated answers were generally closer to the average references but significantly deviated from the starting point given in the juvenile group. This indicates varying strategies in dealing with unfamiliar situations based on the reference type provided. Regardless of the actual distribution of the questioned variable, estimates based on average data were often insufficiently adjusted. Although deviations were more pronounced in lifespan and group-size questions, they still centred around the average anchors. In contrast, estimates in the juvenile group often reached estimates double, triple, or even ten times the initial starting value, implying a multiplication-based calculation process.

These findings suggest that the presence of reference data significantly shapes estimation behaviour, potentially altering the cognitive process of number generation and resulting in varying first-digit patterns. While the Animal Estimation task was not optimally designed for testing theories related to cognitive processes, it nonetheless provided valuable insights guiding future research. Recognising this potential, we have outlined a few proposals for more targeted investigations into this hypothesis. These recommendations are detailed in the later section on future directions.

Evidence Supporting Sensitivity to the First-Digit Distribution

Despite the lack of evidence showing sensitivity to the underlying distribution of variables, the analysis of statistical learning in relation to the first-digit distribution revealed a promising and robust effect, especially in the Dots Displays tasks. A series of visual estimation tasks not only uncovered that the initial Benford bias could be rapidly disrupted through interacting with the feedback but also progressively shifted towards a more accurate representation of the actual first-digit pattern that deviates from BL. This shift became more pronounced with repeated exposure to feedback, in contrast to the control group exhibiting a stabilised Benford bias without any feedback. The observed shift in the first-digit pattern was consistently supported by accuracy analysis, providing coherent explanations. The estimation performance improved after participants learned to gauge the accuracy of their answers relative to the true answer. Notably, it was found that the performance of first-digit accuracy was more responsive to feedback, indicating a more reliable explanation. For instance, we observed a round-by-round enhancement in first-digit accuracy in the group where the true pattern diverged from BL and favoured larger leading digits. This improvement, facilitated by ongoing feedback, mirrored a consistent evolution in the first-digit pattern throughout the rounds. Conversely, the progression in full-number estimates was primarily observed in the

initial blocks, holding insufficient explanatory power for the alterations in the first-digit pattern in subsequent rounds.

The analysis from the video estimation further disassociated the influence of magnitude and domain by expanding the test scope. The findings supported that the feedback effect was not merely an artifact due to learning the size of the answers, which were limited to a three-digit length. When quantities varied, including hundreds, thousands, and tens of thousands of targets, participants still demonstrated a reasonable shift towards a flatter, more accurate pattern by producing more larger leading digits. Nevertheless, unlike the rapid shift in Benford bias seen in Round 2 when perceiving hundreds of items in the Dots Displays, estimating multi-magnitude quantities presented in video clips did not result in an immediate disruption. The anticipated changes in the video estimation only became noticeable by the last round. Additionally, the monotonic decline was not as significantly suppressed as observed in the Dots Displays, with digit-1 still being overproduced. The delayed and minimised learning effect might be due to the increased difficulty and barriers by showing numerosities across magnitudes, which could hinder the learning pace and ability. This was further supported by accuracy analysis, which revealed that full-number estimation performance remained unchanged over the course of four rounds, with an improvement in first-digit accuracy only becoming evident starting from Round 3.

Introducing domain noise in the final part of the experiments further hindered the learning effect. The method used so far involved repeating a set of nine items under the impact of feedback, where a specific domain was consistently paired with the same quantity value in every test set. However, when the domain could be reassigned to any quantity in subsequent repeats, participants needed to recognise regularities among additional noise. This updated design made it less likely for participants to rely on memory for specific items, such as remembering an estimate of 1,000 birds as a “close” answer. Instead, this required them to

focus solely on quantity perception, extracting the numerosity information per se, such as "1,000," regardless of whether it referred to birds, humans, or whales. Accurate estimation demands the ability to transfer feedback knowledge from a past context (e.g., 1,000 birds) to a new and unfamiliar domain (e.g., 1,000 skeletons), which requires extra effort and time (Aslin, 2017). As anticipated, through providing feedback over three rounds given this most challenging version of the test, I failed to obtain a clear trend showing a shift towards the true distribution despite a few fluctuations at certain digits. However, this does not eliminate the possibility of acquiring such regularities, considering that only three repeats were given in the tests with feedback. Future studies, as proposed in the last section, aim to explore this potential further.

Interestingly, among all the tests, we did not find that estimating the quantities from the same or different targets substantially impacted the first-digit pattern or estimation performance. This contradicted our initial assumption that presenting diverse domains in the same set would delay learning compared to the estimation under the same context. This could be attributed to an aggregation effect. While each participant encountered only one type of target consistently across testing blocks, the aggregate of all estimates demonstrated variations in the context of domains among individuals. In other words, the overall pattern emerged was based on estimates accumulated across nine targets, potentially diluting the domain-specific effect. A more effective approach might involve using the same context for everyone, facilitating a straightforward comparison with multi-domain conditions. Although a warning page was supplied prior to the estimation tasks, outlining that participants can withdraw when concerned about watching various colours and shapes within crowded scenes in the video clips, a few students still reported that these dynamic stimuli sometimes triggered dizziness and sickness. According to the mental number line and Approximate Number System (ANS) theory (Dehaene, 1997), the compressive nature of numerosity representation

might reduce sensitivity to changes in large quantities, hampering differentiation between similar magnitudes. If people struggle to tell the differences, such as between 5,500 and 6,500, the feedback merely reinforces the negative experience of fatigue from viewing the same stimuli repeatedly. In contrast, seeing different targets from one to another might mitigate feelings of failure by shifting to a new context for variety and references. This speculation was supported by more positive comments collected from the different domain groups than those from the single domain.

It was encouraging to see that Benford bias could be disturbed in most estimation tasks when perceiving the large numerosity with minimal information supplied in the feedback. These shifts highlighted the role of people's sensitivity to the first-digit distributions. On the other hand, such findings also conveyed important messages about the constraints in the ability or speed to learn, such as the magnitude of perceived numerosity and the noises in the target domain paired.

Summary of Project Findings

The existing tests provided inadequate evidence to demonstrate that people could acquire and utilise the underlying distributions of variables when producing answers to factual questions. However, indirect evidence emerged supporting the notion that Benford bias might originate from cognitive processes in number estimation, influenced by the reference anchors supplied. Finally, the results from the visual estimation tasks consistently captured people's sensitivity to the first-digit distribution through interaction with feedback, which seems to be a promising and reliable explanation for shifted Benford bias.

General Implications

This project focused on examining the relationship between the acquisition of statistical regularities and behavioural patterns in number estimation, especially concerning

the first-digit distribution. Considering the summarised findings, my goal was to demonstrate how this project enriches the fields of Psychology, specifically in understanding the mechanisms of statistical learning and number estimation.

Implications for the Framework of Statistical Learning

The set of four experiments distinctively contrasted the sensitivity to different types of distributional information. While participants did not adhere to the underlying distribution of variables through exposure, such as animal weights, they effectively grasped and adapted to the actual first-digit distribution when provided with feedback. This section aims to delve into the potential implications of these findings, marked by both expected and unexpected results, on the conceptual framework of statistical learning.

Notes From Unexpected Results

The conceptual framework of statistical acquisition was primarily in line with the construct of implicit learning, which addressed the role of mere exposure without explicit instructions or awareness of the statistical pattern acquired (Conway & Christiansen, 2005). In the animal estimation, mere exposure to the knowledge of distributional data, even supplemented with incentive-driven quiz sessions, failed to produce the expected learning pattern. Most studies on statistical learning typically created several simple associations between pairs of novel shapes or sounds to simulate the basic regularities surrounding us (Aslin, 2017). However, when it comes to distributions concerning the probabilities of multiple outcomes, the complexity of this underlying structure might hinder the ability to learn. Prior studies indicate that understanding probabilities and dataset distributions is inherently challenging, requiring consistent effort over time, as observed in undergraduate students (e.g., Garfield & Ahlgren, 1988). This failure, as opposed to previous positive findings, highlighted the effectiveness and limitations of various stimuli in promoting

statistical learning. It outlined the importance of distinguishing between the complexity levels of the materials exposed to.

Nevertheless, the absence of adherence to certain regularities does not necessarily mean a failure to acquire knowledge. There might be a gap between knowing and applying, an area still under-explored in Psychology. Examples of probability learning, focusing on behavioural changes over time, suggested that the different strategies used by adults and children in transitioning from probability matching to maximising rewards were a result of choices rather than the knowledge of associations (Plate, Shutts, Green & Pollak, 2018). Tests revealed that children could accurately identify the option with the highest reward, indicating that the primary difference lay in the amount of information they demanded before deciding to change their approach. Likewise, assuming that participants had adequate knowledge as indicated by the quiz, investigating the conditions under which they apply or do not apply this knowledge could provide deeper insights. Sensitivity to underlying distributions was often reported from responses estimated for everyday activities, like baking times (Griffiths & Tenenbaum, 2006). However, anomalies were also detected, such as in predictions about pharaohs' reigns, where people's predictions failed to follow Bayesian calculations based on optimal statistical judgment.

This made a clear contrast to the strategies employed when facing different levels of uncertainty. In highly unfamiliar situations like ancient Egypt, people tended to rely more on making analogies from well-known categories, such as modern monarchs but often did so without adequately adjusting for specific factors. The current tasks, similar to the items asking about the duration of a pharaoh's reigns, delved into factual knowledge about wild animals, a topic unfamiliar to most. In such uncertain scenarios, participants might not utilise newly acquired data about human-related topics. Instead, they might opt for analogical reasoning based on more familiar animal instances despite not having direct knowledge to

precisely define the parameters of that distribution. Employing analogy-based prediction strategies can be a practical way to navigate judgments in areas beyond one's immediate knowledge and experience, thereby mitigating the fear of the unknown (Griffiths & Tenenbaum, 2006). Therefore, the choice of strategy, influenced by the degree of familiarity or uncertainty, may serve as a bridge between intuitive judgment and statistical thinking.

Sensitivity to the First Digits and Feedback Effect

The findings on estimating the quantities of visual stimuli revealed that individuals are remarkably responsive to actual first-digit distributions and capable of adjusting their behaviour based on feedback. This section discussed the implications of these positive results, covering the innovative aspects of the design, the active responses to the first-digit place, and the incorporation of feedback in my exploration of statistical influence.

Traditionally, statistical or implicit learning research has focused on artificial grammar or vocabulary tests, exposing participants to a continuous stream of linguistic stimuli. These tests usually presented materials combining pairs or triplets of targets for exploring the basic associations and frequency of pairing (Aslin, 2017). In contrast, the present study broadens this approach by examining distributional information related to a series of leading digits, expanding the range of associational probabilities among nine options. This method required not only observing distributional data with multiple outcomes but also introducing quantity estimation as a new dimension in statistical learning. Unlike conventional methods that focused on perceiving underlying patterns in artificially created linguistic or shape patterns, my project sought to uncover universal regularities present in real-life scenarios, bridging laboratory experimental science with the natural laws observed in the real world.

A key finding in this set of experiments was the rapid adaptation of the first-digit pattern, accompanied by improved estimation accuracy, particularly for the first digits of the

estimates. This enhancement in accuracy is more pronounced in the leading digit places than in full-number estimates. Reports indicated that first-digit accuracy could boost to 30.0% in the final round, following consistent feedback, well above mere chance. Studies like those by Lai et al. (2018) have demonstrated that people find it important to talk about the first digit of a number, known as the left digit effect, where people disproportionately attend to the leftmost digit while making judgments or decisions. Feedback hints like “very close” could imply proximity to the correct order of magnitude, but students shall also determine the first digit (Shouse, 2023). Consequently, it was anticipated that most adjustments and reassessments would primarily target the leading digits, as this is where improvement, prompted by such straightforward feedback, is most likely to be initially observed. This sensitivity to changes in the first digit may offer insights for studies previously unable to detect learning effects. I suggest re-examining the performance based on the left digits of numbers generated rather than the full-number estimates, which might help uncover the hidden evidence.

Another significant observation from these studies is the profound impact of feedback. The effectiveness of feedback in enhancing estimation performance has been well-documented in various studies (e.g., Izard & Deane, 2008; Roy et al., 2008). A crucial aspect of learning, however, concerns the extent of information required in feedback to facilitate meaningful change. Previous studies on visual quantity estimation often provided prototypes (Minturn & Reese, 1951), such as showing what 50 dots look like. In contrast, the feedback information was minimal yet powerful. Participants were able to accurately interpret and incorporate simple cues like “higher” or “lower” into their judgments. Interestingly, the impact of feedback was more pronounced and immediate with smaller quantities, such as hundreds of dots, compared to larger ranges shown in videos, spanning hundreds to thousands. Yet, the continuous shifting of the first-digit pattern, particularly in adapting to

distributions that deviate from BL, was still remarkable. To gain a deeper understanding of the feedback effect, future research could focus on designing additional tests to clarify what participants are actually learning from the prompts. It is possible that, in the dot displays, participants were merely familiarizing themselves with the stimuli rather than learning the actual quantities, as the same set of dot patterns was used repeatedly. Introducing variability in the dot patterns while keeping the underlying true values consistent could help determine what participants are genuinely capturing through their interaction with feedback. Additionally, if participants become sensitive to the true first-digit distribution under feedback conditions, discontinuing the feedback or alternating between two different sets of true values (e.g., from a Benford-like to an anti-Benford distribution) during transfer phases may provide further insights into the learning mechanisms at play.

Despite this progress under feedback, participants often remained unaware of their improvement. Participants in the feedback group who demonstrated improved accuracy did not exhibit stronger confidence in their performance in most cases. Also, no consistent correlation was reported between the confidence ratings and estimation performance. These results, aligning with the "guessing criterion" and "zero-correlation criterion" discussed in Chapter 1, supported the notion of implicit statistical learning, where individuals subconsciously assimilate patterns of associations without conscious awareness (Conway & Christiansen, 2005; Turk-Browne et al., 2005While we did not intend to test this, these results may offer some insights into the connection between individuals' awareness and their actual performance.

Implications for the Number Estimation

The phenomenon of Benford bias in psychology was first noted through the analysis of estimates made in response to general knowledge or factual questions, such as a country's electricity consumption (Burns, 2009) or the number of items in a visual scene like jellybeans

in a picture (Chi & Burns, 2022). These studies laid the groundwork for developing the test materials for the current project, which were designed to further explore the causes of the phenomenon. As discussed in Chapter 2, the cognitive process explored so far does not apply to all the estimation behaviour, leading us to investigate what such findings might add to the understanding of number estimation. Beyond the impact of statistical learning, as previously mentioned, this section also aims to provide an overview of other results and how they might shed light on the ways people formulate estimates for decision-making.

Estimation of Factual Questions

The animal estimation task in the current study involved prompting participants with factual questions about aspects like the weight and group-size of living creatures. While learning the distributions of these variables showed less impact, the use of a reference point notably influenced estimation behaviours. Presenting the average value of a variable led to a more normalised distribution shape with fewer variations, unlike the juvenile reference, which resulted in a strongly positive skew. Additionally, most mean estimates hovered around the provided average anchor, whereas those from the juvenile group were often double, triple, or even ten times the starting point. This suggests that people adjust their expectations and processes based on the information they have. Multiplication was more common when projecting future values from a known starting point, while addition or subtraction was more frequently associated with average-based estimations. Such shifts in estimation behaviour indicate that cognitive judgment is significantly influenced by anchored information. The expected distribution could be primed based on inferences drawn from specific reference types. For instance, an average score is often associated with a normal distribution, while a starting point typically suggests a growth pattern following a power law. Stereotypical reasoning about measures of central tendency is widespread among students from primary to tertiary education (Ismail & Chan, 2015). These stereotypes often align with

everyday assumptions, such as the notion that an average score is its midpoint (Mokros & Russell, 1995), which may not always be true.

This reliance on stereotypes in reasoning implies that people tend to favour familiar information over newly acquired factual evidence, especially in uncertain situations. Assuming that participants who passed the quiz had a basic understanding of how each variable should be distributed, they still opted for solutions based on misinterpreted distributions from specific reference types. Whether consciously or unconsciously, they disregarded the factual information presented, showing a preference for stereotypical reasoning used in everyday judgment. This suggests that it is challenging to override the established norms when responding to general knowledge or factual questions, even when confronted with explicit facts. However, similar to representative bias or base-rate neglect (Kahneman & Tversky, 1972), this approach might be a practical way to navigate uncertainty under time pressure, especially when the newly provided factual information requires time and effort to fully comprehend and apply.

Estimation of Quantities in Visual Elements

Distinct from the Animal Estimation task, which involved general knowledge questions, the Dots Displays and Video Estimations primarily focused on the visual perception of quantities. Beyond understanding how quantity estimation is influenced by the statistical acquisition of first-digit distribution, I also uncovered several phenomena related to the mechanisms of visual perception of numerosity.

The estimation range in the visual perception tasks was significantly broadened, allowing observations beyond the typical limit of 300 targets. It was observed that presenting larger numerosities often led to a more pronounced underestimation in the mean values generated for each dot display, accompanied by greater variability in estimates. This aligns with the Approximate Number System (ANS) theory, which posits that numerosity

representation is logarithmic for large sets of elements (Dehaene, 1997; Booth & Siegler, 2006). The ability to distinguish between two quantities depends on their ratio, not the absolute difference. For example, it is much easier to distinguish between 350 and 650 dots (ratio 7:13) than between 650 and 950 dots (ratio 13:19), even though the absolute difference remains constant. Similar to the notion of the mental number line, this compressive nature of numerosity perception might transfer to symbolic representation during the quantification of items, often resulting in underestimation (Chesney & Matthews, 2018; Crollen, Castronovo & Seron, 2011). Consistent with prior research, systematic underestimation was observed in the free estimation tasks when participants were explicitly asked to generate a number for large numerosities. The nature of free estimation tasks allowed participants to determine quantities without being constrained by predefined numerical limits, replicating the findings from earlier studies.

However, judgments of numerosity are also susceptible to the physical characteristics of the visual elements. In the video estimation featuring animated 3D objects, the pattern of underestimation persisted, especially when using the median as a measure of central tendency due to strong skewness and outliers. While numerosity judgments related to 3D objects often resulted in overestimation (Aida, Kusano, Shimono & Tam, 2015), dynamic stimuli were usually associated with underestimation (Au & Watanabe, 2013; Poom, Lindskog, Winman & van den Berg, 2019). The video materials, integrating both dynamic and 3D properties, replicated the underestimation trend, providing further evidence for the ANS mapping theory in free estimation contexts.

Nevertheless, not all tasks resulted in mean estimates significantly lower than the actual quantities. In the second experiment, when the first digits of true values were uniformly distributed, estimates in the Dots Displays closely matched the actual quantities presented. This could be attributed to the “wisdom of crowds” effect, where collective input

from multiple independent sources balances personal biases and extreme responses, resulting in a more accurate average (Yi, Steyvers, Lee & Dry, 2012). Conversely, estimates in the anti-Benford group were consistently underestimated. It was argued that the observed differences in estimation accuracy could be related to the range of quantities displayed. The uniform group displayed numerosities typically ranging from 250 to 1,050, whereas the anti-Benford condition presented a narrower range, predominantly spanning the higher hundreds, from 450 to 990. According to ANS, larger numerosities that are closer together are more challenging to discriminate; hence, a broader range, as in the uniform condition, allows for better discrimination and more accurate collective judgments. Interestingly, these differences in estimation accuracy were not replicated in the third experiment with a larger sample size, hinting at possible limitations in applying collective judgment principles like the wisdom of crowds in certain contexts.

Regularities in Nature vs. Intuitive Judgment

The literature review in Chapter 2 regarding the cognitive mechanism of number estimation revisits a debate in psychology: Is behaviour primarily guided by regularities present in the environment or by intuitive judgment? My study has shown that estimation can be influenced by the regularities of first-digit distribution given the feedback information, yet there are also indications that intuitive reasoning plays a role.

Beyond the typical monotonic decline of the Benford bias, a notable pattern was the recurrent elevation of digit-5 in estimates. This aligns with observations from the reports in the previous studies (e.g., Burns, 2009; Diekmann, 2007), where Benford bias was first demonstrated in human data. Burns (2009) speculated that the over-represented digit-5 might be a common choice in uncertain situations, as it represents half of the magnitude. Although digit-5 was not always the secondary peak in these datasets, the spike was consistently evident in most instances where the true values spanned across a wider range of magnitudes.

This was particularly noticeable when estimating weights, lifespans, and group-size of animals or quantities in the video clips. When faced with greater uncertainty, individuals seem more likely to use a modal number, such as digit-5, as a solution. Studies on the utility of prominent numbers suggest that mental shortcuts like halving are commonly employed as straightforward calculations that meet the need for cognitive ease (Converse & Dennis, 2018).

Apart from the elevation of digit-5, specific bias was detected to link to specific questions asked. First digits 3 and 7, for example, were elevated in lifespan-related questions, reflecting common estimates of days equivalent to one, two, or twenty years. Additionally, despite feedback facilitating a shift towards accurate distributions, digits 1 and 9 showed resistance to change, indicating a natural inclination towards rounding. Such reactions particularly agreed with the notion of relying on heuristic cues in unfamiliar situations, given the benefit of adopting mental shortcuts (Honda, Kagawa & Shirasunam, 2022).

The impact of reference points in animal estimation further illustrates heuristic thinking. The reliance on stereotypical reasoning based on average or juvenile anchors often skewed away from factual information about true probabilities. In line with the observation of base-rate neglect, judgments are more influenced by the perceived characteristics of a category rather than by consensus data when faced with uncertainty.

In conclusion, while the existing findings exhibited supporting evidence for statistical acquisition, such as adapting to first-digit distribution changes due to feedback, intuitive and heuristic judgments continued to subtly affect decision-making. Thus, this project highlights that the estimation pattern cannot be solely attributed to a single factor. Instead, it demonstrates an inherent interplay and competitive relationship between these elements, underling the unique complexity of human behaviour.

Why Does Benford Bias Matter?

While the main aim of this project is to uncover the underlying reasons for Benford bias emerged from number estimation, it remains crucial to clarify the significance and applicability of investigating this bias within the realm of psychology. Put simply, what value does the exploration of this bias bring to the field? To address this question, this section outlines arguments from three distinct perspectives, emphasising the theoretical and practical implications for the investigation of this bias.

Distinctive Patterns from Number Estimation

The Benford bias has been posited as a psychological phenomenon that reflects a strong preference for smaller leading digits in numerical responses, specifically in number generation tasks. This observation has been extended to various forms of estimation tasks, such as answering factual or trivia questions or approximating quantities of pictured items with multi-digit values across diverse contexts. Remarkably, this project has successfully disrupted this well-established pattern through laboratory interventions, thereby establishing a valid link between universal models and specific human judgments. Unlike the use of artificial regularities shortly created in the test of statistical learning, the integration of BL as an analytical tool creates a unique intersection between real-world regularities and short-term experimental manipulations. This also enhances understanding of how individuals generate numbers for decision-making.

Despite the monotonic decline exhibited by the Benford bias, the elevation of digit-5 was also another robust feature associated with estimation behaviour. The spikes at five were sometimes observed as the secondary peak (e.g., Burns, 2009) that strongly diverged from the expected value. A bias in favour of digit-5 has been documented in many real-life scenarios. A recent intensive examination of off-balance sheet records from 22 commercial banks highlighted the prevalent use of the number 5 as the initial digit due to excessive rounding

and intentional control of percentages (Parnes, 2022). Such deviation and preferences also signalled a unique aspect of human judgment, distinguishing it from natural laws.

The Benford bias replicated in the present study, including the frequent elevation of digit-5, shows that humans can mimic patterns observed in natural settings, such as Benford's law. However, they do not strictly adhere to it, displaying unique characteristics that set them apart from the theoretical expectation. These patterns may hint at an underlying pattern guiding individuals when navigating uncertainty, especially when estimating unknown values. Given the ubiquity of estimation tasks, this project could provide additional insights into the norms established from healthy populations by researchers in perception and clinical settings who wish to utilise these estimates as a measure of cognitive abilities.

Benford Bias and Fraud Detection

A key point of concern regarding the reliability of Benford's law as a lie detector centres on the possibility of humans naturally exhibiting a bias towards smaller first digits, a notion that has not been thoroughly examined in relation to human behaviour in the past decades. Current research suggests that this first-digit phenomenon could be manifested in psychological data, particularly when estimates are being produced for decision-making. The strong preferences towards the smaller leading digits of numerical responses have been constantly detected and replicated since Diekmann (2007) and Burns (2009) when asking people to produce numbers to general knowledge and factual questions or quantities viewed in the visual scenes. Consequently, the validity of utilising BL in fraud detection, as proposed by Nigrini (1992), warrants further investigation. Departure from BL may suggest fraud, but adherence to it does not necessarily indicate the absence of fraud.

Observations of consistent leading digit patterns in psychological studies questioned the fundamental belief that deviations from Benford's distribution in financial data are largely due to human interference. This belief has been extensively adopted in accounting

and auditing practices. Nevertheless, anomalies like an elevation of the leading digit-5 could provide additional references for assessing risks associated with fraud due to the overproduction of first-digit 5 observed through collective generation behaviour. The inclusion of this additional parameter in screening with the first-digit law might facilitate a more holistic approach to validating the credibility of datasets. Recognising how human-generated values diverge from natural regularities might offer valuable insights, particularly when the effectiveness of Benford's frequencies is challenged in addressing fraudulent activities.

Benford Bias - A Cognitive Bias?

The nature of Benford's law highlights a pronounced bias towards smaller leading digits when generating unknown values. This raises the question of whether such a pattern might distort responses in the same way a typical cognitive bias does, potentially impairing the accuracy of predicted answers. This is particularly relevant given the prevalent use of similar paradigms in perception (i.e., dot displays) and clinical settings (i.e., CET items), inevitably leading to questions about the potential consequences of these preferences on numerical estimates.

Cognitive biases are typically characterised as systematic deviations from norms or rationality during decision-making and judgmental processes (Haselton, Nettle & Andrews, 2005). While these biases can sometimes be beneficial by offering mental shortcuts that enhance efficiency and effectiveness in problem-solving (Shah, Oppenheimer & Daniel, 2008), much of the work attempted to disclose their pitfalls, particularly how they can cloud rational choices and judgments.

The current dataset did not provide evidence to support the idea that a stronger bias towards smaller first digits correlated with poorer task performance. In fact, accuracy was quite impressive sometimes, as seen in Experiment 2, where all mean estimates in the

uniform group were close to the true values. This could be attributed to the aggregation effect, similar to the wisdom of crowds. While the leftmost digit of a number is undoubtedly crucial in determining the precision of a numerical response, other factors, particularly the number magnitude, also play a significant role.

We recognise that referring to this psychological phenomenon as “Benford bias” could potentially be misconstrued as implying a detrimental effect on estimation performance. However, while underestimation was a common occurrence in these studies, much like in other tasks involving the perception of numerosity, I did not find conclusive evidence linking a stronger bias to a decline in performance. Hence, further systematic research is essential to comprehensively clarify its relationship with estimation accuracy.

Limitations

The datasets in this study were sourced from online responses aimed at increasing statistical power. However, unlike supervised experiments in a laboratory setting, participants might not always be fully dedicated to the tasks. This concern is shared by earlier studies, such as Oppenheimer, Meyvis and Davidenko (2009), who found that over forty-six percent of their participants failed to comply with the request from instructional manipulation checks provided in their tasks. To address this, a catch page was in cooperated into the online materials to test if participants were really paying attention to the tasks and demonstrating a minimum level of engagement. On this page, they were instructed to click the title rather than the “Continue” button to proceed, which was explicitly stated at the end of the instructions. Unfortunately, these findings echoed past studies, with many participants failing to follow the instructions given. Across four separate experiments, the proportion of students who precisely followed the catch page instructions without triggering any error warnings only varied between 26.1% to 41.5%, significantly lower than half of the sample population.

Additionally, in each of the experiments, at least one participant was found to fail more than nine times, receiving excessive repeats of the warnings while ignoring them. Although those outliers were removed, such extreme cases suggested either reckless behaviour that disregards the instructions or inadequate language skills to complete English tasks.

A noticeable pattern of potential invalid responses was also observed in the Animal Estimation tasks, particularly among participants assigned to the Juvenile group. These tasks presented a starting point for the weight, age, or group-size of a young animal in a picture and asked participants to predict the weight at maturity, total lifespan, and final group-size after gathering in the wild. A reasonable answer is always expected to be larger than the starting point, given a basic understanding of growth, though one might argue this point, especially regarding the number of animals gathered in the wild. However, for the total lifespan question, it would be clearly anticipated that the predicted lifespan should not be less than the current age. Nonetheless, I noted that forty students produced a total lifespan estimate smaller than the young animal's current age at least once among nine items. Among these, six students made such irrational responses in more than half of their answers, leading to the complete exclusion of their lifespan data for nine animals.

With regard to the validity of the responses in Animal Estimation tasks, the answers to the quiz test regarding the statistical information acquired during exposure stages might also suggest the difficulty of understanding the distributions, which potentially had an impact on the effectiveness of acquiring statistics. Learning about probabilities and how datasets are distributed is essential for undergraduate students enrolled in statistical courses. However, this skill has been shown to be challenging to master, as evidenced by previous research (e.g., Garfield & Ahlgren, 1988); thus, it often leads to feelings of anxiety and distress (Peiró-Signes, Trull, Segarra-Oña & García-Díaz, 2021). After being exposed to the basic concepts associated with normal, skewed, and logarithmic distributions through human data examples,

less than 20.0% of students successfully passed all six quiz items on the first attempt. The quiz was designed to align with a threshold level of understanding, assuming that recalling simple contexts, such as recognising that weight follows a normal distribution while age is skewed, would enable students to achieve correct answers. Each student was given a maximum of five attempts, but it took them more than two and a half attempts on average to correctly answer all the quiz items before proceeding to the estimation tasks.

As the final estimation test involved new video clip materials, student comments provided additional insights into how they perceived the limitations of such testing materials. Some participants encountered technical issues commonly found during online tests, such as videos not loading properly or lagging. These issues could have impacted their experience and performance. Some participants found the task less engaging and repetitive as it progressed, affecting their attention and interest. However, others saw the repetition as beneficial for improving their estimation skills. Certain stimuli, such as skeletons or dense crowds, caused anxiety or discomfort for some participants, with physical or emotional reactions like dizziness or headaches reported, especially when the stimuli involved spinning or dense groups. I recognise the need to refine the video presentation as making a smooth display with numerous elements within seconds is challenging and requires constant adjustments in parameters like rotation speed. Participants employed varied strategies in their estimation, from imagining themselves in a similar situation at a stadium to counting methods. One participant even questioned whether their estimates were influenced by the presentation method.

Considering the factors that could potentially impact the reliability and validity of the collected responses, I developed a data cleaning guide that tries to strike a balance between preserving data volume and ensuring data quality. However, this cannot be fully assured to minimise the impact of undetected anomalies, such as the risk of participants using external

information to answer questions about animal estimates. Thus, a reasonable level of supervision (i.e., laboratory experiments) is recommended for future studies to ensure valid attempts and genuine responses. Additionally, modifications to the video materials may be necessary to enhance the user experience and facilitate the presentation of large quantities of estimation targets. The limitations highlighted in this section reveal the possible noise factors in the current research. Although it is challenging to precisely quantify the extent of noise in the online test, I still obtained the anticipated patterns in most cases.

Future Directions

In light of the unexpected results observed in this project, along with the discussed limitations, the current project also outlines several viable avenues for future research. These directions will not only offer testable solutions for aspects that have not been thoroughly explored but will also enhance the theoretical framework and practical implications related to statistical learning and number estimation.

Further Explorations in Explaining Benford Bias

The test assessing the acquisition of the underlying distribution of variables in Animal tasks did not provide encouraging evidence to support optimal statistical inferences in cognitive judgment. Unlike the positive results with responses to everyday predictions, such as movie grosses or baking time, estimating the weight, lifespan, or group-size of wild animals involved less direct experience in daily life. This is similar to the anomaly reported with answers regarding the reigns of pharaohs by Griffith and Tenenbaum (2006) in their study. They noted that people might employ a different strategy to generate estimates in response to trivia or factual questions with greater uncertainty. Additionally, the quiz assessing participants' understanding of learned distributional statistics failed to demonstrate sufficient comprehension through simple exposure, as most participants required more than

two attempts on average to correctly answer the basic-level questions. Those proficient in remembering surface details may also pass the test without a deeper grasp. Therefore, future studies should revise the estimation items and include more life-experience questions for contrast. Moreover, to facilitate the effective acquisition of statistical information, one should consider refining the information quality or framing of descriptions with reader-friendly illustrations, such as using natural "frequencies" instead of complex terms like "probabilities" (Lesage et al., 2013).

The current project established a reliable observation between sensitivity to the distributional information of first digits and behaviour in number estimation. The improved accuracy of the first-digit estimation could explain shifts in Benford bias, especially when it started to approach the true pattern. Several studies (e.g., Lai et al., 2018) found that people feel it is important to talk about the first digit of a number, as seen in the left digit effect, where people disproportionately attend to the leftmost digit while making judgments or decisions. Thus, it should be common to observe some reactions like re-evaluating and adjusting the leading digit of quantities when the participants learn that their answers deviated largely from reality. Thus, the protocol analysis can serve as a valuable tool to further gain a deeper understanding of how people cognitively interact with feedback and alter their subsequent behaviour. It elicits verbal thoughts, capturing the elaborate details of cognitive processes, though it requires proper training (Austin & Delaney, 1998). Unlike simple comments or survey responses, protocol analysis seizes the complexity and nuance of thought processes in real time, allowing researchers to gain direct insight into the cognitive processes underlying specific tasks such as information processing and strategy selection (Ericsson & Simon, 1993). This approach may additionally clarify the different estimation behaviour reported due to the use of reference points. In the animal task, the Benford bias only arose from the estimates to the questions given starting point values, not average data,

with each set of estimates having a distinct underlying distribution. The think-aloud protocols enable step-by-step recordings of the mental process, such as multiplying a factor or making minor adjustments based on existing values, thus offering rich qualitative data that complements quantitative findings.

However, the frequent exposure to the first-digit pattern still could not cover the entire picture of the story. Despite consistent interaction with the true value patterns under feedback, there was no significant shift in the initial Benford bias when each round involved a random pairing of the domain with each true value, leading to the most challenging learning task in the last experiment. This could be potentially resolved by increasing exposure, such as extending the trials with two more blocks or slowing down the presentation rate, especially if the new pattern becomes more unfamiliar (Gebhart, Newport & Aslin, 2009). However, this also highlighted that the efficacy of learning can be limited by seemingly irrelevant yet accompanying features of statistical information. Similar effects have been documented in numerosity perception tasks, where the perceived quantities of visual elements covary with physical attributes like clustering or size (Crollen, Castronovo & Seron, 2010). Therefore, a thorough investigation into the conditions that influence how feedback affects estimation behaviour, particularly in relation to the display target, could provide more profound insights into the boundaries of generalisation in statistical learning.

In examining people's sensitivity to the first-digit distribution using Dot Displays, the improved accuracy in first-digit estimation was the key evidence in supporting a progressive shift in Benford bias. This was particularly notable under conditions where the true first digit did not conform to BL. However, feedback also effectively improved the first-digit accuracy in the group where the true values of the first digits agreed with BL. This appeared promising initially but was contradicted by our expectations. If better accuracy in first-digit estimation were to bring the pattern closer to the true distribution, one would anticipate a stronger

Benford bias in subsequent rounds with feedback in this case. However, this was not observed in the Benford-like group. In fact, the monotonic decline remained fairly stable without notable changes, irrespective of the feedback information. Therefore, the observation in Benford-like conditions raises questions about whether the enhanced performance in estimation accuracy is sufficient to fully account for the shifts in Benford bias. Thus, alternative explanations are demanded, such as changes in cognitive processes that have not yet been fully explored in this project.

The potential role of cognitive processes in number generation can also be linked to the mathematical fact that when the random samples are chosen from random distributions (in an unbiased way), the combined samples typically follow BL (Berger & Hill, 2015, 2020). This underlines the importance of integrating independent pieces of information to achieve a specific outcome. Psychology offers similar insights from the research studying strategies that enhance performance in estimation tasks, such as unpacking (Kruger & Evans, 2004), summing (Gasper, Roy & Flowe, 2019) or groupitising (Moscoso Castaldi, Burr, Arrighi & Anobile, 2020), as reviewed in Chapter 2. Identical to the solution to Fermi's problem, these approaches converge on the principle of consolidating judgments on smaller clusters of individual tasks to improve overall quality in estimation and decision-making. Therefore, it is plausible to consider that Benford bias may cognitively arise from combining multiple pieces of information during the estimation process. Further studies could utilise the protocol analysis to look for the relevant evidence, or directly design an experiment to manipulate the number of sub-components of judgmental steps within number estimation tasks, thereby testing this speculation.

Culture and Individual Differences

In examining the first-digit pattern of estimates, analyses regularly searched for any gender or language effect. Yet, no consistent result was recorded showing that any specific

gender or language group exhibited a distinctive behaviour, though occasionally, a difference might be signalled under certain conditions. However, the review of numerical cognition in Chapter 2 underlined that cultural, educational, and linguistic factors could shape how people cognitively perceive and process numerosity. Research suggested that symbolic comprehension and processing of numerals could benefit from more transparent linguistic representations, such as the place-value system used in Chinese (Miller et al., 2005). Recent studies have also revealed a relationship between the left digit effect and the structure of language. The prominence of the left digit effect is reduced among speakers of languages with inverted numerical expressions, such as Dutch (Savelkoul, Williams & Barth, 2020). These findings suggested that linguistic characteristics can impact the cognitive processing of multi-digit numbers.

Future research should explore the interaction between linguistic expression and number estimation, particularly to examine whether Benford bias can also emerge when individuals produce estimates using number words from their own languages. Neurological studies based on the Triple Code Model have shown that separate neural networks are responsible for processing two forms of symbolic numerical representation: the visual representation of Arabic numerals and the auditory verbal representation of numbers (Dehaene, 1992; Dehaene & Cohen, 1995). This further supports the need to investigate estimates generated with number words separately from those using Arabic numerals. Nevertheless, BL has shown remarkable robustness in natural settings, invariant to the changes of systems such as bases or units of measures. Therefore, the explorations with number words used in linguistic representation could additionally help to distinguish between psychological behaviours such as Benford bias and natural phenomenon.

Generating reasonable estimates to unknown questions and quantities of visual elements is also supported by executive functioning (EF), as demonstrated through the

studies of the impaired estimation performance in patients with damaged areas in the frontal lobe (Shallice & Evans, 1978) and other related mental health illness (e.g., Liss, Fein, Bullard, Robins, & Waterhouse, 2000; Jackson, Fein, Essock, & Mueser, 2001). Individual differences in executive functioning have been commonly recognised and reported across multiple cognitive tests; for instance, the ability to keep goal-relevant information actively maintained in the working memory varies from person to person (Tuholski, Engle & Baylis, 2001). Relevant studies collectively proposed that individual variations in EF are related (i.e., unity) yet separable (i.e., diversity), and this could be influenced by genetic factors (Friedman et al., 2008), as well as clinically and socially relevant phenomena (Miyake & Friedman, 2012). Thus, estimation processes associated with EF are likely to reflect individual differences. As a result, the Benford bias, when assumed to be a product of a cognitive process in number estimation, could covary with EF performance measured by cognitive tests.

Previously, the data from the Dots Display tasks have effectively shown a Benford bias aggregated from nine estimates from each individual, which offers a desirable and reliable paradigm for investigating individual differences. The advantage of this task with dots lies in its capacity to present participants with numerous trials using a massive array of stimuli, enabling a detailed analysis for estimating large quantities at the individual level.

Practice in Fraud Detection and Machine Learning

Earlier research (e.g., Nigrini, 1992) posited that Benford's distribution could serve as a tool for identifying fraudulent financial data, predicated on the belief that only naturally occurring datasets follow this pattern of monotonic decline. Yet, recent studies in Psychology, including the existing observations, presented a challenge that even generated estimates could align with Benford's pattern, casting doubt on the exclusive applicability of BL for fraud detection. However, it is critical to note that the simple act of creating numbers

in response to an unknown question in the current study design shall not be directly comparable to the intentional act of fabricating financial data, which often involves more complex considerations, such as psychological factors (Nigrini, 2005).

To test the assumption that fabricated numbers in financial contexts do not follow Benford's distribution, a revised design is required to directly provide a context where people are likely to deliberately distort financial figures, such as capitals on balance sheets. Such investigations could benefit from recruiting data to simulate real-life situations, for example, asking financial professionals to fill in missing numbers for a ledger account under normal business conditions or the specific purpose of what they believe would appear believable to others. Moreover, in modern auditing practices, the first two digits have been more frequently used in data verification than the simple model at the first digit place. Diekmann (2007) and Nigrini (1992; 2015) suggested that the second digit could reveal more about potential data manipulation. Hence, subsequent research should delve into the patterns of second digits created by individuals and how they relate to the first digits to truly gauge when BL can be most effectively used to detect deception.

In this project, several paradigms from the classic learning models in Psychology were applied to train and shape human behaviour and responses through different approaches like exposure to the statistical regularities in the animal estimation or reinforcing the performance via feedback when estimating the quantities of dots displayed. Such acquisition processes parallel the developmental paths of Artificial intelligence (AI), particularly in machine learning, which focuses on implicitly learning and generalising the statistical pattern for use in making decisions or predictions (Samuel, 1959). In fact, machine learning is deeply rooted in statistical science, supplying the analytical frameworks and methodologies necessary for data examination and interpretation (Bzdok, 2018). Traditional statistical analysis enables people to make inferences from sampling given pre-selected models. In

machine learning, the data itself informs and shapes the model as it learns from underlying structures and patterns. If human beings could demonstrate behaviour approximating BL of natural frequencies and adaptively update this distributional information through feedback in limited instances, it would also be intriguing to observe how powerfully the artificial neural networks react to similar stimuli. With sufficient simulations and inputs, data mining should be able to capture unknown properties of large datasets and distinguish human-generated information from genuine patterns. This unsupervised or semi-supervised learning approach promotes the use of AI as an anomaly detector, which is typically supported in screening outliers, noises, and unexpected activities (Hodge & Austin, 2004). Coupled with psychological evidence regarding people's sensitivity to statistical information and distinct behavioural patterns, such as the overproduction of digit-5 in number estimation, I believe that the integration of AI into the practice of data extraction and verification might further enhance the effectiveness, efficiency and reliability of the first digit phenomenon in fraud detection.

Conclusion

In exploring the underlying causes of Benford bias in number estimation, a series of experiments were conducted to examine the influence of statistical information on how people generate estimates in response to factual knowledge questions or when assessing large quantities of visual elements. The findings highlighted several key points: 1) There was a lack of conclusive evidence that individuals could acquire and utilise the distributions of variables when answering factual questions; 2) Indirect evidence supported that Benford bias might be a product of cognitive processes in number estimation, shaped by reference points when presented; and 3) Participants exhibited a pronounced sensitivity to the first-digit distribution when given vague feedback in visual estimation tasks. The improved accuracy in first-digit estimation could potentially explain the shift in Benford bias due to learning. However, these learning processes may face certain limitations, such as estimating quantities across a wide range of magnitudes and managing noise when paired with different targets.

Overall, the present project not only extended the phenomenon of Benford bias by applying it to diverse number estimation contexts but also disrupted this first-digit pattern through laboratory interventions. By incorporating Benford's distribution as an analytical tool, this study forged a distinctive connection between naturally occurring regularities and controlled, short-term experimental interventions. While the current findings supported the notion of statistical learning, as seen in the adaptation to first-digit distribution changes following feedback, consistent reports also captured that behaviour was still swayed by intuitive and heuristic judgments, such as spikes at digit-5. Thus, these findings highlighted a complex interplay between various influencing factors. The exploration into Benford bias and its cognitive underpinnings could enrich the field of statistical learning, broadening its relevance to number estimation tasks. In turn, this enhances comprehension of the methods individuals use in number estimation for decision-making.

References

Aida, S., Kusano, T., Shimono, K., & Tam, W. J. (2015). Overestimation of the number of elements in a three-dimensional stimulus. *Journal of Vision*, *15*(9), 23-23.

Alexopoulos, T., & Leontsinis, S. (2014). Benford's Law in astronomy. *Journal of Astrophysics and Astronomy*, *35*(4), 639–648.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological science*, *15*(2), 106-111.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological review*, *79*(2), 97.

Anobile, G., Castaldi, E., Moscoso, P. A. M., Burr, D. C., & Arrighi, R. (2020). "Groupitizing": a strategy for numerosity estimation. *Scientific Reports*, *10*(1), 13436.

Antell, S. E., & Keating, D. P. (1983). Perception of numerical invariance in neonates. *Child development*, 695–701.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *Quarterly Journal of Economics*, *118*, 73-105

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, *8*(1), 69–80.

Arshadi, L., & Jahangir, A. H. (2014). Benford's law behavior of Internet traffic. *Journal of Network and Computer Applications*, *40*, 194–205.

Aslin, R. N. (2017). Statistical learning: A powerful mechanism that operates by mere exposure. *WIREs Cognitive Science*, *8*, e1373.

Au, R. K., & Watanabe, K. (2013). Numerosity underestimation with item similarity in dynamic visual display. *Journal of Vision*, *13*(8), 5-5.

Bakker, A. (2004). Design research in statistics education: On symbolizing and computer tools (Doctoral dissertation).

Balamurugan, A., Harish, A., & Gandhi, S. (2019). Application of Benford's Law on stock turnover. *International Journal of Business and Management Invention (IJBMI), 8*(01), 63-67. 2319–801X.

Barabassy, A., Beinhoff, U. & Riepe, M. Cognitive estimation in mild Alzheimer's disease. *Journal of Neural Transm*, 114, 1479–1484 (2007).

Bargh, J. A., & Ferguson, M. J. (2000). Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin, 126*(6), 925–945.

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of memory and language*, *83*, 62-78.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society, 78*, 551-572.

Berger, A., & Hill, T. P. (2015). A Short Introduction to the Mathematical Theory of Benford's Law. In: S.J. Miller (ed.) *Benford's Law: Theory and Applications* (pp. 23-6)., Princeton University Press: Princeton, NJ.

Berger, A., & Hill, T. P. (2020). *The mathematics of Benford's law: a primer*. Statistical Methods & Applications.

Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2009). The Role of Simplification and Information in College Decisions: Results and Implications from the H&R Block FAFSA Experiment. An NCPR Working Paper. *National Center for Postsecondary Research*.

Blankenship, K. L., Wegener, D. T., Petty, R. E., Detweiler-Bedell, B., & Macy, C. L. (2008). Elaboration and consequences of anchored estimates: An attitudinal

perspective on numerical anchoring. *Journal of Experimental Social Psychology*, *44*(6), 1465-1476.

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental psychology*, *42*(1), 189.

Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. *Child development*, *79*(4), 1016-1031.

Brand, M., Fujiwara, E., Kalbe, E., Steingass, H. P., Kessler, J., & Markowitsch, H. J. (2003). Cognitive estimation and affective judgments in alcoholic Korsakoff patients. *Journal of Clinical and Experimental Neuropsychology*, *25*(3), 324-334.

Brooks, J. G., & Brooks, M. G. (1993). *In search of understanding: The case for constructivist classrooms.* Alexandria, VA: Association of Supervision and Curriculum Development.

Burns, B. D. (2009). Sensitivity to Statistical Regularities: People (largely) Follow Benford's Law. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 2872-2877). Austin, TX: Cognitive Science Society.

Burns, B. D., & Krygier, J. (2015). Psychology and Benford's Law. In S. J. Miller (Ed.), *The Theory and Applications of Benford's Law* (pp. 259-268). Princeton University Press.

Bullard, S. E., Fein, D., Gleeson, M. K., Tischer, N., Mapou, R. L., & Kaplan, E. (2005). Corrigendum to "The Biber Cognitive Estimation Test"[Archives of Clinical Neuropsychology 19 (2004) 835–846]. *Archives of Clinical Neuropsychology*, *20*(2), 279-279.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234.

Campbell, K. L., Zimerman, S., Healey, M. K., Lee, M. M., & Hasher, L. (2012). Age differences in visual statistical learning. *Psychology and Aging, 27*(3), 650–656.

Cavdaroglu, S., Katz, C., & Knops, A. (2015). Dissociating estimation from comparison and response eliminates parietal involvement in sequential numerosity perception. *NeuroImage*, *116*, 135-148.

Castronovo, J., Crollen, V., & Seron, X. (2010). Visual enumeration: A bi-directional mapping process between symbolic and non-symbolic representations of number?. *Journal of Vision*, *10*(7), 1327-1327.

Chambliss, D. F., & Schutt, R. K. (2006). *Making sense of the social world: Methods of investigation* (pp. 106-135). Chapter 5: Causation and experimental design.

Chang, S. A., & Chang, C. (Year). Identification of behavioral signature: Benford's law in online auctions. *International Journal of Business & Management Studies*, *4*(9).

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. *Heuristics and biases: The psychology of intuitive judgment*, (120-138).

Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica*, *148*, 163-172.

Chesney, D. L., & Haladjian, H. H. (2011). Evidence for a shared mechanism used in multiple-object tracking and subitizing. *Attention, Perception, & Psychophysics, 73*, 2457–2480.

Chesney, D. L., & Matthews, P. G. (2018). Task constraints affect mapping from Approximate Number System estimates to symbolic numbers. *Frontiers in Psychology*, *9*, 1801.

Cheyette, S. J., & Piantadosi, S. T. (2019). A primarily serial, foveal accumulator underlies approximate numerical estimation. *Proceedings of the National Academy of Sciences, 116*(36), 17729-17734.

Chesney, D. L., & Matthews, P. G. (2022). Circling around number: People can accurately extract numeric values from circle area ratios. *Psychonomic Bulletin & Review*, *29*(4), 1503-1513.

Ciccione, L., & Dehaene, S. (2020). Grouping mechanisms in numerosity perception. *Open Mind, 4*, 102-118.

Chi, D. (2020). *First Digit Phenomenon in Number Generation Under Uncertainty: Through the Lens of Benford's Law*. Unpublished master's thesis, The University of Sydney, New South Wales, Australia.

Chi, D., & Burns, B. D. (2022). Why do people fit to Benford's Law? – A test of the recognition hypothesis. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society* (pp. 3648-3654).

Choo, H., & Franconeri, S. L. (2014). Enumeration of small collections violates Weber's law. *Psychonomic bulletin & review*, *21*, 93-99.

Cipolotti, L., Warrington, E. K., & Butterworth, B. (1995). Selective impairment in manipulating Arabic numerals. *Cortex, 31*(1), 73-86.

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and instruction*, *21*(1), 1-78.

Compton, B. J., & Logan, G. D. (1999). Judgments of perceptual groups: Reliability and sensitivity to stimulus transformation. *Perception & Psychophysics*, *61*(7), 1320-1335.

Converse, B. A., & Dennis, P. J. (2018). 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*, *147*, 94-107.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 24.

Corbetta, M., Shulman, G. L., Miezin, F. M., & Petersen, S. E. (1995). Superior parietal cortex activation during spatial attention shifts and visual feature conjunction. *Science*, *270*(5237), 802-805.

Coubart, A., Izard, V., Spelke, E. S., Marie, J., & Streri, A. (2014). Dissociation between small and large numerosities in newborn infants. *Developmental Science*, *17*(1), 11-22.

Crollen, V., Castronovo, J., & Seron, X. (2011). Under-and over-estimation. *Experimental psychology*.

D'Aniello, G. E., Castelnuovo, G., & Scarpina, F. (2015). Could cognitive estimation ability be a measure of cognitive reserve? *Frontiers in Psychology*, *6*, 608.

Decarli, G., Paris, E., Tencati, C., Nardelli, C., Vescovi, M., Surian, L., & Piazza, M. (2020). Impaired large numerosity estimation and intact subitizing in developmental dyscalculia. *PLoS One*, *15*(12), e0244578.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44(1-2)*, 1-42.

Dehaene, S. (1997). The number sense: How the mind creates mathematics. Oxford, UK: Oxford University Press.

Dehaene, S. (2001). Précis of the number sense. *Mind & Language*, *16*(1), 16-36.

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. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. *From monkey brain to human brain*, 133-157.

Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical cognition, 1*(1), 83-120.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition, 43*(1), 1-29.

Dehaene, S., Sergent, C., & Changeux, J. P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences*, *100*(14), 8520-8525.

de Finetti, B. (1974). The true subjective probability problem. In *The concept of probability in psychological experiments* (pp. 15-23). Dordrecht: Springer Netherlands

Della Sala, S., MacPherson, S. E., Phillips, L. H., Sacco, L., & Spinnler, H. (2004). The role of semantic knowledge on the cognitive estimation task: evidence from Alzheimer's disease and healthy adult aging. *Journal of Neurology*, *251*, 156-164.

Diekmann, A. (2007). Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. *Journal of Applied Statistics, 34*(3), 321-329.

Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting*, *5*(1), 17-34.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*(2), 188-200.

Enough, B., & Mussweiler, T. (2001). Sentencing under uncertainty: Anchoring effects in the courtroom 1. *Journal of Applied Social Psychology*, *31*(7), 1535-1551.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological science*, *12*(5), 391-396.

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making*, *18*(3), 199-212.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT Press.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83*(1), 37–64.

Eysenck, M. W., & Keane, M. T. (2015). *Cognitive psychology: A student's handbook* (7th ed.). Psychology Press.

Fayol, M., Seron, X., & Campbell, J. I. D. (2005). Handbook of mathematical cognition.

Farhadi, N., & Lahooti, H. (2021). Are COVID-19 data reliable? A quantitative analysis of pandemic data from 182 countries. *COVID*, *1*, 137-152.

Farhadi, N., & Lahooti, H. (2022). Forensic analysis of COVID-19 data from 198 countries two years after the pandemic outbreak. *COVID*, *2*(4), 472-484.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*(7), 307-314.

Fiser, J. Z., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 458-467.

Fornaciai, M., Brannon, E. M., Woldorff, M. G., & Park, J. (2017). Numerosity processing in early visual cortex. *NeuroImage*, *157*, 429-438.

Forsyth, D. K., & Burt, C. D. (2008). Allocating time to future tasks: The effect of task segmentation on planning fallacy bias. *Memory & Cognition, 36,* 791-798.

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General, 137*(2), 201–225.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-economics, 40*(1), 35-42.

Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of personality and social psychology*, *81*(4), 657.

Galton, F. (1907). Vox populi. *Nature, 75(1949),* 450-451.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for research in Mathematics Education, 19*(1), 44-63.

Gasper, H. L., Roy, M. M., & Flowe, H. D. (2019). Improving time estimation in witness memory. *Frontiers in Psychology, 10,* Article 1452.

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing structures in midstream: Learning along the statistical garden path. *Cognitive Science, 33*(7), 1087–1116.

Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review, 16*(3), 486–490.

Geary, D. C. (2013). Early foundations for mathematics learning and their relations to learning disabilities. *Current directions in psychological science, 22*(1), 23-27.

Geary, D. C., & Moore, A. M. (2016). Cognitive and brain systems underlying early mathematical development. *Progress in brain research, 227,* 75-103.

Gebuis, T., Kadosh, R. C., & Gevers, W. (2016). Sensory-integration system rather than approximate number system underlies numerosity processing: A critical review. *Acta psychologica*, *171*, 17-35.

Geyer, C. L., & Williamson, P. P. (2004). Detecting fraud in data sets using Benford's Law. Communications in Statistics: Simulation & Computation, *33*, 229-246.

Gibbon, J., & Church, R. M. (1990). Representation of time. *Cognition*, *37*(1-2), 23-54.

Gigerenzer, G., & Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999).

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & The ABC Research Group, *Simple heuristics that make us smart* (pp. 3–34). Oxford University Press.

Ginsburg, N., & Goldstein, S. R. (1987). Measurement of visual cluster. *The American Journal of Psychology*, 193-203.

Ginsburg, N. (1976). Effect of item arrangement on perceived numerosity: Randomness vs regularity. *Perceptual and motor skills*, *43*(2), 663-668.

Green, C. S., Benson, C., Kersten, D., & Schrater, P. (2010). Alterations in choice behavior by manipulations of world model. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 16401–16406.

Griesser, M., Ma, Q., Webber, S., Bowgen, K., & Sumpter, D. J. (2011). Understanding animal group-size distributions. *PloS One*, *6*(8).

Griffith, T., & Tenenbaum, J. (2006). Optimal Predictions in Everyday Cognition. *Psychological Science*, *17*(9), 767-73.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). Thousand Oaks, CA: Sage.

Hales, B., Terblanche, M., Fowler, R., & Sibbald, W. (2008). Development of medical checklists for improved quality of patient care. *International Journal for Quality in Health Care*, *20*(1), 22-30.

Hamdan, N., & Gunderson, E. A. (2017). The number line is a critical spatial-numerical representation: Evidence from a fraction intervention. *Developmental Psychology*, *53*(3), 587.

Harvey, B. M. (2016). Quantity cognition: Numbers, numerosity, zero and mathematics. *Current Biology*, *26*(10).

Haselton, M. G., Nettle, D., & Andrews, P. W. (2005). The evolution of cognitive bias, Buss, DM. *The handbook of evolutionary psychology*, 724-746.

Hassan, B. (2002). Assessing data authenticity with Benford's law. *Information Systems Control Journal*, *6*, 41-46.

Hauser, J. R., & Wernerfelt, B. (1990). An evaluation cost model of consideration sets. *Journal of consumer research*, *16*(4), 393-408.

He, L., Zhang, J., Zhou, T., & Chen, L. (2009). Connectedness affects dot numerosity judgment: Implications for configural processing. *Psychonomic bulletin & review*, *16*(3), 509-517.

Hill, T. P. (1995a). Base-invariance implies Benford's law. *Proceedings of the American Mathematical Society*, *123*, 887-895.

Hill, T. P. (1995b). A statistical derivation of the significant-digit law. *Statistical Science*, 354-363.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology, 103*(1), 17-29.

Honda, H., Kagawa, R., & Shirasuna, M. (2022). On the round number bias and wisdom of crowds in different response formats for numerical estimation. *Scientific Reports, 12*, 8167.

Horiuchi, S. (2003). Interspecies differences in the life span distribution: Humans versus invertebrates. *Population and Development Review, 29*, 127–151.

Hyde, D. C., & Spelke, E. S. (2009). All numbers are not equal: an electrophysiological investigation of small and large number representations. *Journal of Cognitive Neuroscience*, *21*(6), 1039-1053.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.

Im, H. Y., Zhong, S. H., & Halberda, J. (2016). Grouping by proximity and the visual impression of approximate number in random dot arrays. *Vision Research*, *126*, 291-307.

Ismail, Z., & Chan, S. W. (2015). Malaysian students' misconceptions about measures of central tendency: An error analysis. *AIP Conference Proceedings, 1643*, 93-100.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing?. *Journal of Personality and Social Psychology*, *79*(6), 995.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221-1247.

Jackson, C. T., Fein, D., Essock, S. M., & Mueser, K. T. (2001). The effects of cognitive impairment and substance abuse on psychiatric hospitalizations. *Community Mental Health Journal*, *37*, 303-312.

Jevons, W. S. (1871). The power of numerical discrimination. *Nature*, *3*(67), 281-282.

Jha, T., Mendel, J., Cho, H., & Choudhary, M. (2022). Prediction of bacterial sRNAs using sequence-derived features and machine learning. *Bioinformatics and Biology Insights*, *16*, 1–15.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449-1475.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: a flaw in human judgment*. Hachette UK.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*(3), 430-454.

Kahneman, D., & Tversky. A. (1985). "Evidential impact of base rates". In Daniel Kahneman, Paul Slovic & Amos Tversky (ed.). *Judgment under uncertainty: Heuristics and biases*. Science. 185, 153–160.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*(4), 498-525.

Kayton, K., Williams, K., Stenbaek, C., Gwiazda, G., Bondhus, C., Green, J & Patalano, A. L. (2022). Summary accuracy feedback and the left digit effect in number line estimation. *Memory & Cognition*, *50*(8), 1789-1803.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, *7*(5), e36399.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35-B42.

Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. (2013). Early numerical development and the role of non-symbolic and symbolic skills. *Learning and Instruction, 25*, 95-103.

Konold, C., & Higgins, T. (2003). Reasoning about data. *A research companion to principles and standards for school mathematics*, 193215.

Kouw, W. M., & Loog, M. (2019). A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, *43*(3), 766-785.

Kreuzer, M., Jordan, D., Antkowiak, B., Drexler, B., Kochs, E. F., & Schneider, G. (2014). Brain electrical activity obeys Benford's law. *Anesthesia & Analgesia*, *118*(1), 183-91.

Kruger, J., & Evans, M. (2004). If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology*, *40*(5), 586-598.

Kubovy, M. (1977). Response availability and the apparent spontaneity of numerical choices. *Journal of Experimental Psychology: Human Performance and Performance, 2,* 359-364.

Kubovy, M., & Psotka, J. (1976). The predominance of seven and the apparent spontaneity of numerical choices. *Journal of Experimental Psychology: Human Perception and Performance, 2*(2), 291-294.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, 21(8), 1134-1140.

Lai, M., Zax, A., & Barth, H. (2018). Digit identity influences numerical estimation in children and adults. *Developmental Science, 21*(5), e12657.

Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. *Cognitive Science, 37*(5), 775–799.

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive science*, *33*(6), 969-998.

Lesage, E., Navarrete, G., & De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: The role of general cognitive resources. *Thinking & Reasoning*, *19*(1), 27-53.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of experimental psychology: Human Learning and Memory*, *4*(6), 551.

Liss, M., Fein, D., Bullard, S., & Robins, D. (2000). Brief report: Cognitive estimation in individuals with pervasive developmental disorders. *Journal of autism and developmental disorders*, *30*(6), 613.

Liu, Y., Li, S., & Sun, Y. (2014). Unpacking a time interval lengthens its perceived temporal distance. *Frontiers in Psychology*, *5*, 1345.

Macan, T. H. (1994). Time management: Test of a process model. *Journal of Applied Psychology*, *79*(3), 381.

Miller, K. F., Kelly, M., & Zhou, X. (2005). Learning mathematics in China and the United States. *Handbook of mathematical cognition*, *8*(35), 163-178.

Miller. S. (Eds.). (2015). *The Theory and Applications of Benford's Law*. Princeton, NJ, Princeton University Press.

Minturn, A. L., & Reese, T. W. (1951). The effect of differential reinforcement on the discrimination of visual number. *The Journal of Psychology: Interdisciplinary and Applied*, *31*, 201–231.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21*(1), 8-14.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*(1), 20–39.

Moore, A. M., & Ashcraft, M. H. (2015). Children's mathematical performance: Five cognitive tasks across five grades. *Journal of Experimental Child Psychology, 135*, 1-24.

Moscoso, P. A., Castaldi, E., Burr, D. C., Arrighi, R., & Anobile, G. (2020). Grouping strategies in number estimation extend the subitizing range. *Scientific Reports*, *10*(1), 14979.

Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds?. *Cognitive Science, 32(7),* 1133-1147.

Mundy, E., & Gilmore, C. K. (2009). Children's mapping between symbolic and nonsymbolic representations of number. *Journal of Experimental Child Psychology*, *103*(4), 490-502.

Muradian, K. H. (1989). Distribution of the species-specific lifespan in various taxonomic animal groups. *Zhurnal Obshchei Biologii, 50*(1), 22–26.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology, 35*(2), 136-164.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, *26*(9), 1142-1150.

Negi, A., & Singh, K. (2023). Detection of red flags in coronavirus vaccination data available in public domain. *Juni Khyat* (UGC Care Group I Listed Journal), 13(08).

Neisser, U. (1967). Cognitive Psychology, Appleton-Century-Crofts, New York, 1967. *The functions and nature of imagery (ed. PW Sheehan). Academic Press*, New York. Novick, R., and Lazar, 955-61.

Nelson, S. P., Wickman, B., Null, J., & Gazin, E. (Year). An evaluation of music's adherence to Benford's Law throughout history. *Journal of Mathematical Sciences: Advances and Applications*, *67*(1), 73-84.

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and brain sciences*, *37*(1), 1-19.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, *4*(1), 39-40.

Nieder, A., & Miller, E. K. (2004). A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences*, *101*(19), 7457-7462.

Nigrini, M. J. (1992). *The detection of income tax evasion through an analysis of digital distributions*. (Unpublished Dissertation), University of Cincinnati, United States of America.

Nigrini, M. J. (2015). Detecting fraud and errors using Benford's law. In S. J. Miller (Ed.), *The Theory and Applications of Benford's Law* (pp. 180-200). Princeton, NJ: Princeton University Press.

Oakes, L. M., Hurley, K. B., Ross-Sheehy, S., & Luck, S. J. (2011). Developmental changes in infants' visual short-term memory for location. *Cognition*, *118*(3), 293–305.

Odic, D., & Starr, A. (2018). An introduction to the approximate number system. *Child Development Perspectives*, *12*(4), 223-229.

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, *72*(1), 147-152.

Oppenheimer, D. M., LeBoeuf, R. A., & Brewer, N. T. (2008). Anchors aweigh: a demonstration of cross-modality anchoring and magnitude priming. *Cognition*, *106*(1), 13–26.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745-2750.

Park, J. H., Choi, C. H., & Cho, E. (2016). Preliminary study to detect match-fixing: Benford's law in badminton rally data. *Journal of Physical Education*, *3*(1), 64-77.

Parnes, D. (2022). Banks' off-balance sheet manipulations. *The Quarterly Review of Economics and Finance*, *86*, 314-331.

Peiró-Signes, Á., Trull, O., Segarra-Oña, M., & García-Díaz, J. C. (2021). Anxiety towards statistics and its relationship with students' attitudes and learning approach. *Behavioral Sciences*, *11*(3), 32.

Piazza, M. (2011). Neurocognitive start-up tools for symbolic number representations. In S. Dehaene & E. M. Brannon (Eds.), *Space, time and number in the brain* (pp. 267-285). Academic Press. (Reprinted from *Trends in Cognitive Sciences, 14*, 542–551, 2010, with permission from Elsevier.)

Piazza, M., Fumarola, A., Chinello, A., & Melcher, D. (2011). Subitizing reflects visuo-spatial object individuation capacity. *Cognition*, *121*(1), 147-153.

Plate, R. C., Fulvio, J. M., Shutts, K., Green, C. S., & Pollak, S. D. (2018). Probability learning: Changes in behavior across time and development. *Child development*, *89*(1), 205-218.

Pomerantz, M. J. (1981). Do “Higher Order Interactions” in Competition Systems Really Exist?. *The American Naturalist*, *117*(4), 583-591.

Poom, L., Lindskog, M., Winman, A., & Van den Berg, R. (2019). Grouping effects in numerosity perception under prolonged viewing conditions. *PloS One*, *14*(2), e0207502.

Pope, D., & Simonsohn, U. (2011). Round numbers as goals: Evidence from baseball, SAT takers, and the lab. *Psychological Science*, *22*(1), 71-79.

Price, G. R., & Ansari, D. (2011). Symbol processing in the left angular gyrus: evidence from passive perception of digits. *Neuroimage*, *57*(3), 1205-1211.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*(1), 30-54.

Redko, I., Morvant, E., Habrard, A., Sebban, M., & Bennani, Y. (2019). *Advances in domain adaptation theory*. Elsevier.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, *19*(6), 607-614.

Roggeman, C., Santens, S., Fias, W., & Verguts, T. (2011). Stages of nonsymbolic number processing in occipitoparietal cortex disentangled by fMRI adaptation. *Journal of Neuroscience*, *31*(19), 7168-7173.

Roy, M. M., Mitten, S. T., & Christenfeld, N. J. (2008). Correcting memory improves accuracy of predicted task duration. J*ournal of Experimental Psychology: Applied*, *14(3)*, 266.

Russo, D., Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2017). A Tutorial on Thompson Sampling. *Foundations and Trends in Machine Learning*, *11*(1), 1-96.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science, 274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27-52.

Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229.

Sanchez, D. J., & Reber, P. J. (2013). Explicit pre-training instruction does not improve implicit perceptual-motor sequence learning. *Cognition*, *126*(3), 341-351.

Savelkouls, S., Williams, K., & Barth, H. (2020). Linguistic inversion and numerical estimation. *Journal of Numerical Cognition*, *6*(3), 263-274.

Schmithorst, V. J., & Brown, R. D. (2004). Empirical validation of the triple-code model of numerical processing for complex math operations using functional MRI and group Independent Component Analysis of the mental addition and subtraction of fractions. *NeuroImage*, *22*(3), 1414-1420.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*(2), 195-202.

Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1264.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 592.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, *24*(4), 449-474.

Shallice, T., & Evans, M. E. (1978). The involvement of the frontal lobes in cognitive estimation. *Cortex, 14*(2), 294-303.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. *Psychological Bulletin, 134*(2), 207.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child development, 75*(2), 428-444.

Siegler, R. S., & Opfer, J. E. (2003). The Development of Numerical Estimation: Evidence for Multiple Representations of Numerical Quantity. *Psychological Science, 14*(3), 237-250.

Silverman, S., & Ashkenazi, S. (2016). Deconstructing the cognitive estimation task: a developmental examination and intra-task contrast. *Scientific Reports, 6*(1), 39316.

Smets, K., Sasanguie, D., Szücs, D., & Reynvoet, B. (2015). The effect of different methods to construct non-symbolic stimuli in numerosity estimation and comparison. *Journal of Cognitive Psychology, 27*(3), 310-325.

Stael von Holstein, C. A. (1972). Probabilistic forecasting: An experiment related to the stock market. *Organisational Behavior & Human Performance, 8*(1), 139–158.

Starr, A., DeWind, N. K., & Brannon, E. M. (2017). The contributions of numerical acuity and non-numerical stimulus features to the development of the number sense and symbolic math achievement. *Cognition, 168*, 222-233.

Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences, 110*(45), 18116-18120.

Strauss, E., Sherman, E., & Spreen, O. (2006). *A compendium of neuropsychological tests*. Oxford University Press.

Striga, D., & Podobnik, V. (2018). Benford's Law and Dunbar's Number: Does Facebook Have a Power to Change Natural and Anthropological Laws? *IEEE Access, 6,* 14629-14642.

Stuss, D. T., & Levine, B. (2002). Adult clinical neuropsychology: lessons from studies of the frontal lobes. *Annual Review of Psychology*, *53*(1), 401-433.

Tao, M., Yang, D., & Liu, W. (2019). Learning effect and its prediction for cognitive tests used in studies on indoor environmental quality. *Energy and Buildings, 197*, 87-98.

Thomas, M., Morwitz, V., & [Dawn Iacobucci served as editor and Kent Monroe served as associate editor for this article.]. (2005). Penny Wise and Pound Foolish: The Left-Digit Effect in Price Cognition. *Journal of Consumer Research*, *32*(1), 54–64.

Thompson, C. A., & Opfer, J. E. (2010). How 15 Hundred Is Like 15 Cherries: Effect of Progressive Alignment on Representational Changes in Numerical Cognition. *Child Development*, *81*(6), 1768–1786.

Thompson, C. A., & Siegler, R. S. (2010). Linear numerical-magnitude representations aid children's memory for numbers. *Psychological Science*, *21*(9), 1274-1281.

Thorndike, E. L. (1931). *Human Learning.* The Century Co.

Timmermans, D. (1993). The impact of task complexity on information use in multi-attribute decision making. *Journal of Behavioural Decision Making*, *6*(2), 95-111.

Tobin, S., & Grondin, S. (2015). Prior task experience affects temporal prediction and estimation. *Frontiers in Psychology*, *6*, 916.

Tripodi, M. (2016). *Why do people produce Benford's law?* (Unpublished Honours Thesis), University of Sydney, Australia.

Tuholski, S. W., Engle, R. W., & Baylis, G. C. (2001). Individual differences in working memory capacity and enumeration. *Memory & Cognition*, *29*, 484-492.

Tulving, E. (1976). Ecphoric processes in recall and recognition. In J. Brown (Ed.), *Recall and recognition.* John Wiley & Sons.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*(4), 552.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105–110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Uchmański, J. (1985). Differentiation and frequency distributions of body weights in plants and animals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *310*(1142), 1–75.

van Hoogmoed, A. H., Huijsmans, M. D., & Kroesbergen, E. H. (2021, June). Non-symbolic numerosity and symbolic numbers are not processed intuitively in children: Evidence from an event-related potential study. In *Frontiers in Education* (Vol. 6, p. 629053). Frontiers Media SA.

van Hoogmoed, A. H., & Kroesbergen, E. H. (2018). On the difference between numerosity processing and number processing. *Frontiers in Psychology*, *9*, 1650.

Van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: Counting by groups. *Memory & Cognition*, *10*, 396-404.

van Opstal, F., Gevers, W., De Moor, W., & Verguts, T. (2008). Dissecting the symbolic distance effect: Comparison and priming effects in numerical and nonnumerical orders. *Psychonomic Bulletin & Review*, *15*, 419-425.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of economic surveys*, *14*(1), 101-118.

Waismeyer, A., Meltzoff, A. N., & Gopnik, A. (2015). Causal learning from probabilistic events in 24-month-olds: an action measure. *Developmental science*, *18*(1), 175-182.

Wegener, D. T., Petty, R. E., Detweiler-Bedell, B. T., & Jarvis, W. B. G. (2001). Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology*, *37*(1), 62-69.

Weinstein, L., & Adam, J. A. (2008). *Guesstimation: Solving the world's problems on the back of a cocktail napkin*. Princeton University Press.

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2009). Speech segmentation in a simulated bilingual environment: A challenge for statistical learning? *Language Learning and Development*, *5*(1), 30–49.

Weiss, Y., Simoncelli, E.P., Adelson E.H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5, 598–604.

Wells, H. G. (1903). *Mankind in the Making*. Chapman & Hall, Ld. London.

Wender, K. F., & Rothkegel, R. (2000). Subitizing and its subprocesses. *Psychological Research*, *64*, 81-92.

Wertheimer, M. (1924). Gestalt Theory. Address Before the Kant Society, Berlin, 7th December, 1924; reprinted in Wertheimer M (1950) in A Sourcebook of Gestalt Psychology, W.D. Ellis (ed.) (pp. 1–11), Humanities, New York.

Whynes, D. K., Philips, Z., & Frew, E. (2005). Think of a number… any number?. *Health Economics*, *14*(11), 1191-1195.

Williams, K., Zax, A., Patalano, A. L., & Barth, H. (2022). Left digit effects in numerical estimation across development. *Journal of Cognition and Development*, *23*(2), 188-209.

Willingham, D. B., & Goedert-Eschmann, K. (1999). The relation between implicit and explicit learning: Evidence for parallel development. *Psychological Science*, *10*(6), 531-534.

Wong, T. T. Y., Ho, C. S. H., & Tang, J. (2017). Defective number sense or impaired access? Differential impairments in different subgroups of children with mathematics difficulties. *Journal of Learning Disabilities*, *50*(1), 49-61.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(13), 5012–5015.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1-B11.

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive science*, *36*(3), 452-470.

Yeo, D. J., Wilkey, E. D., & Price, G. R. (2017). The search for the number form area: a functional neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, *78*, 145-160.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1-B11.

Zawojewski, J. S., & Shaughnessy, J. M. (2000). Take time for action: Mean and median: Are they really so easy? *Mathematics Teaching in the Middle School*, *5(7)*, 436-440.

Appendices

Appendix A

Instructions for the Animal Estimation Task in Experiment 1

Animal Estimation

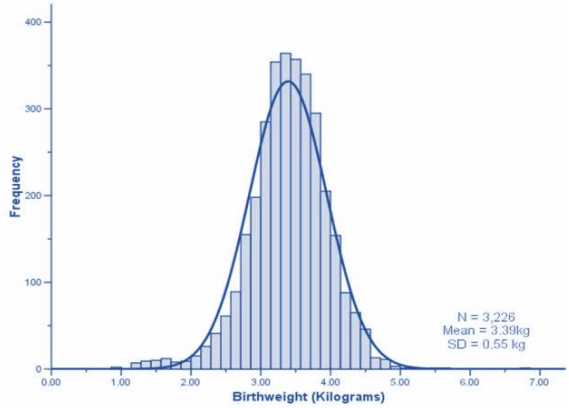
In this task, you will be asked to estimate the weight, lifespan, and group size of an animal, based on the picture provided. There are nine animal pictures in total.

Evidence (e.g., Griffiths & Tenenbaum, 2006) has suggested that when people make estimations, they might utilise and follow the underlying distributions. The distribution tends to show all the possible values (or intervals) of the data and how often each value occurs. In general, the weight is normally distributed, the lifespan is negatively skewed, and the group size typically follows a logarithm distribution.

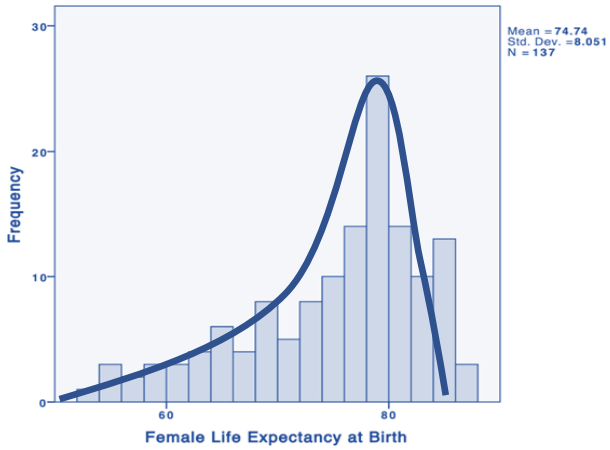
Hence, to assist your estimation, we would like to show you what the distributions regarding weight, lifespan and group size look like, with examples from the data of human beings. You will learn three different patterns with more detailed explanations in the following pages.

Appendix B

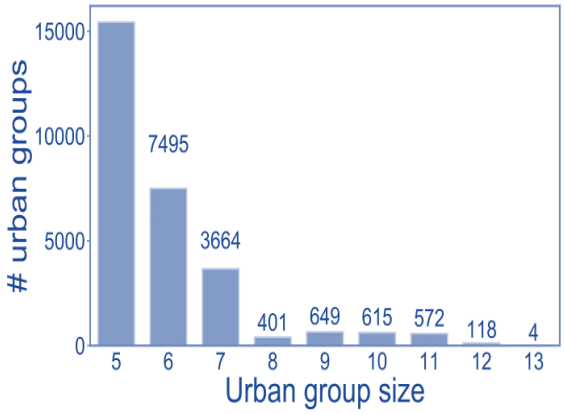
Materials of the Distributional Information of Three Variables Supplied in the Exposure Stage to Learn for the Animal Estimation



This is a distribution of birth weight in 3,226 newborn babies (data from O'Cathain et al 2002).
The x-axis displays the birthweight in kilograms, and the y-axis shows the number of newborn babies weighted in each value.
This is an example of a normal distribution, which is symmetrical around its centre (i.e. around the mean, μ). That is, the right side of the centre is a mirror image of the left side. The mean, mode and median are all equal.



This is a distribution of female life expectancy at birth (data from World Health Organization, 2012).
The x-axis displays different ages, and the y-axis shows the number of females for each value (i.e., expected lifespan).
This is an example of a skewed (non-symmetric) distribution, which one tail of the distribution considerably longer or drawn out relative to the other tail. This histogram of female expected lifespan is a left-skewed distribution, where most values are clustered around the right tail of the distribution while the left tail of the distribution is longer.



This is a distribution of the urban groups size measured on WhatsApp in mobile phone networks (data from Zignani, Quadri, Gaito, et al., 2019).
The x-axis displays the number of members in the group (i.e., group size), and the y-axis shows the number of groups in each group size
This is an example of a logarithm distribution, which is used for modelling the behaviour of items with a constant failure rate. As the right-skewed curve in the picture, it highlights that small groups (k=5,6) are predominant in mobile phone networks.

Appendix C

Questions Items in the Animal Estimation Task in Experiment 1

It is a picture of a Capybara.
Please give an estimation to the questions below:

The weight of this Capybara is _____ (g)
The lifespan of this Capybara is _____ (days)
How many Capybaras are usually grouped in the wild?
_____

It is a picture of an Atlantic Herring.
Please give an estimate to the questions below:

The weight of this Atlantic Herring is _____ (g)
The lifespan of this Atlantic Herring is
_____(days)
How many Atlantic Herrings are usually grouped in the wild? _____

It is a picture of a Cape Buffalo.
Please give an estimate to the questions below:

The weight of this Buffalo is _____ (g)
The lifespan of this Buffalo is _____(days)
How many Buffalos are usually grouped in the wild?
_____

It is a picture of a Red-winged Blackbird.
Please give an estimate to the questions below:

The weight of this Blackbird is _____(g)
The lifespan of this Blackbird is _____(days)
How many Blackbirds are usually grouped in the wild?
_____

It is a picture of a Purple Sea Urchin.
Please give an estimate to the questions below:

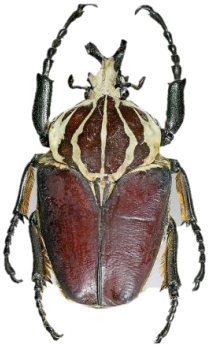
The weight of this Purple Sea Urchin is _____ (g).
The lifespan of this Purple Sea Urchin is _____(days)
How many Purple Sea Urchins are usually grouped in the wild? _____

It is a picture of a Plains Garter Snake.
Please give an estimate to the questions below:

The weight of this Plains Garter Snake is _____ (g).
The lifespan of this Plains Garter Snake is _____(days)
How many Plains Garter Snakes are usually grouped in the wild? _____

It is a picture of a Fiddler Crab.
Please give an estimate to the questions below:

The weight of this Fiddler Crab is _____ (g).
The lifespan of this Fiddler Crab is _____(days)
How many Fiddler Crabs are usually grouped in the wild? _____



It is a picture of a Goliath beetle.
Please give an estimate to the questions below:

The weight of this Goliath beetle is _____(g).
The lifespan of this Goliath beetle is _____(days)
How many Goliath beetles are usually grouped in the wild? _____

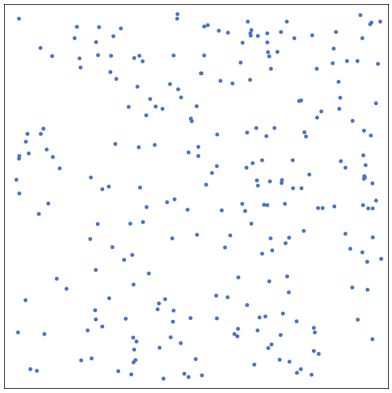
It is a picture of a Greater flamingo.
Please give an estimate to the questions below:

The weight of this Greater flamingo is _____ (g).
The lifespan of this Greater flamingo is _____(days)
How many Greater flamingos are usually gathered in the wild? _____

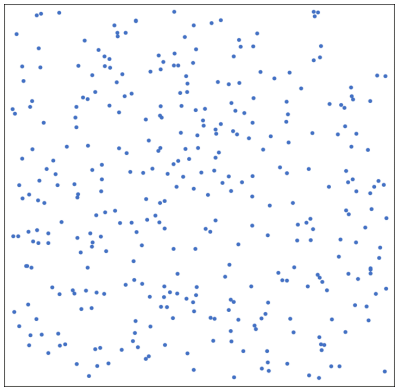
Appendix D

Items of Dots Displays Present in Experiments 1 and 2 in the Uniform Group

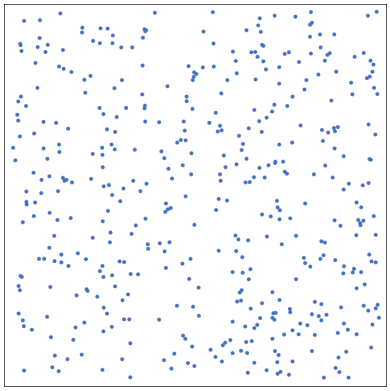
250 Dots:



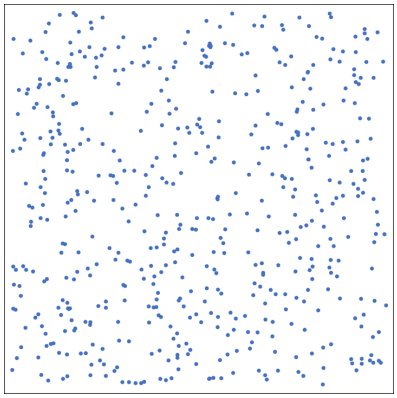
350 Dots:



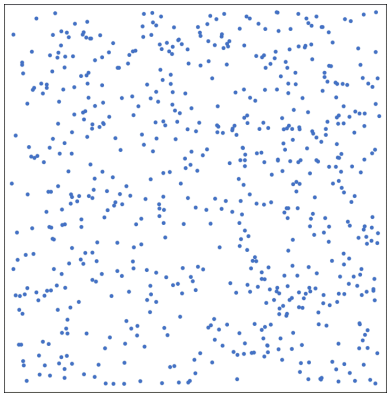
450 Dots:



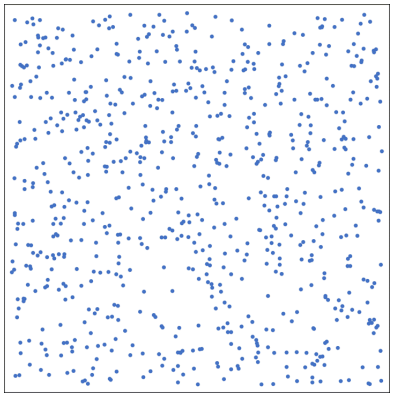
550 Dots:



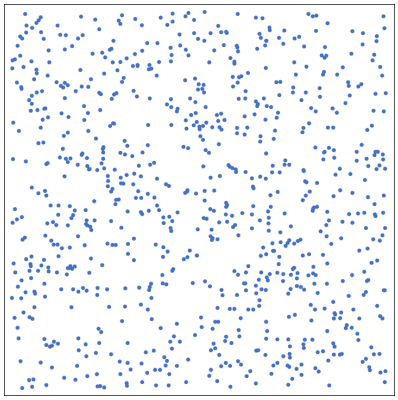
650 Dots:



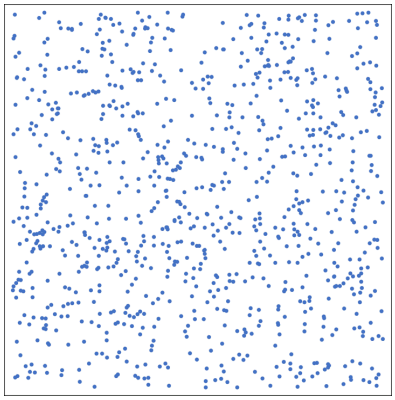
750 Dots:



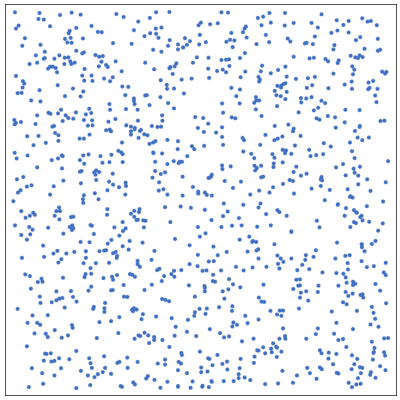
850 Dots:



950 Dots:



1050 Dots:



Appendix E

Demographic Questionnaires

Please answer the following questions about yourself:

What is your gender? ○ Male ○ Female ○ Non-binary/other

What is your age in years? ☐

What was the first language that you learned to speak? Please select one language from the menu below:
Afrikaans
Albanian
Arabic
Armenian
Basque
Bengali
Bulgarian
Catalan
Cambodian
Chinese
Croatian
Czech
Danish
Dutch
English
Estonian
Fiji
Finnish
French
Georgian
German
Greek
Gujarati
Hebrew
Hindi
Hungarian
Icelandic
Indonesian
Irish
Italian
Japanese
Javanese
Korean
Latin
Latvian
Lithuanian

Macedonian
Malay
Malayalam
Maltese
Maori
Marathi
Mongolian
Nepali
Norwegian
Persian
Polish
Portuguese
Punjabi
Quechua
Romanian
Russian
Samoan
Serbian
Slovak
Slovenian
Spanish
Swahili
Swedish
Tamil
Tatar
Telugu
Thai
Tibetan
Tonga
Turkish
Ukrainian
Urdu
Uzbek
Vietnamese
Welsh
Xhosa
OTHER
If some other language, then please specify: [ ]

Which of these fields best describes your major, current major or field of expertise? Please select one option from the menu below:
Arts and Humanities (Linguistics & Literatures, Philosophy, Music, etc.)
Business (Accounting, Finance, etc.)
Health and Medicine (Nursing, Pharmacy, etc.)
Public and Social Services (Law, Security, etc.)
Science and Math (Biology, Statistics, etc.)
Technology & Engineering (Communication technologies, etc.)
Social sciences (Political science, Education, etc.)
OTHER
If some other field, then please specify: [ ]

Appendix F

General Instructions

Welcome to the experiment!

On some pages it will be necessary to scroll down to see all the questions. You won't be able to proceed to the next page until you have answered every question on a page. (The button allowing you to proceed to the next page is always at the bottom of the page.) You will not be able to return to a previous page once you have proceeded to the next page.

Throughout this experiment, you will be met with difficult to answer questions, they are designed to be hard to answer. Even if you have no idea, please just guess.

If you are using the browser "Internet Explorer" it is best if you change to a different browser.

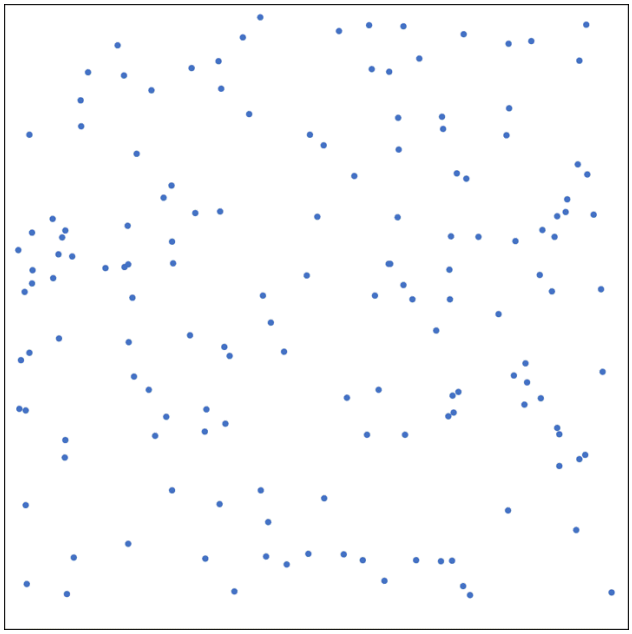
When ready to continue, just click on the button below.

Appendix G

Instructions with a Demonstration for the Dots Display Tak

In this task, you will be asked to estimate the number of dots displayed for a few seconds. There will be two rounds of dots, which makes 27 (3*9) items in total. Please answer every question to the best of your ability by entering the *Arabic numerals.*

This is a sample of the dots (after clicking the button)



Click to show the dots

***In the test, please click the action button first, and the dots will be presented for a few seconds and then disappear in the display.*

Appendix H

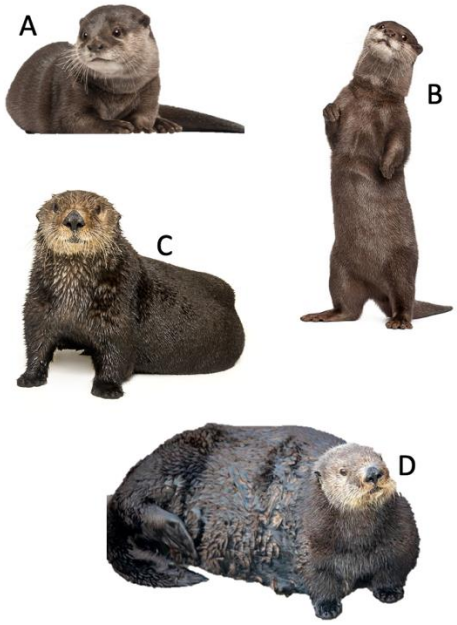
Instructions on the Catch Page

Doing online experiments

Online experiments can be risky for the experimenter because we can't be sure that participants are paying full attention to what is on the screen or responding carefully. So we urge you to put aside distractions while you complete this experiment, to read carefully what is written on each page, and to think about your answers (even though we are going to ask you some questions for which you may not have an answer). On each page you won't be able to proceed to the next page until you have answered all the questions and clicked on the button at the bottom of the page. To make sure that you have read the instructions on this page, on this page alone, instead of clicking on the button at the bottom (the obvious response) you should click on the title above in order to proceed to the next page. Although the title doesn't look like a link, it is.

Appendix I

Questions in the Animal Estimation Task with Average References in Experiment 2

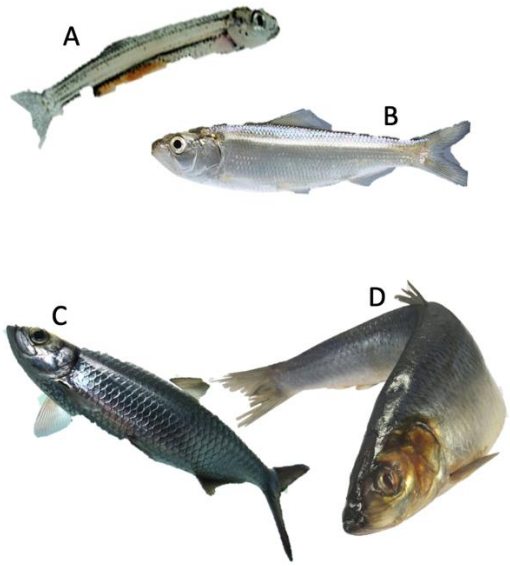


Here is a series of pictures of a Sea Otter.
Please generate the estimates to Picture A/B/C/D only:

If the average weight of the Sea Otter is 25,000 g, please estimate the weight for this Sea Otter.
_____ (g)

If the average lifespan of the Sea Otter is 4,500 days, please estimate the lifespan for this Sea Otter.
_____(days)

If the average group size of the Sea Otters in the wild is 250, please estimate the group size for this Sea Otter. _____

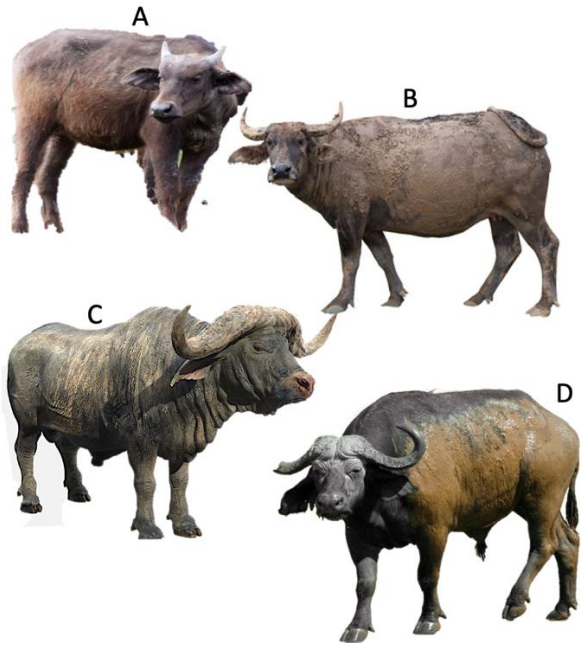


Here is a series of pictures of an Atlantic Herring.
Please generate the estimates to Picture A/B/C/D only:

If the average weight of the Atlantic Herring is 850 g, please estimate the weight for this Atlantic Herring. _____ (g)

If the average lifespan of the Atlantic Herring is 5500 days, please estimate the lifespan for this Atlantic Herring. _____(days)

If the average group size of the Atlantic Herring in the wild is 4,500,000,000, please estimate the group size for this Atlantic Herring _____

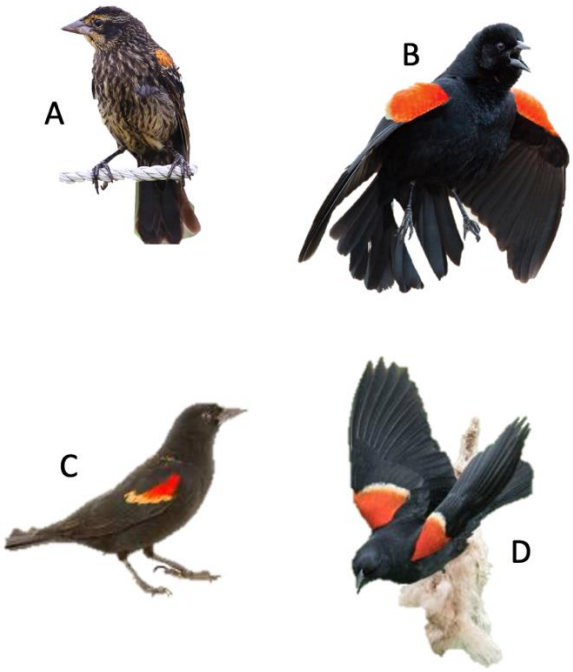


Here is a series of pictures of a Cape Buffalo. Please generate the estimates to Picture A/B/C/D only:

If the average weight of the Cape Buffalo is 750,000 g, please estimate the weight for this Cape Buffalo. ____ (g)

If the average lifespan of the Cape Buffalo is 9500 days, please estimate the lifespan for this Cape Buffalo. _____(days)

If the average group size of the Cape Buffalo in the wild is 75, please estimate the group size for this Cape Buffalo. _____

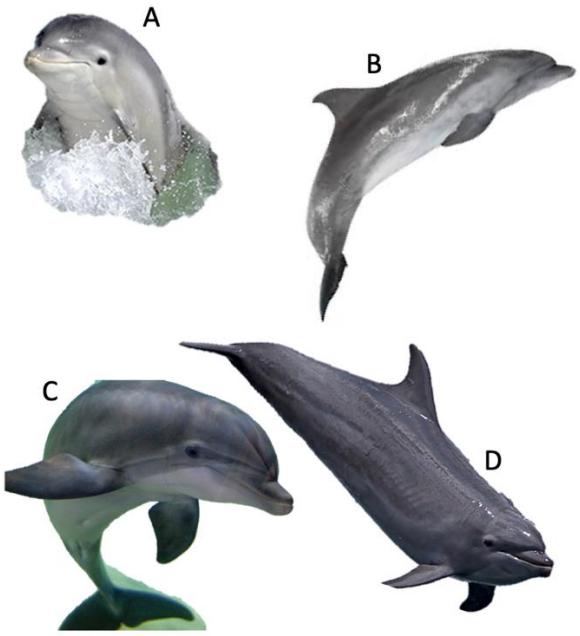


Here is a series of pictures of a Red-winged Blackbird.
Please generate the estimates to Picture A/B/C/D only:

If the average weight of the Red-winged Blackbird is 55 g, please estimate the weight for this Red-winged Blackbird. _____ (g)

If the average lifespan of the Red-winged Blackbird is 850 days, please estimate the lifespan for this Red-winged Blackbird. _____(days)

If the average group size of the Red-winged Blackbird in the wild is 650,000, please estimate the group size for this Red-winged Blackbird.
_____

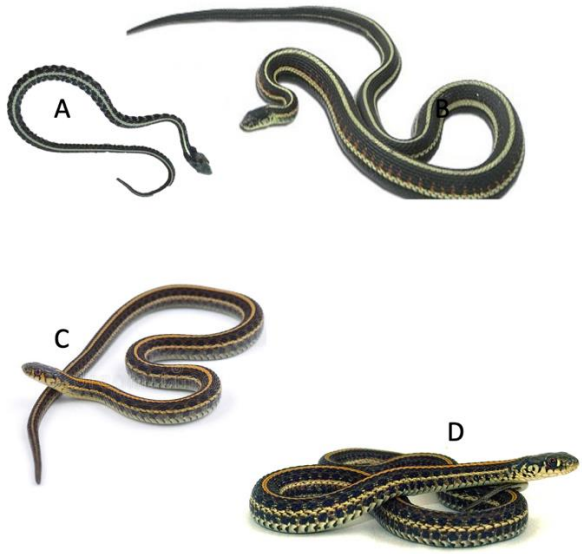


Here is a series of pictures of a Bottlenose dolphin.
Please generate the estimates to Picture A/B/C/D only:

If the average weight of the Bottlenose dolphin is 650,000 g, , please estimate the weight for this Bottlenose dolphin._____ (g)

If the average lifespan of the Bottlenose dolphin is 15,000 days, please estimate the lifespan for this Bottlenose dolphin. _____(days)

If the average group size of theBottlenose dolphin in the wild is 550, please estimate the group size for this Bottlenose dolphin. _____

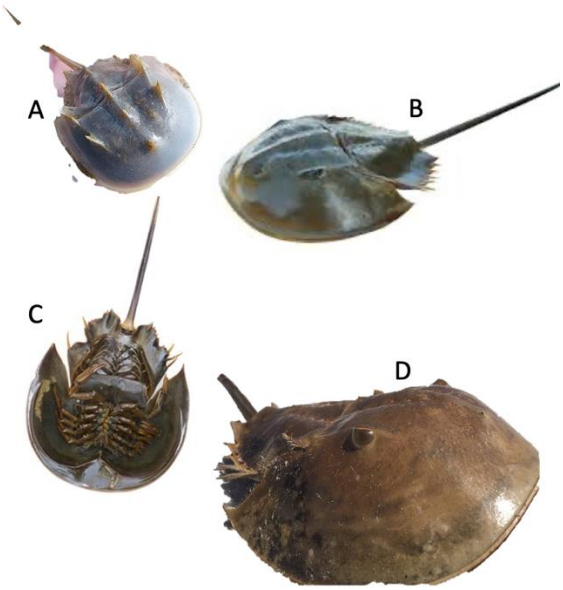


Here is a series of pictures of a Plains Garter Snake.
Please generate the estimates to Picture A/B/C/D only:

If the average weight of the Plains Garter Snake is 150 g, please estimate the weight of this Plains Garter Snake. _____ (g)

If the average lifespan of the Plains Garter Snake is 3,500 days, please estimate the lifespan for this Plains Garter Snake. _____(days)

If the average group size of the
Plains Garter Snake in the wild is 950, please estimate the group size for this Plains Garter Snake. _____

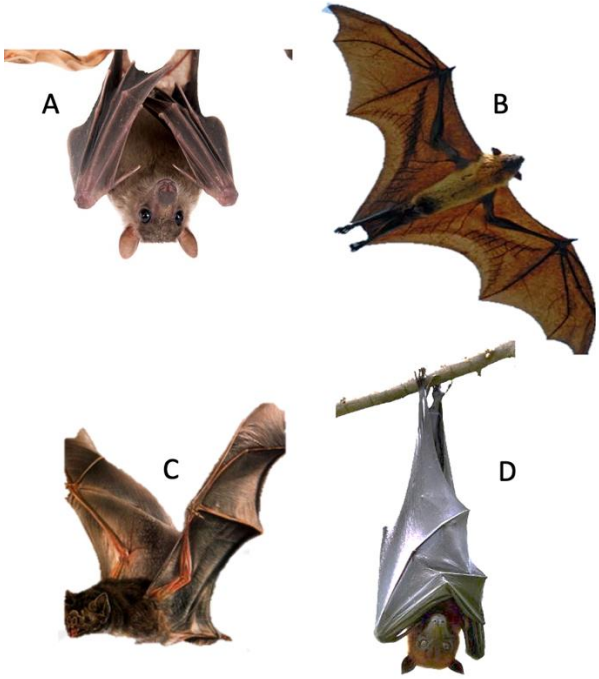


This is a picture of an Atlantic Horseshoe Crab.
Please give an estimation to Picture A/B/C/D only:

If the average weight of the Atlantic Horseshoe Crab is 4,500 g, please estimate the weight for this Atlantic Horseshoe Crab. _____ (g)

If the average lifespan of the Atlantic Horseshoe Crab is 7,500 days, please estimate the lifespan for this Atlantic Horseshoe Crab. _____(days)

If the average group size of the Atlantic Horseshoe Crab in the wild is 3,500, please estimate the group size for this Atlantic Horseshoe Crab. _____

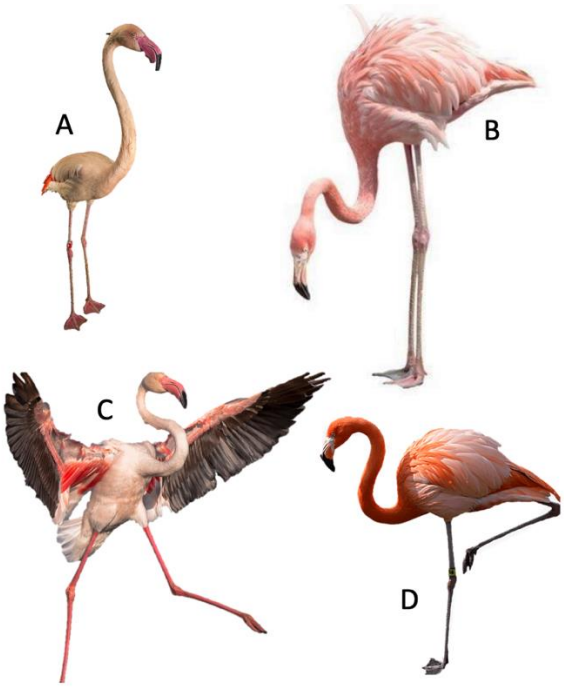


Here is a series of pictures of a Megabat.
Please generate the estimates to Picture A/B/C/D only:

If the average weight of the Megabat is 950 g, please estimate the weight for this Megabat. _____ (g)

If the average lifespan of the Megabat is 6500 days, please estimate the lifespan for this Megabat. _____(days)

If the average group size of the Megabat in the wild is 15,000, please estimate the group size for this Megabat. _____



Here is a series of pictures of a Greater Flamingo.
Please generate the estimates to Picture A/B/C/D only:

If the average weight of the Greater Flamingo is 3500 g, please estimate the weight for this Greater Flamingo. _____ (g)

If the average lifespan of the Greater Flamingo is 20500 days, please estimate the lifespan for this Greater Flamingo. _____(days)

If the average group size of the Greater Flamingo in the wild is 850, please estimate the group size for this Greater Flamingo. _____

Appendix J

Questions in the Animal Estimation Task with Juvenile References in Experiment 2

Here is a picture of a juvenile Sea Otter.
Please give an estimation to the questions below:

If the weight of this young Sea Otter is 9,500 g, please estimate its weight once it matures. _____ (g)

If this young Sea Otter has been living for 2,500 days, please estimate its total lifespan. _____ (days)

If 95 Sea Otters were observed at a point of time during their gathering in the wild, please estimate the final group size _____

It is a picture of a juvenile Atlantic Herring.
Please give an estimate to the questions below:

If the weight of this young Atlantic Herring is 650 g, please estimate its weight once it matures. _____ (g)

If this young Atlantic Herring has been living for 3500 days, please estimate its total lifespan. _____ (days)

If 2,500,000,000 Atlantic Herring were observed at a point of time during their gathering in the wild, please estimate the final group size. _____.

It is a picture of a juvenile Cape Buffalo.
Please give an estimate to the questions below:

If the weight of this young Cape Buffalo is 550,000 g, please estimate its weight once it matures. _____ (g)

If this young Cape Buffalo has been living for 7500 days, please estimate its total lifespan._____ (days)

If 55 Cape Buffalos were observed at a point of time during their gathering in the wild, please estimate the final group size. _____

It is a picture of a juvenile Red-winged Blackbird.
Please give an estimate to the questions below:

If the weight of this young Red-winged Blackbird is 35 g, please estimate its weight once it matures._____ (g)

If this young Red-winged Blackbird has been living for 650 days, please estimate its total lifespan. _____(days)

If 450,000 Red-winged Blackbirds were observed at a point of time during their gathering in the wild, please estimate the final group size. _____

It is a picture of a juvenile Bottlenose dolphin.
Please give an estimate to the questions below:

If the weight of this young Bottlenose dolphin is 450,000 g, please estimate its weight once it matures. _____ (g)

If this young Bottlenose dolphin has been living for 8,500 days, please estimate its total lifespan. ___ (days)

If 350 Bottlenose dolphins were observed at a point of time during their gathering in the wild, please estimate the final group size. _____

It is a picture of a juvenile Plains Garter Snake.
Please give an estimate to the questions below:

If the weight of this young Plains Garter Snake is 85 g, please estimate its weight once it matures. _____ (g)

If this young Plains Garter Snake has been living for 1,500 days, please estimate its total lifespan. ___ (days)

If 750 Plains Garter Snakes were observed at a point of time during their gathering in the wild, please estimate the final group size. _____

It is a picture of a juvenile Atlantic Horseshoe Crab.
Please give an estimate to the questions below:

If the weight of this young Atlantic Horseshoe Crab is 2,500 g, please estimate its weight once it matures. _____ (g)
If this young Atlantic Horseshoe Crab has been living for 5,500 days, please estimate its total lifespan. _____(days)

If 1,500 Atlantic Horseshoe Crabs were observed at a point of time during their gathering in the wild, please estimate the final group size. _____

It is a picture of a juvenile Megabat.
Please give an estimate to the questions below:

If the weight of this young Megabat is 750 g, please estimate its weight once it matures._____ (g)

If this young Megabat has been living for 4,500 days, please estimate its total lifespan. _____ (days)

If 8,500 Megabats were observed at a point of time during their gathering in the wild, please estimate the final group size. _____

It is a picture of a juvenile Greater Flamingo.
Please give an estimate to the questions below:

If the weight of this young Greater Flamingo is 1500 g, please estimate its weight once it matures. _____ (g)

If this young Greater Flamingo has been living for 9,500 days, please estimate its total lifespan. _____ (days)

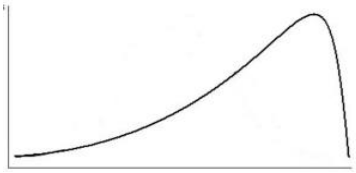
If 650 Greater Flamingos were observed at a point of time during their gathering in the wild, please estimate the final group size. _____

Appendix K

Quiz items for the Animal Estimation task in Experiment 2

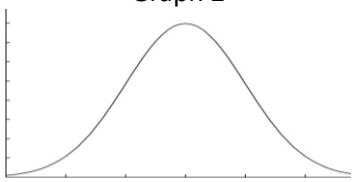
Graph 1

The distribution of lifespan is graph _____
(Answer: 1)



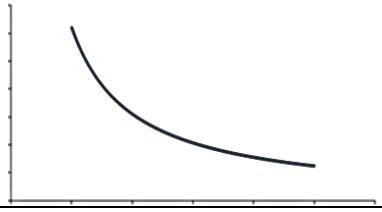
Graph 2

The distribution of group size is graph ____
(Answer: 3)



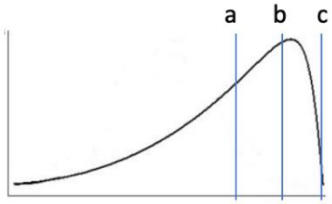
Graph 3

The distribution of weight is graph ____
(Answer: 2)



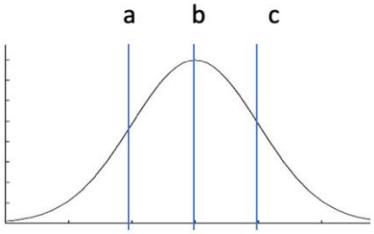
Graph 1

The average lifespan in Graph 1 is more likely to agree with the value in line _____
(Answer: a)



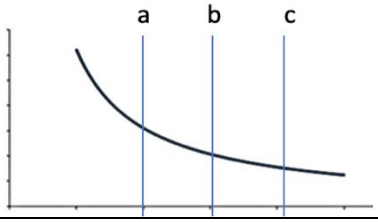
Graph 2

The average weight in Graph 2 is more likely to agree with the value in line _____
(Answer: b)



Graph 3

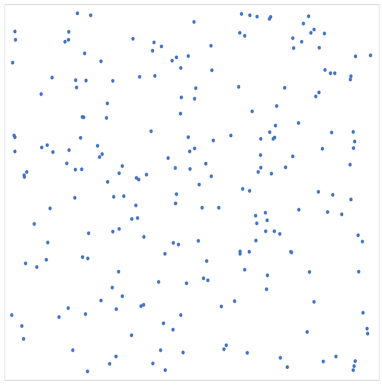
The average group-size in Graph 3 is more likely to agree with the value in line _____
(Answer: a)



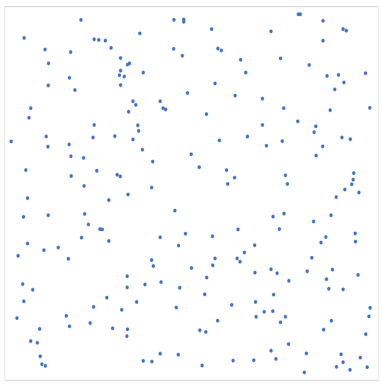
Appendix L

Items of Dots Displays Present in the Benford-like Group in Experiments 2 and 3.

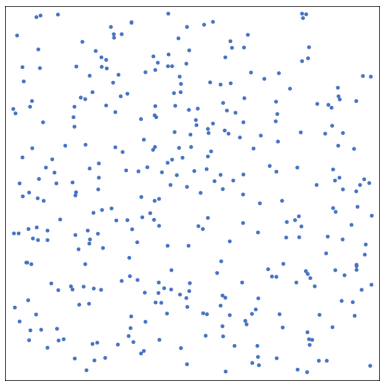
220 Dots



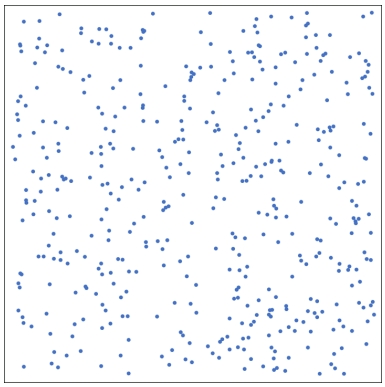
280 Dots



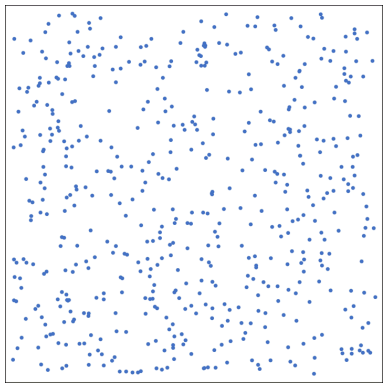
350 Dots



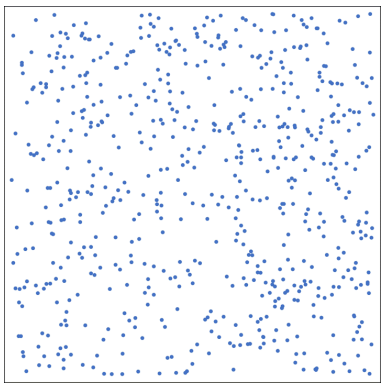
450 Dots



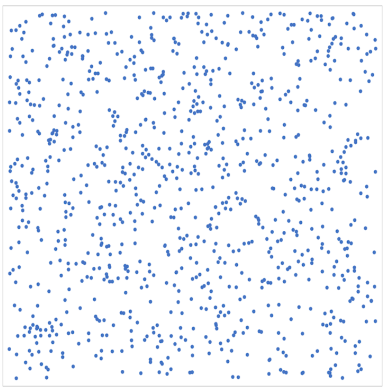
550 Dots



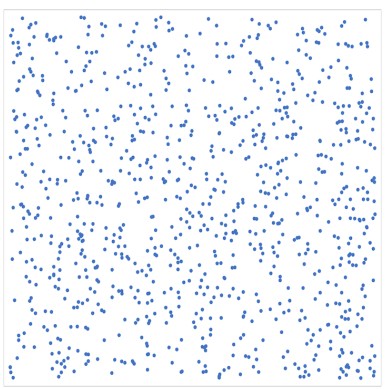
650 Dots



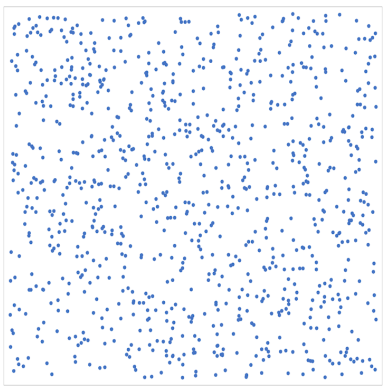
1030 Dots



1060 Dots



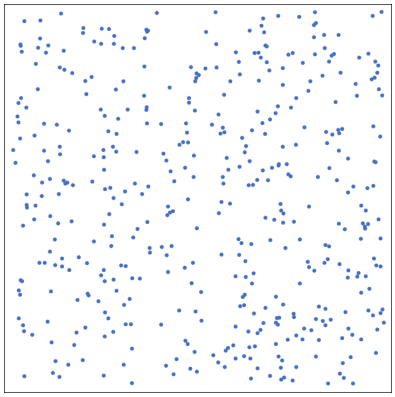
1090 Dots



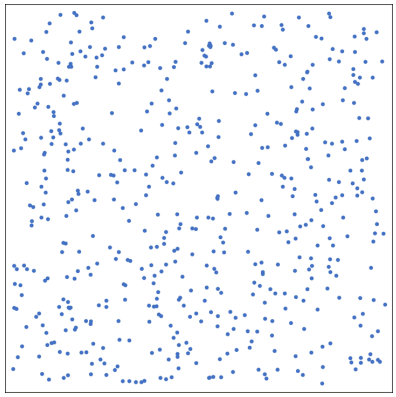
Appendix M

Items of Dots Displays Present in the Anti-Benford Group in Experiments 2 and 3.

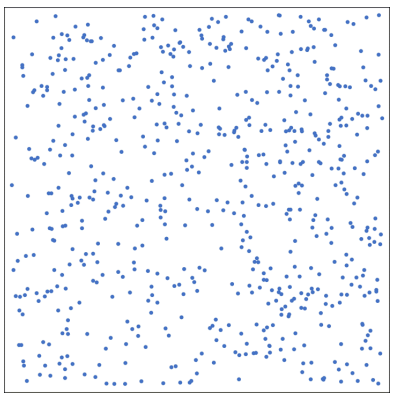
450 Dots



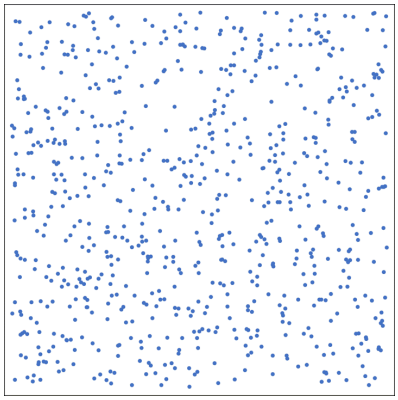
550 Dots



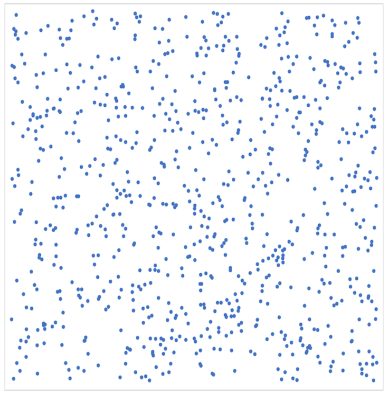
650 Dots



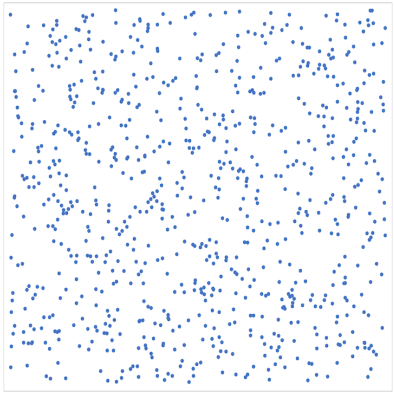
750 Dots



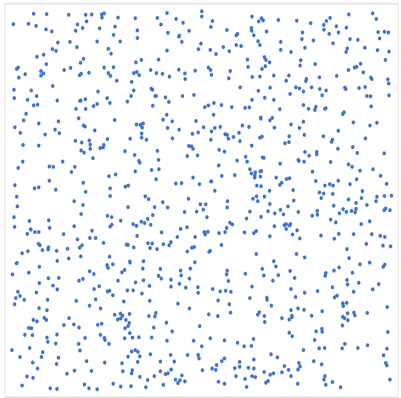
820 Dots



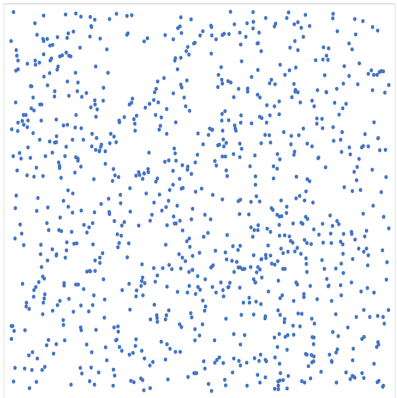
880 Dots



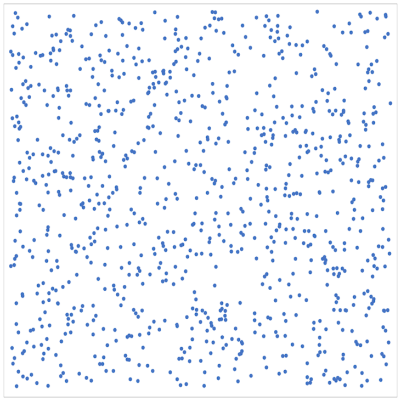
930 Dots



960 Dots



990 Dots

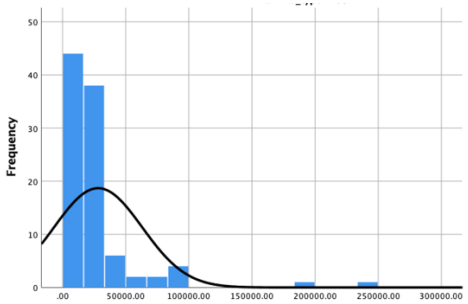


Appendix N

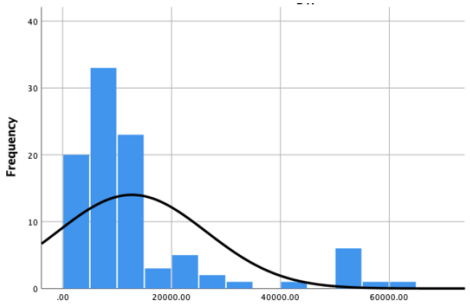
Distribution Graphs (SPSS Outputs) of Full-number Generated for Three Variables Questioned for Nine Animals in Experiment 2 based on the Juvenile References

1. Sea Otter

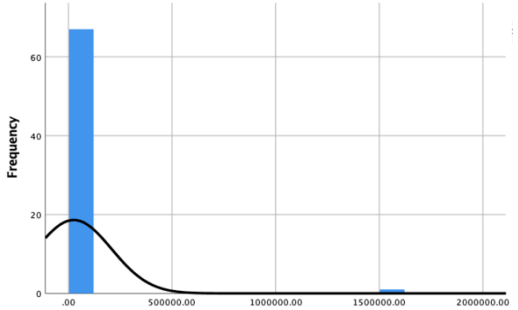
Weight



Lifespan

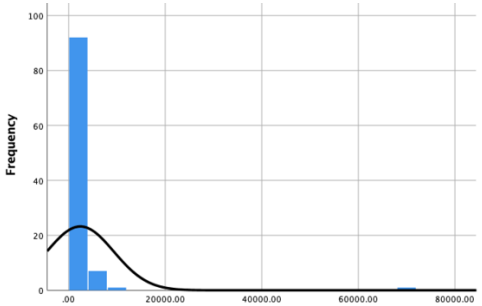


Group-size

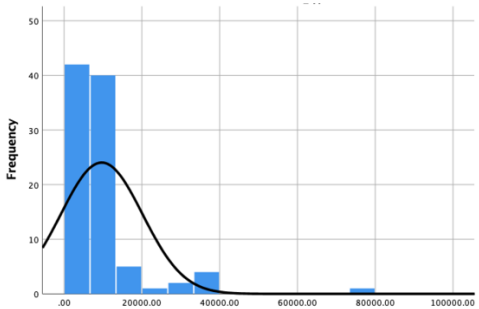


2. Atlantic Herring:

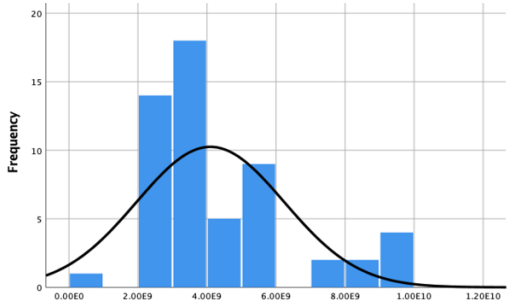
Weight



Lifespan

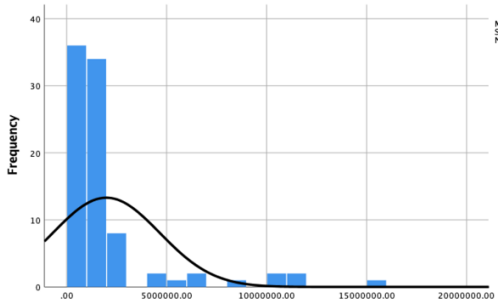


Group-size

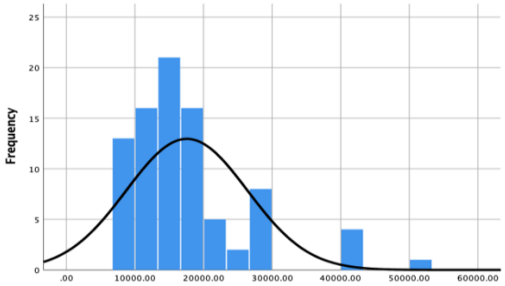


3. Cape Buffalo

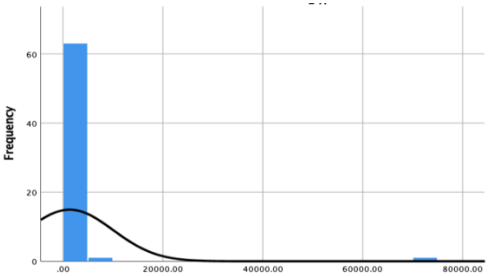
Weight



Lifespan

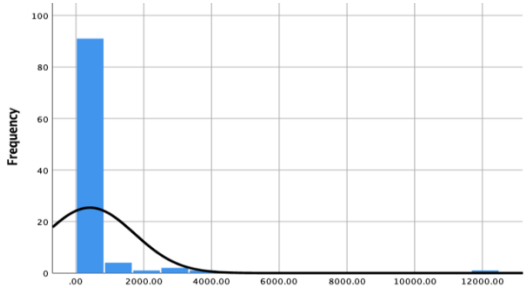


Group-size

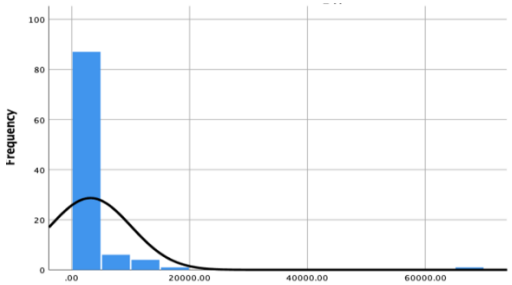


4. Red-winged Blackbird

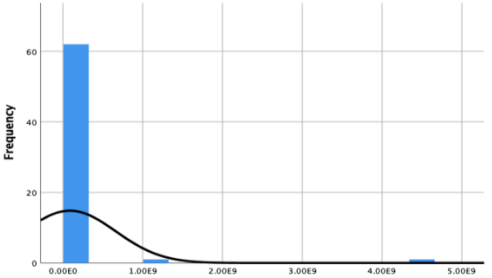
Weight



Lifespan

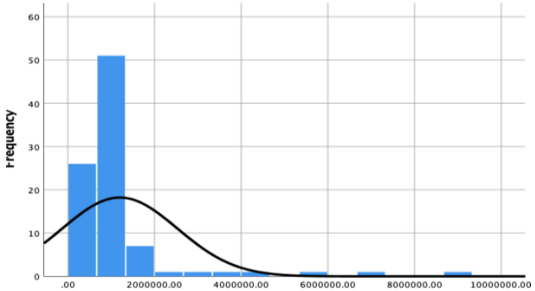


Group-size

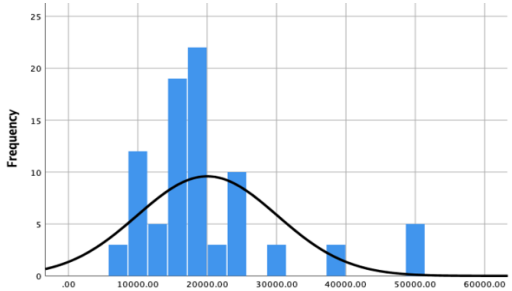


5. Bottlenose Dolphin

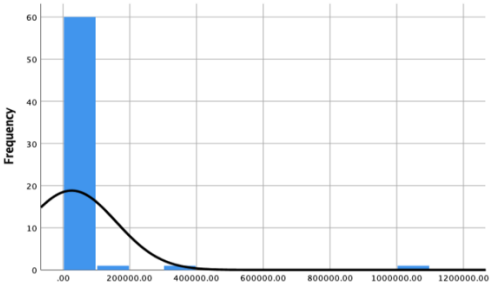
Weight



Lifespan

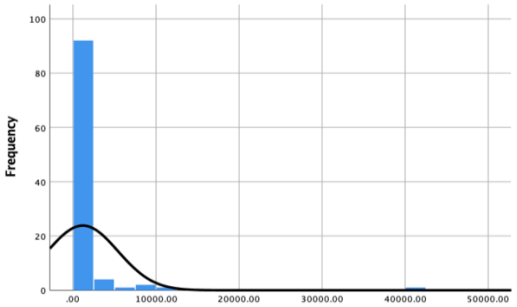


Group-size

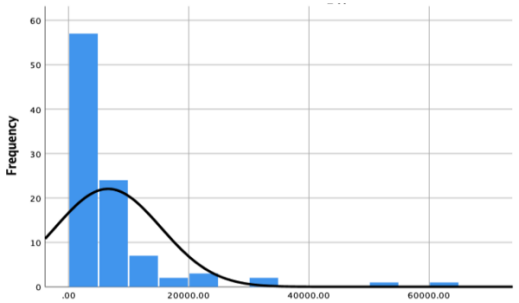


6. Plains Garter Snake

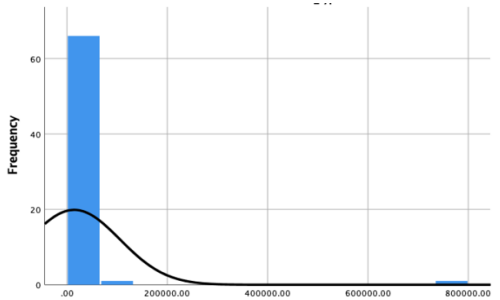
Weight



Lifespan

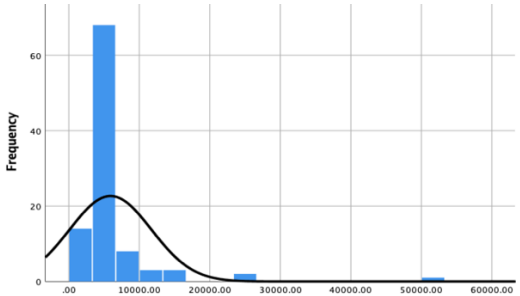


Group-size

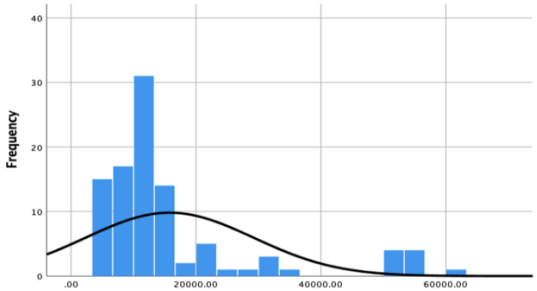


7. Atlantic Horseshoe Crab

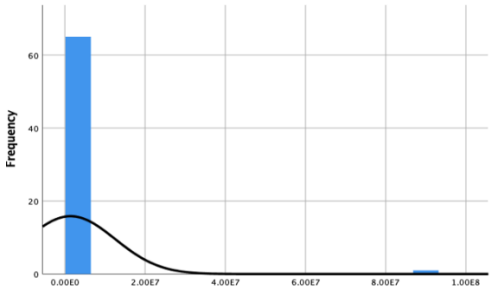
Weight



Lifespan

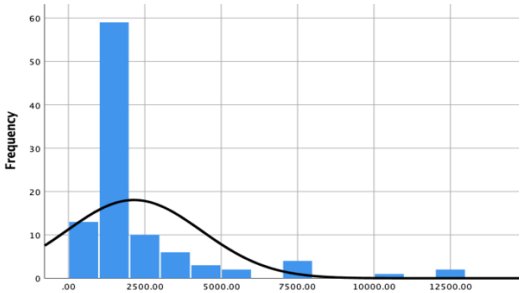


Group-size

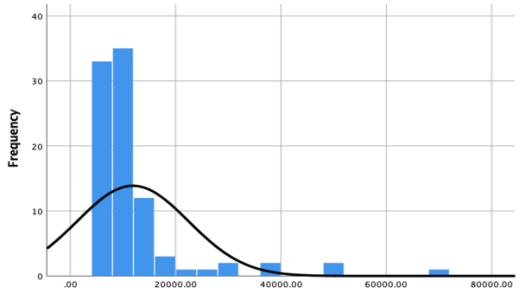


8. Megabat

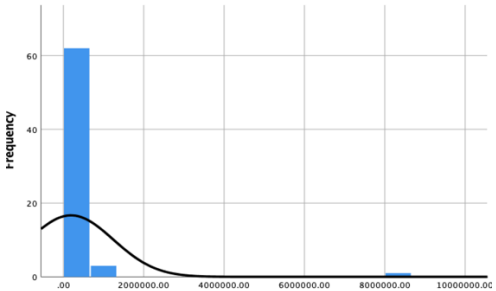
Weight



Lifespan

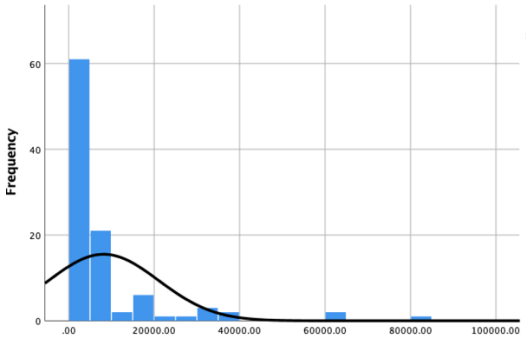


Group-size

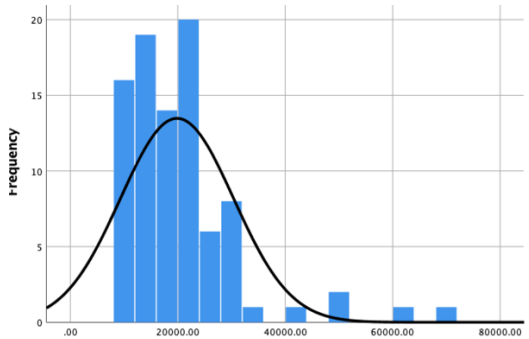


9. Greater Flamingo

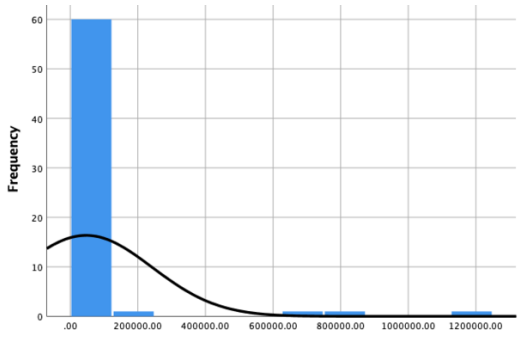
Weight



Lifespan



Group-size

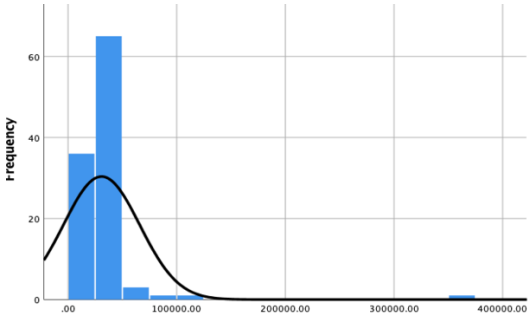


Appendix O

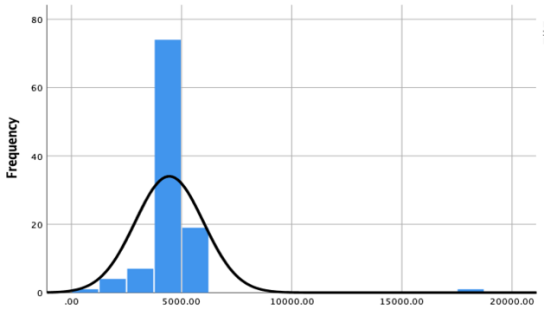
Distribution Graphs (SPSS Outputs) of Full-number Generated for Three Variables Questioned for Nine Animals in Experiment 2 based on the Average References

1. Sea Otter

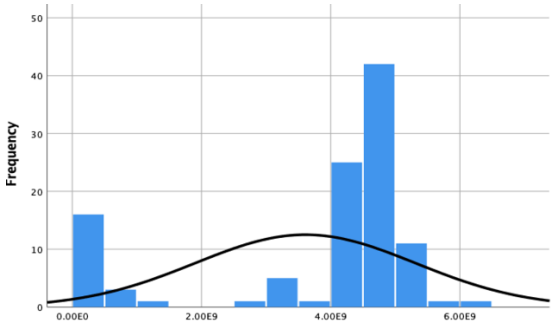
Weight



Lifespan

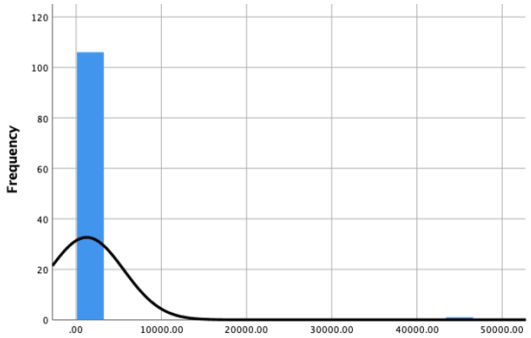


Group-size

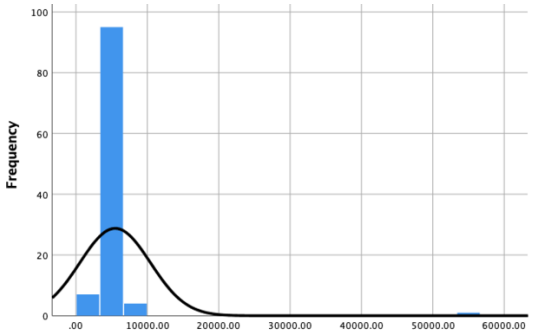


2. Atlantic Herring

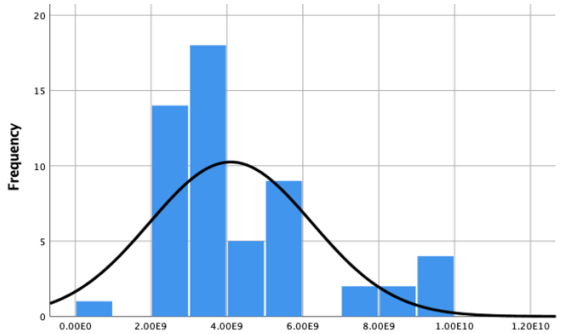
Weight



Lifespan

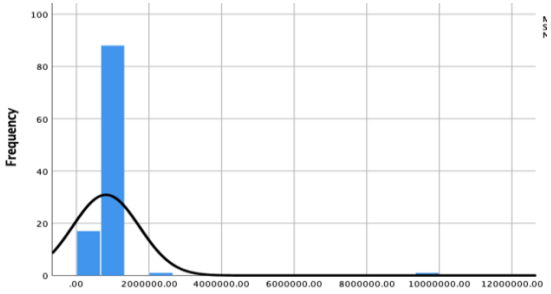


Group-size

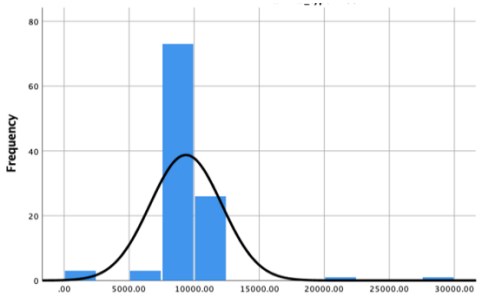


3. Cape Buffalo

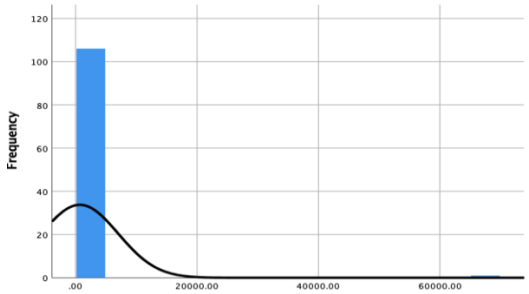
Weight



Lifespan

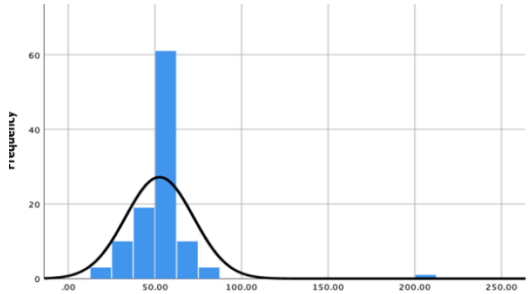


Group-size

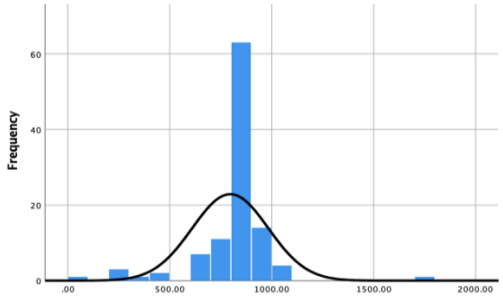


4. Red-winged Blackbird

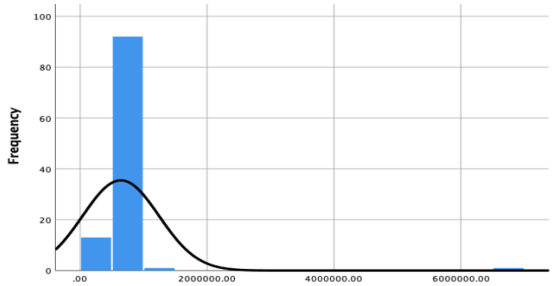
Weight



Lifespan

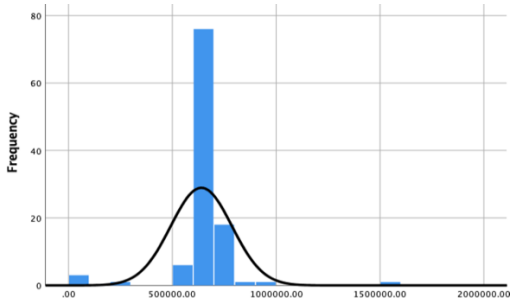


Group-size

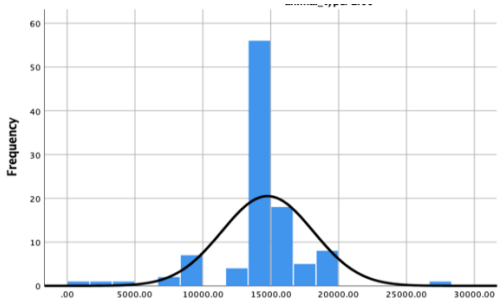


5. Bottlenose Dolphin

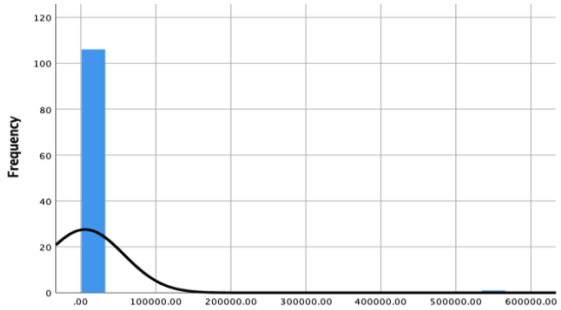
Weight



Lifespan

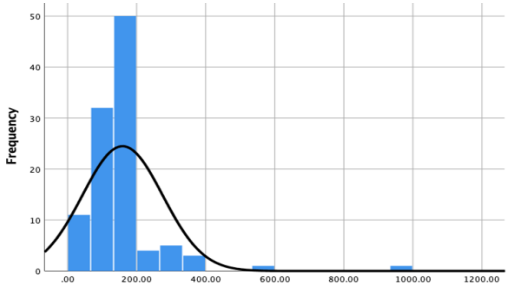


Group-size

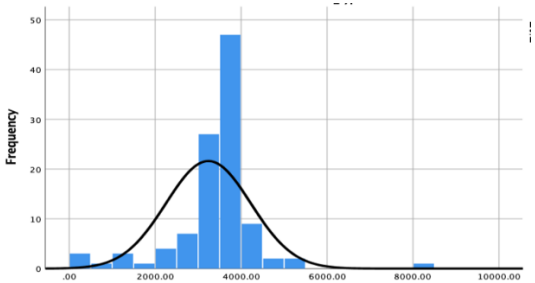


6. Plains Garter Snake

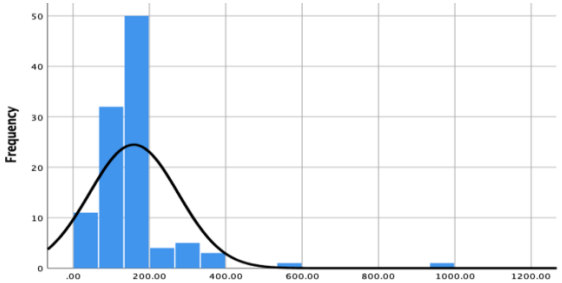
Weight



Lifespan

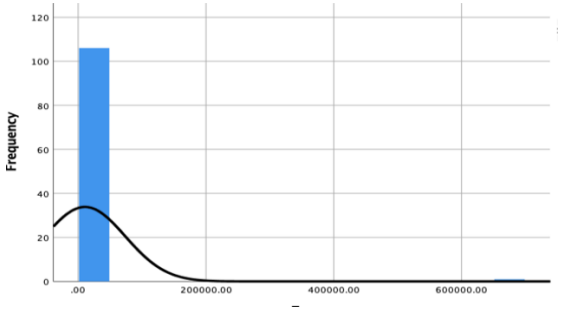


Group-size

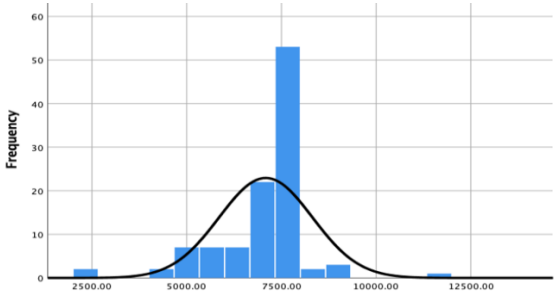


7. Atlantic Horseshoe Crab

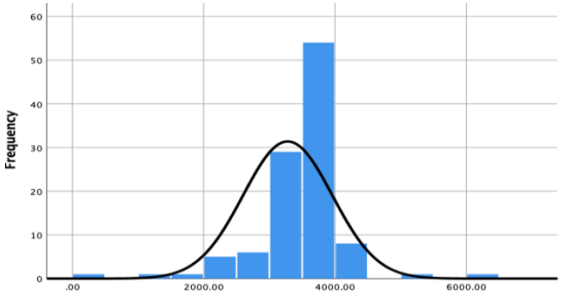
Weight



Lifespan

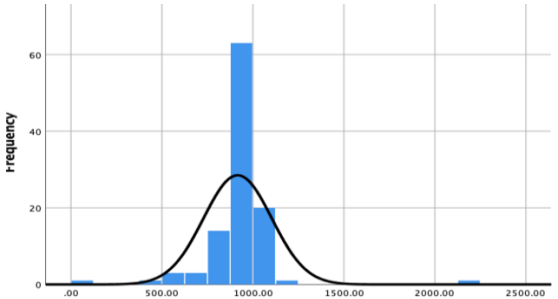


Group-size

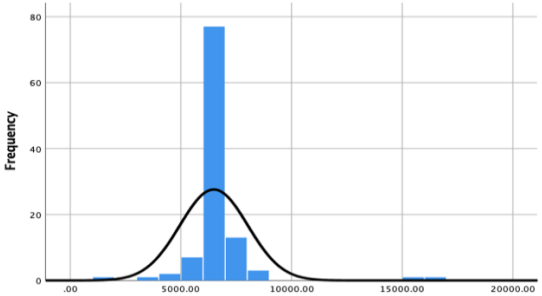


8. Megabat

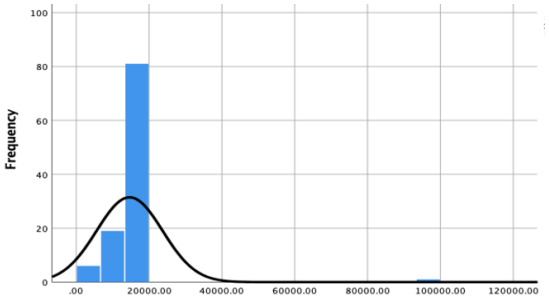
Weight



Lifespan

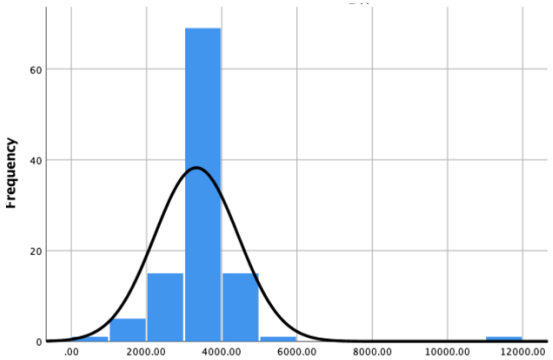


Group-size

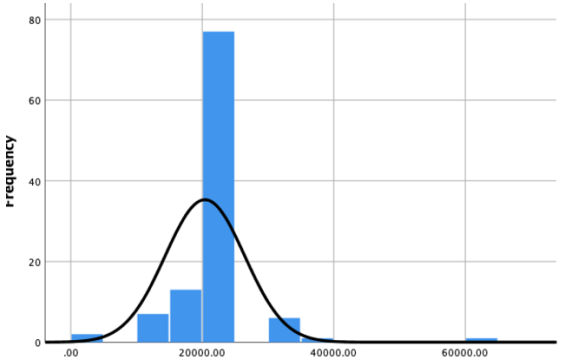


9. Greater Flamingo

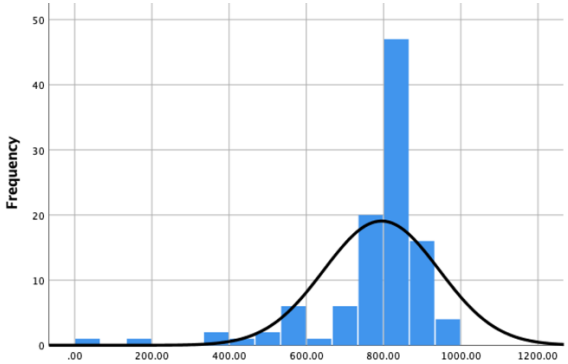
Weight



Lifespan



Group-size



Appendix P

The List of item pool for Video Clips Present in Experiments 4a and 4b, Partially for the Preliminary Study

750 Audience:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Audience_750.mp4

850 Audience:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Audience_850.mp4

950 Audience:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Audience_950.mp4

4500 Audience:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Audience_4500.mp4

5500 Audience:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Audience_5500.mp4

6500 Audience:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Audience_6500.mp4

15000 Audience:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Audience_15000.mp4

25000 Audience:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Audience_25000.mp4

35000 Audience:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Audience_35000.mp4

750 Birds:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Birds_750.mp4

850 Birds:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Birds_850.mp4

950 Birds:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Birds_950.mp4

4500 Birds:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Birds_4500.mp4

5500 Birds:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Birds_5500.mp4

6500 Birds:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Birds_6500.mp4

15000 Birds:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Birds_15000.mp4

25000 Birds:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Birds_25000.mp4

35000 Birds:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Birds_35000.mp4

750 Cheering_Men:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Cheering_Men_750.mp4

850 Cheering_Men:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Cheering_Men_850.mp4

950 Cheering_Men:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Cheering_Men_950.mp4

4500 Cheering_Men:
https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Cheering_Men_4500.mp4

5500 Cheering_Men:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Cheering_Men_5500.mp4

6500 Cheering_Men:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Cheering_Men_6500.mp4

15000 Cheering_Men:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Cheering_Men_15000.mp4

25000 Cheering_Men:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Cheering_Men_25000.mp4

35000 Cheering_Men:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Cheering_Men_35000.mp4

750 Crows:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Crows_750.mp4

850 Crows:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Crows_850.mp4

950 Crows:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Crows_950.mp4

4500 Crows:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Crows_4500.mp4

5500 Crows:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Crows_5500.mp4

6500 Crows:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Crows_6500.mp4

15000 Crows:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Crows_15000.mp4

25000 Crows:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Crows_25000.mp4

35000 Crows:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Crows_35000.mp4

750 Dancers:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Dancers_750.mp4

850 Dancers:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Dancers_850.mp4

950 Dancers:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Dancers_950.mp4

4500 Dancers:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Dancers_4500.mp4

5500 Dancers:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Dancers_5500.mp4

6500 Dancers:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Dancers_6500.mp4

15000 Dancers:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Dancers_15000.mp4

25000 Dancers:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Dancers_25000.mp4

35000 Dancers:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Dancers_35000.mp4

750 Moving_Dots:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Moving_Dots_750.mp4

850 Moving_Dots:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Moving_Dots_850.mp4

950 Moving_Dots:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Moving_Dots_950.mp4

4500 Moving_Dots:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Moving_Dots_4500.mp4

5500 Moving_Dots:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Moving_Dots_5500.mp4

6500 Moving_Dots:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Moving_Dots_6500.mp4

15000 Moving_Dots:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Moving_Dots_15000.mp4

25000 Moving_Dots:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Moving_Dots_25000.mp4

35000 Moving_Dots:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Moving_Dots_35000.mp4

750 Protestors:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Protestors_750.mp4

850 Protestors:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Protestors_850.mp4

950 Protestors:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Protestors_950.mp4

4500 Protestors:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Protestors_4500.mp4

5500 Protestors:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Protestors_5500.mp4

6500 Protestors:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Protestors_6500.mp4

15000 Protestors:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Protestors_15000.mp4

25000 Protestors:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Protestors_25000.mp4

35000 Protestors:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Protestors_35000.mp4

750 Skeletons:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Skeletons_750.mp4

850 Skeletons:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Skeletons_850.mp4

950 Skeletons:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Skeletons_950.mp4

4500 Skeletons:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Skeletons_4500.mp4

5500 Skeletons:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Skeletons_5500.mp4

6500 Skeletons:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Skeletons_6500.mp4

15000 Skeletons:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Skeletons_15000.mp4

25000 Skeletons:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Skeletons_25000.mp4

35000 Skeletons:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Skeletons_35000.mp4

750 Whales:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Whales_750.mp4

850 Whales:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Whales_850.mp4

950 Whales:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Whales_950.mp4

4500 Whales:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Whales_4500.mp4

5500 Whales:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Whales_5500.mp4

6500 Whales:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Whales_6500.mp4

15000 Whales:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Whales_15000.mp4

25000 Whales:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Whales_25000.mp4

35000 Whales:

https://psi-cognitive.sydney.edu.au/staff/bburns/Program_23s1dc/clips/Whales_35000.mp4

Appendix Q

The Warning Instruction Page in Experiment 4

Warning

This online experiment will need to present a series of short video clips, which might consume a considerable amount of your internet data - **EACH will take between 2 MB and 32 MB. The total download will be about 400 MB.**

The content of clips might contain **a variety of different colours/shapes in a dense crowd**.

It requires a **stable** internet connection to run the tasks smoothly.

Please use a **laptop/PC** for this online experiment, not a phone or tablet device.

If any of these issues are a concern for you then you can withdraw from the experiment.

Appendix R

Instructions for the Feedback Page at the End of Experiment 4

Feedback Page

We appreciate it if you can provide some comments and feedback below.

For example: Have you experienced any problems? Have you done a similar task before? Feelings and thoughts about the visual estimation and your performance?