The cognitive and neural bases of graph perception and comprehension

Lorenzo Ciccione

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée au Collège de France

Les bases cognitives et neurales de la perception et compréhension des graphiques

The cognitive and neural bases of graph perception and comprehension

Soutenue par
Lorenzo Ciccione
Le 19/12/2022

Ecole doctorale n° 474
Frontières de l'innovation en recherche et éducation

Spécialité
Sciences cognitives
Cognitive science

COLLÈGE DE FRANCE
1530

Composition du jury:

Pascal, MAMASSIAN Research director, Ecole Normale Supérieure, PSL	*Président*
Manuela, PIAZZA Full professor, Università degli Studi di Trento	*Rapporteur*
Steven, FRANCONERI Full professor, Northwestern University	*Rapporteur*
Elizabeth, SPELKE Full professor, Harvard University	*Examinateur*
Véronique, IZARD Researcher, Université de Paris	*Examinateur*
Stanislas, DEHAENE Full professor, Collège de France, PSL	*Directeur de thèse*

TABLE OF CONTENTS

ACKNOWLEDGMENTS

As any scientific endeavor, work is never accomplished alone and many people deserve my gratitude and recognition. First and above all, my great supervisor, professor Stanislas Dehaene (Stan). Working with him is one of the greatest achievements of my educational and professional career and every meeting with him is a true learning experience. He taught me to be rigorous, passionate and ambitious. I sincerely thank Stan for letting me work with him. And thanks to the ENS, the LPI and the Mind Science Foundation for funding my work.

I then want to thank all the people in the lab I have been working in (Neurospin/Unicog): Mathias for the long discussions, the meta-science thoughts and the work that we have done together (part of this thesis' work is the result of a wonderful collaboration with him); Tiffany for the coffee breaks (with no coffee), the walks, and the philosophical quests for meaning (and gossip); Vanna for the true help, the politically incorrect conversations, and the presence; Marie for her coding skills, her cooking skills, and her friendship skills; Cassandra, to be an inspiration for my oral presentations, and for her contagious smile and laughs. Special thanks as well to my PhD adventure colleagues: Yvan, Theo M., Alexis, Harish, Lucas, Caroline, Cedric, Maxime, Theo D., Alex, Christos, Audrey, Andrea, Pauline, Marie, Valentine, François, Alexandre, Timo. Huge thanks for their incredible help to Minye, Antonio, Christophe, Evelyn, Isabelle. And thanks to all people working in participants' recruitment, HR, and administration and to the hundreds of people that participated in my experiments.

I would also like to thank: my thesis' advisors (Veronique and Valerian); the people from my doctoral school (LPI, ED 474: Camille, Chiara, Ana, Dragana); my dear friends from the ENS, where I spent three wonderful years for my master studies: Marco, Camille, Ioanna, Emma, Lena, Michele, Zeynep, Quentin; and my science/beer friends: Camille and Raph.

Another round of important acknowledgments: in Paris, thanks to Basak, Garance and Izel, the friends that everyone should deserve but not many people have; you are an inspiration of loyalty, kindness, and love. Thanks to my Italian Parisian friends: Nicola and Gauthier, who helped me de-connect from my French/English brain. Thanks to my friends from Andora, the small town I grew up in: Michela, Giulia, Federica for making my days there happy and fun. Thanks to Marianna, the best friend that was always there: we saw each other grow up along these years and I am proud of who we are. In Trento, where I spent my undergraduate years, thanks to everyone I had the privilege to live with (Francesco, Linda and Chenfu among others) and especially to Dario, one of the few people in my life that would always answer my calls, no matter where and when: meeting him was a gift.

A special note to those who allowed me to teach at university (and especially to Serge, who also became a mentor and, together with Mathilde and Esther, a collaborator) and to the many students I had (Université de Paris, Paris 8, IA School, EDC Business School and American University of Paris), since you made me realize that teaching is the career I want to purse in life. Thanks as well to the teachers I had in school and at university, and particularly to the good ones (they made the difference).

Thanks to Thomas, without whom these years would not have been the same: he made me travel not just out of Paris, but of my comfort zone. Thanks to my cat Cleo, who has the most beautiful and caring eyes I have ever met. Thanks to my papà Gianfranco, who taught me how to stand up and be proud of my ideas and abilities, how to forget bad things fast, and how to be always ambitious. Thanks to my mamma Ornella, who taught me what it means to love and being loved, the importance of education and culture, the value of coherence and that I am never alone. Lastly, thanks to child and teenager Lorenzo, who was strong and patient and understood he was brave, lucky, and smart enough to keep on going.

STATEMENT OF ORIGINALITY

The work presented in this manuscript has been conducted by myself, under the guidance of my supervisor, Professor Stanislas Dehaene, and sometimes in collaboration with several colleagues: Mathias Sablé-Meyer (software implementation of study 2 and study 9; modelling of study 9); Guillaume Dehaene (algorithm of outlier creation in study 7); Serge Caparos, Mathilde Josserand, Esther Boissin (data collection in Namibia for study 3); Cassandra Potier-Watkins (data collection at school for study 4).

Most of the findings described in this manuscript have already been published during the three years of my PhD program and can be found here:

Ciccione, L., & Dehaene, S. (2020). Grouping Mechanisms in Numerosity Perception. *Open Mind*, *4*, 102–118. https://doi.org/10.1162/opmi_a_00037

Ciccione, L., & Dehaene, S. (2021). Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots. *Cognitive Psychology*, *128*, 101406. https://doi.org/10.1016/j.cogpsych.2021.101406

Ciccione, L., Sablé-Meyer, M., & Dehaene, S. (2022). Analyzing the misperception of exponential growth in graphs. *Cognition*, *225*, 105112. https://doi.org/10.1016/j.cognition.2022.105112

Ciccione, L., Dehaene, G., & Dehaene, S. (2022). Outlier detection and rejection in scatterplots: Do outliers influence intuitive statistical judgments? *Journal of Experimental Psychology: Human Perception and Performance.* https://doi.org/10.1037/xhp0001065

Ciccione, L., & Dehaene, S. (2022). Graphicacy skills across ages and cultures: a new assessment tool of intuitive statistical abilities. *Proceedings of the Annual Meeting of the Cognitive Science Society.* (Vol. 44, No. 44).

Ciccione, L., Sablé-Meyer, M., Boissin, E., Josserand, M., Potier-Watkins, C., Caparos, S., & Dehaene, S. (2022). Graphicacy across age, education, and culture: a new tool to assess intuitive graphics skills. *BioRxiv*

The manuscript does not follow the strict order of the articles. In order to improve readability, I re-organized the single studies in the way that I considered most logical and coherent. Therefore, studies from a single article might be presented in different chapters.

SUMMARY

Despite being a recent cultural product, graphs are ubiquitous in our life: they appear in newspapers, in school books, on television. Economic reports and scientific papers usually include multiple graphical representations, and professionals spend a considerable amount of time conceiving and designing charts and plots. The extraordinary increase in data availability, together with powerful digital visualization techniques, made graphs one of the most suited tools to transmit complex information in a complete and efficient manner. A growing body of research is investigating how to make graphs easier to read and more memorable, what graphical features guide readers' attention, and how people make subjective estimations of correlation when facing a graph. However, little is known about the cognitive bases of our ability to intuitively extract statistical information from graphs. In a series of 9 behavioral studies, we applied psychophysical methods and analyses to characterize the perception of one of the simplest and most used graphical representation: the scatterplot.

With a novel trend judgment task (study 1, N=10: "does this graph go up or down?"), I was able to model participants' accuracy and response time as a function of the t-value of the graph, showing that our perceptual system acts closely to an optimal observer that would base its decision on a statistical test. Then, I psychometrically operationalized human abilities at performing such task through what I called the "graphicacy index", and found that such measure widely varies in a large-scale online sample and strongly correlates with mathematical knowledge (study 2, N=3943). I also replicated the reliance on the t-value in two specific populations: the Himba people, who live in remote villages in Namibia and do not receive formal schooling (study 3, N=87); and 6-years-old French children attending their

first year of primary school, who were never taught what a scatterplot was (study 4, N=27). Taken together, these results suggest that the cognitive precursors of graph perception, although variable, are available independently from education, age, and culture.

With a line adjustment task (study 5, N=10: “adjust the line to make it pass through the points”) I further characterized human mental regression, by finding that humans do not minimize the vertical distance of the points from the fit (as it would be predicted by a classic OLS model) but they rather minimize the orthogonal distance (as done by Deming regression). I replicated such Deming bias with an extrapolation task (study 6, N=10: “place a point as the best continuation of the plot”) and I explored its concrete implications.

I then probed the limits of mental regression by investigating whether it was robust to the presence of outliers (study 7, N=30), as it would be predicted in the literature if graphs were treated like an ensemble. I found that humans spontaneously include outliers in their intuitive trend judgments and line adjustments and are able to exclude those with a large z score, only if explicitly asked to pay attention to them.

I lastly investigated whether humans were able to go beyond linearity, by asking them to extrapolate from noisy scatterplots generated from non-linear functions (study 8, N=10): I found that they could, with a performance dependent on the level of evidence in the plot, except for quadratics, which were underestimated. I then considered another example of accelerating functions that are known to be misperceived: exponentials (study 9, N=625). Participants underestimated them but only if hidden in noise, suggesting that human bias at extrapolating from complex accelerating functions do not derive from a lack of understanding of the function itself, but rather from an inability to extract it from noise. I also found that such bias correlates with mathematical knowledge and, on the basis of a Bayesian simulation, that is likely to be due to a prior against non-linear functions.

RESUME

Bien qu'ils soient un produit culturel récent, les graphiques sont omniprésents dans notre vie: ils apparaissent dans les journaux, dans les manuels scolaires, à la télévision. Les rapports économiques et les articles scientifiques comportent généralement de multiples représentations graphiques, et les professionnels passent un temps considérable à concevoir et à dessiner des diagrammes et des graphiques. L'extraordinaire augmentation de la disponibilité des données, ainsi que les puissantes techniques de visualisation numérique, ont fait des graphiques l'un des outils les plus adaptés pour transmettre des informations complexes de manière complète et efficace. De plus en plus de recherches s'intéressent à la façon de rendre les graphiques plus faciles à lire et plus mémorables, aux aspects qui guident l'attention des lecteurs et à la façon dont les gens font des estimations subjectives de la corrélation. Cependant, on sait peu de choses sur les bases cognitives de notre capacité à extraire intuitivement des informations statistiques des graphiques. Dans une série de 9 études comportementales, nous avons appliqué des méthodes et des analyses psychophysiques pour caractériser la perception de l'une des représentations graphiques les plus simples et les plus utilisées: le nuage de points.

Avec une nouvelle tâche de jugement de tendance (étude 1, N=10: "ce graphique va-t-il vers le haut ou vers le bas ?"), j'ai pu modéliser la précision et le temps de réponse des participants en fonction de la valeur *t* du graphique, montrant que notre système perceptif agit de manière proche d'un observateur optimal qui baserait sa décision sur un test statistique. Ensuite, j'ai opérationnalisé psychométriquement les capacités humaines à effectuer une telle tâche par le biais de ce que j'ai appelé "indice de graphicacité", et j'ai constaté que cette mesure varie largement (dans un grand échantillon en ligne) et qu'elle est fortement corrélée

aux connaissances mathématiques (étude 2, N=3943). J'ai également reproduit le recours à la valeur *t* dans deux populations spécifiques: chez les Himba, qui vivent dans des villages reculés de Namibie et qui ne reçoivent pas d'éducation formelle (étude 3, N=87); et chez les enfants français de 6 ans en première année d'école primaire, à qui on n'a jamais appris ce qu'était un nuage de points (étude 4, N=27). L'ensemble de ces résultats suggère que les précurseurs cognitifs de la perception des graphes, bien que variables, sont disponibles indépendamment de l'éducation, de l'âge et de la culture.

Avec une tâche d'ajustement de ligne (étude 5, N=10: "ajuster la ligne pour qu'elle passe entre les points"), j'ai caractérisé davantage la régression mentale humaine, en trouvant que les participants ne minimisent pas la distance verticale des points par rapport à la droite (comme prédit par un modèle classique OLS) mais ils minimisent plutôt la distance orthogonale (comme fait la régression de Deming). J'ai reproduit ce biais de Deming avec une tâche d'extrapolation (étude 6, N=10: "placer un point comme continuation du graphique") et j'ai exploré ses implications concrètes.

J'ai ensuite sondé les limites de la régression mentale en cherchant à savoir si elle était robuste à la présence de valeurs aberrantes (étude 7, N=30), comme cela serait prédit dans la littérature si les graphiques étaient traités comme un ensemble. J'ai constaté que les humains incluent spontanément les valeurs aberrantes dans leurs jugements intuitifs de tendance et leurs ajustements de lignes et qu'ils sont capables d'exclure celles qui ont un score z élevé, uniquement si on leur demande explicitement d'y prêter attention.

Enfin, j'ai cherché à savoir si les humains étaient capables d'aller au-delà de la linéarité, en leur demandant d'extrapoler à partir de nuages de points bruités générés par des fonctions non linéaires (étude 8, N=10): j'ai constaté qu'ils le pouvaient, avec une performance dépendant du niveau d'évidence du graphique, sauf pour les quadratiques, qui étaient sous-

estimées. J'ai ensuite considéré un autre exemple de fonctions connues pour être mal perçues: les exponentielles (étude 9, N=625). Les participants les ont sous-estimées, mais seulement si elles étaient cachées dans le bruit, ce qui suggère que le biais humain à extrapoler à partir de fonctions complexes ne provient pas d'un manque de compréhension de la fonction elle-même, mais plutôt d'une incapacité à l'extraire du bruit. J'ai également constaté que ce biais est corrélé aux connaissances mathématiques et, sur la base d'une simulation bayésienne, qu'il est probablement dû à un a priori contre les fonctions non linéaires.

RIASSUNTO

Nonostante siano un prodotto culturale recente, i grafici sono onnipresenti nella nostra vita: compaiono sui giornali, nei libri di scuola, in televisione. I rapporti economici e i documenti scientifici spesso includono molteplici rappresentazioni grafiche e i professionisti dedicano una notevole quantità di tempo a ideare e progettare grafici e diagrammi. Lo straordinario aumento della disponibilità di dati, insieme alle potenti tecniche di visualizzazione digitale, ha reso i grafici uno degli strumenti più adatti a trasmettere informazioni complesse in modo completo ed efficiente. Un numero crescente di ricerche sta studiando come rendere i grafici più facili da leggere e più memorabili, quali caratteristiche grafiche guidano l'attenzione dei lettori e come le persone effettuano stime soggettive di correlazione quando si trovano di fronte a un grafico. Tuttavia, poco si sa sulle basi cognitive della nostra capacità di estrarre intuitivamente informazioni statistiche dai grafici. In una serie di 9 studi comportamentali, abbiamo applicato metodi e analisi psicofisiche per caratterizzare la percezione di una delle rappresentazioni grafiche più semplici e più utilizzate: il diagramma di dispersione. Con un nuovo compito di giudizio di tendenza (studio 1, N=10: "questo grafico sale o scende?"), sono stato in grado di modellare l'accuratezza e il tempo di risposta dei partecipanti in funzione del valore *t* del grafico, dimostrando che il nostro sistema percettivo si comporta in modo simile a un osservatore ottimale che basa la sua decisione su un test statistico. In seguito, ho operazionalizzato psicometricamente le capacità umane a svolgere tale compito attraverso quello che ho chiamato "indice di graficità", e ho scoperto che tale misura varia ampiamente in un campione online su larga scala ed è fortemente correlata alle conoscenze matematiche (studio 2, N=3943). Ho anche replicato la dipendenza dal valore *t* in due popolazioni specifiche: il popolo Himba, che vive in villaggi remoti della Namibia e non

riceve un'istruzione formale (studio 3, N=87); e i bambini francesi di 6 anni che frequentano il primo anno di scuola primaria, ai quali non è mai stato insegnato cosa fosse un diagramma di dispersione (studio 4, N=27). Nel complesso, questi risultati suggeriscono che i precursori cognitivi della percezione dei grafici, sebbene variabili, siano disponibili indipendentemente dall'istruzione, dall'età e dalla cultura.

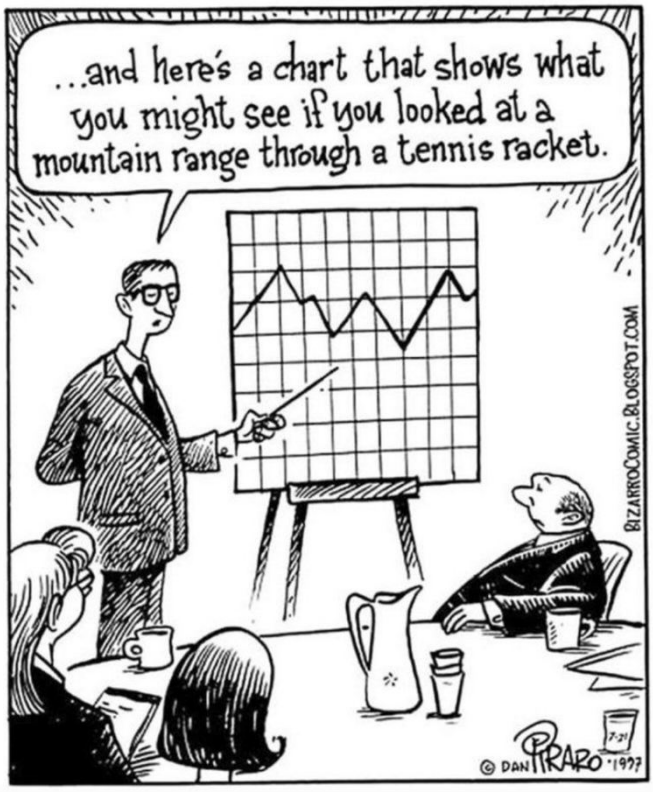
Con un compito di aggiustamento della linea (studio 5, N=10: "aggiusta la linea per farla passare attraverso i punti") ho caratterizzato ulteriormente la regressione mentale umana, scoprendo che gli esseri umani non minimizzano la distanza verticale dei punti dal fit (come sarebbe previsto da un modello OLS classico), ma piuttosto minimizzano la distanza ortogonale (come fa la regressione di Deming). Ho replicato questo bias di Deming con un compito di estrapolazione (studio 6, N=10: "posizionare un punto come migliore continuazione del grafico") e ne ho esplorato le implicazioni concrete.

Ho poi sondato i limiti della regressione mentale indagando se fosse robusta alla presenza di outlier (studio 7, N=30), come sarebbe previsto dalla letteratura se i grafici fossero trattati come un insieme. Ho scoperto che gli esseri umani includono spontaneamente gli outlier nei loro giudizi intuitivi sulla tendenza e negli aggiustamenti delle linee e sono in grado di escludere quelli con un punteggio z elevato, solo se viene loro chiesto esplicitamente di prestarvi attenzione.

Infine, ho indagato se gli esseri umani fossero in grado di andare oltre la linearità, chiedendo loro di estrapolare da diagrammi di dispersione (con rumore) generati da funzioni non lineari (studio 8, N=10): ho scoperto che sono in grado di farlo, con prestazioni dipendenti dal livello di evidenza di informazione nel grafico, tranne che per le quadratiche, che venivano sottostimate. Ho poi preso in considerazione un altro esempio di funzioni che notoriamente sono percepite in modo errato: le esponenziali (studio 9, N=625). I partecipanti le hanno

sottostimate, ma solo se nascoste nel rumore, suggerendo che la sottostima umana nell'estrapolazione di funzioni complesse non deriva da una mancanza di comprensione della funzione stessa, ma piuttosto dall'incapacità di estrarla dal rumore. Ho anche riscontrato che tale sottostima è correlata alle conoscenze matematiche e, sulla base di una simulazione bayesiana, che è probabile sia dovutoa a un prior contro le funzioni non lineari.

INTRODUCTION



Dan Piraro

WHAT IS A GRAPH?

Consider the following sentences: “stock prices are dropping today”; “the number of new Covid-19 cases is exponentially growing”; “life expectancy correlates with wealth”. Although their content is different, they share two fundamental aspects: first, they express a relation between (two) variables; second, they can be easily represented through graphs, which are, by definition, *pictures that show how (two) sets of information or variables are related* (Cambridge Dictionary of English). It seems even reasonable to assume that many readers, when facing such type of sentences, would probably mentally represent them as graphs (or at least think of graphs). Graphical representations come in all shapes and sizes: they can be colorful histograms illustrating the distribution of an enterprise budget among its departments or simple scatterplots expressing the increase in new deaths during a pandemic. Histograms and scatterplots are just two examples of the wide variety of graphs that are used today. The key aspect is that the above definition of graph still holds for all of them: they are revealing a relation between variables (in the examples above: the amount of money and the enterprise department; or the number of deaths and time). Interestingly, the relation between those variables is not necessarily made explicit, but it rather emerges from the information being displayed. Crucially, each datapoint (i.e., each irreducible portion of the dataset, such as the money given to a single department or the number of deaths in a single day) needs at least one other datapoint to express a meaningful relation between the variables at play and the higher the number of information datapoints is, the richer the interpretation of the relation becomes. Based on these assumptions, graphs are particularly useful when the relation is not known beforehand: in other words, when the graph creator wants to make sense of the amount of information they have, possibly discovering trends that they could not discover without a plot. The historical excursus presented in the following

paragraphs would provide evidence for this claim: indeed, throughout their history, graphs have been mainly used to interpret large amounts of empirical observations whose relations could not easily emerge without their visual representation.

A SHORT HISTORY OF GRAPHS

From the origins to the 18th century

Graphs are a cultural product, meaning that they are a human invention with defined rules and syntax. In this respect, they are very similar to written words and numbers, probably the two most famous cultural inventions: they all are symbolic representations based on shared conventions; they take advantage of the speed and capacity of the human visual channel to allow for the fast transmission of complex information; they require considerable learning experience; they are commonly taught at school. However, unlike numbers and words, graphs have been invented much more recently and they became widespread only in the last two centuries (Spence, 2006). The precursors of the wide variety of graphical representations that exist today can be traced back to Réné Descartes, the famous French mathematician that invented and formalized the system of Cartesian coordinates. The legend tells that, while lying on the bed and looking at his rectangular roof, flies would move around the surface of the ceiling; Descartes thought about a way to precisely represent the position of the flies on that surface. To do so, he conceptualized the west and south ceiling's borders as two axes and he divided them in fixed units (starting from zero), thus organizing the entire space as a grid. By doing so, he could easily determine the location of each fly in terms of their position over such a grid. The position was thus subsumed by a pair of numbers: one indicating the location on the south "x" axis, the other indicating the location on the west "y" axis. Thanks to the invention of the Cartesian plane, Descartes made one of the most important breakthroughs

in the history of mathematics: he found a way to bridge the gap between geometry and algebra, therefore making math equations visually representable. It is then not surprising that graphs depicting mathematical and physical laws were the first to appear, although they did not meet an immediate success until very late. Edmund Halley, the English astronomer, was the first to use a line graph in 1686 to represent the change in atmospheric pressure with altitude (figure 1).

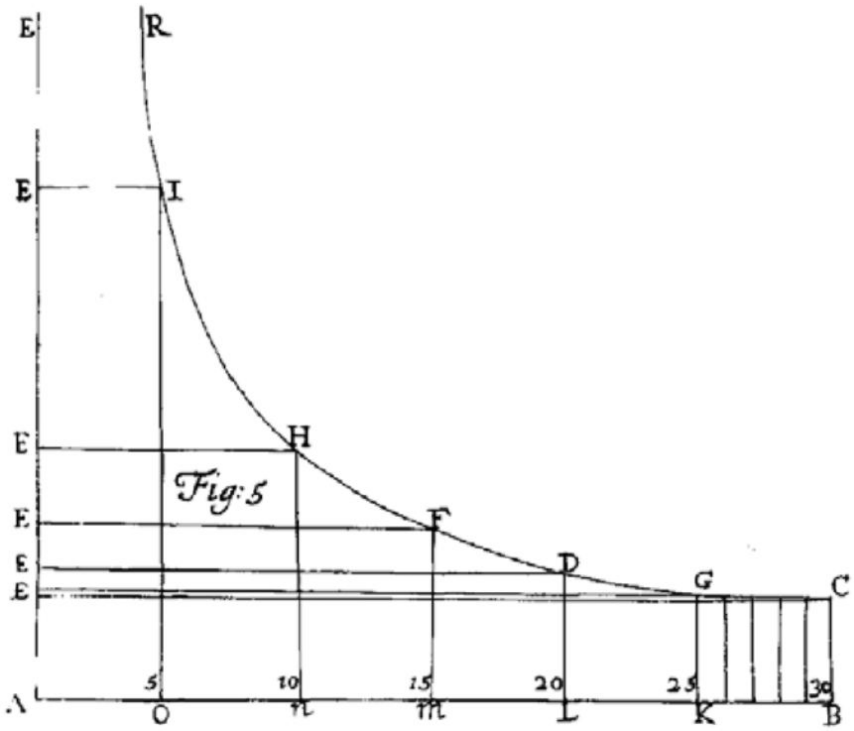


Figure 1. Bivariate plot of the theoretical relation between barometric pressure and altitude (*On the height of the mercury in the barometer at different elevations above the surface of the earth, and on the rising and falling of the mercury on the change of weather.* 1686, Edmund Halley).

No empirical data was represented in this plot though and, most importantly, the graph was used by Halley simply as a visualization of a (theoretical) physical argument rather than as a tool to reveal an undiscovered trend or relationship between two variables. In order to find graphs of this sort we have to wait until 1786: William Playfair, a Scottish engineer, decided

to graphically represent, by means of all types of graphs, the expenditures, revenues and debts of England. Interestingly, the text that accompanies his charts gives us a clear idea of his drawings' purposes. Take the example in figure 2, depicting the interest of national debt; Playfair explicitly declares his desire to show the *ruinous folly* of England to finance the war through debt, whose rocketing increase is made clear through the graph. In other words, Playfair was the first to use plots and charts in the "statistical" sense (interestingly, the term statistics was used for the first time in the contemporary sense only one year later, in 1787): to reveal a trend that summarizes a set of data, possibly in order to convey a message to the reader.

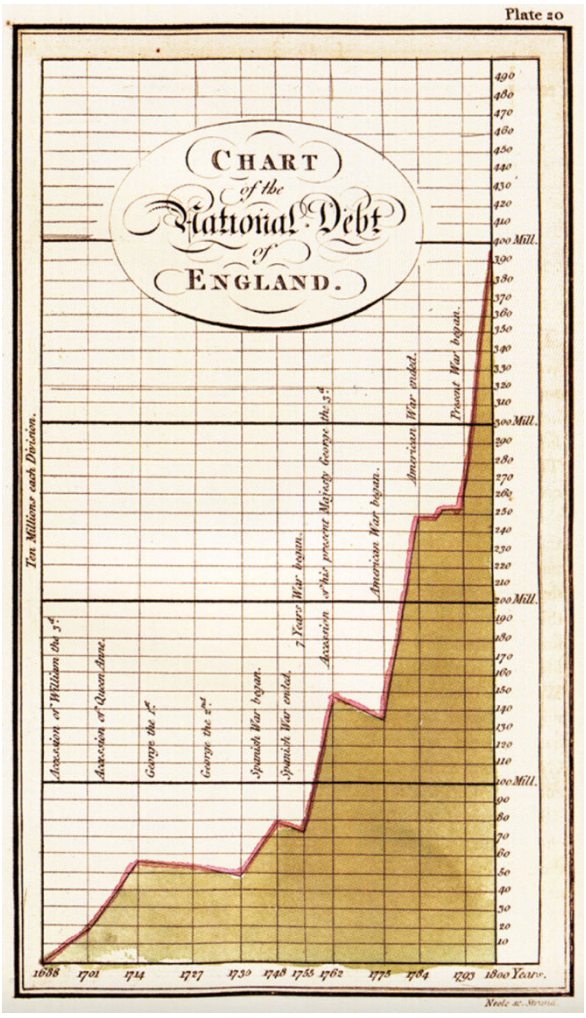


Figure 2. Evolution of the interest of the national debt (*The commercial and political atlas;* 1786, William Playfair).

The 19th century

During the 19th century, graphical representations with a statistical purpose slowly became more common: they were used to convey a message that would have not been evident from simple text or mathematical notation alone. I am going to consider a few remarkable examples in the following paragraphs.

John Snow, in 1854, realized a spatialized version of a histogram, thanks to which he discovered the origin of a cholera epidemics (figure 3). In fact, many people were dying of

cholera in the Soho district of London and authorities had a hard time finding the cause of such a major outbreak. Snow decided to more closely look at the data by taking a map of the district and drawing a small line on top of each building where a death had occurred. Therefore, higher stocks of lines indicated that those households had suffered from a higher number of deaths. The genial aspect of Snow's graph was to go beyond a simple histogram: in fact, it did not only provide the magnitude of the number of deaths per household but it also visually conveyed the location of the most affected buildings in the district. These were all around a water pump, which was thus immediately identified as the source of contaminated water. The pump was eliminated and the cholera epidemics came to an end.

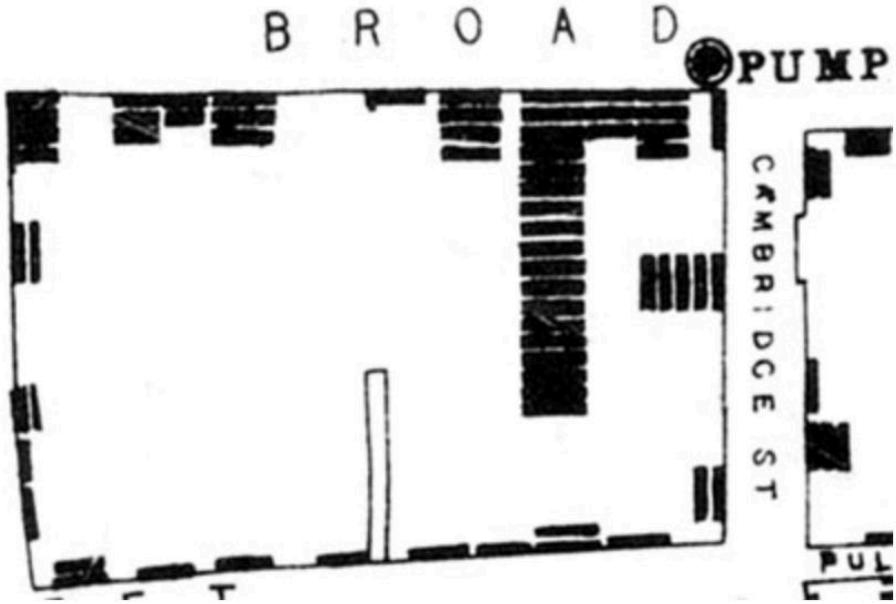


Figure 3. Households touched by Cholera deaths (*On the mode of communication of Cholera;* 1854, John Snow).

Florence Nightingale, in 1858, proposed her famous "rose" graph, by which she visually demonstrated that most deaths in the army were due to preventable causes (figure 4). This is a great example of chart in which a certain amount of data is consistently magnified in order to more strongly convey the underlying message. Indeed, rose graphs are simply histograms

presented on a polar grid: this means that each unit increase in the occurrence of a certain factor (for example, the number of people who died in November 1854) results in a non-linear increase of the area occupied in the grid, making it more salient for the reader. It is worth noting that Nightingale accompanied her graphs with a short text explaining how to read and interpret them.

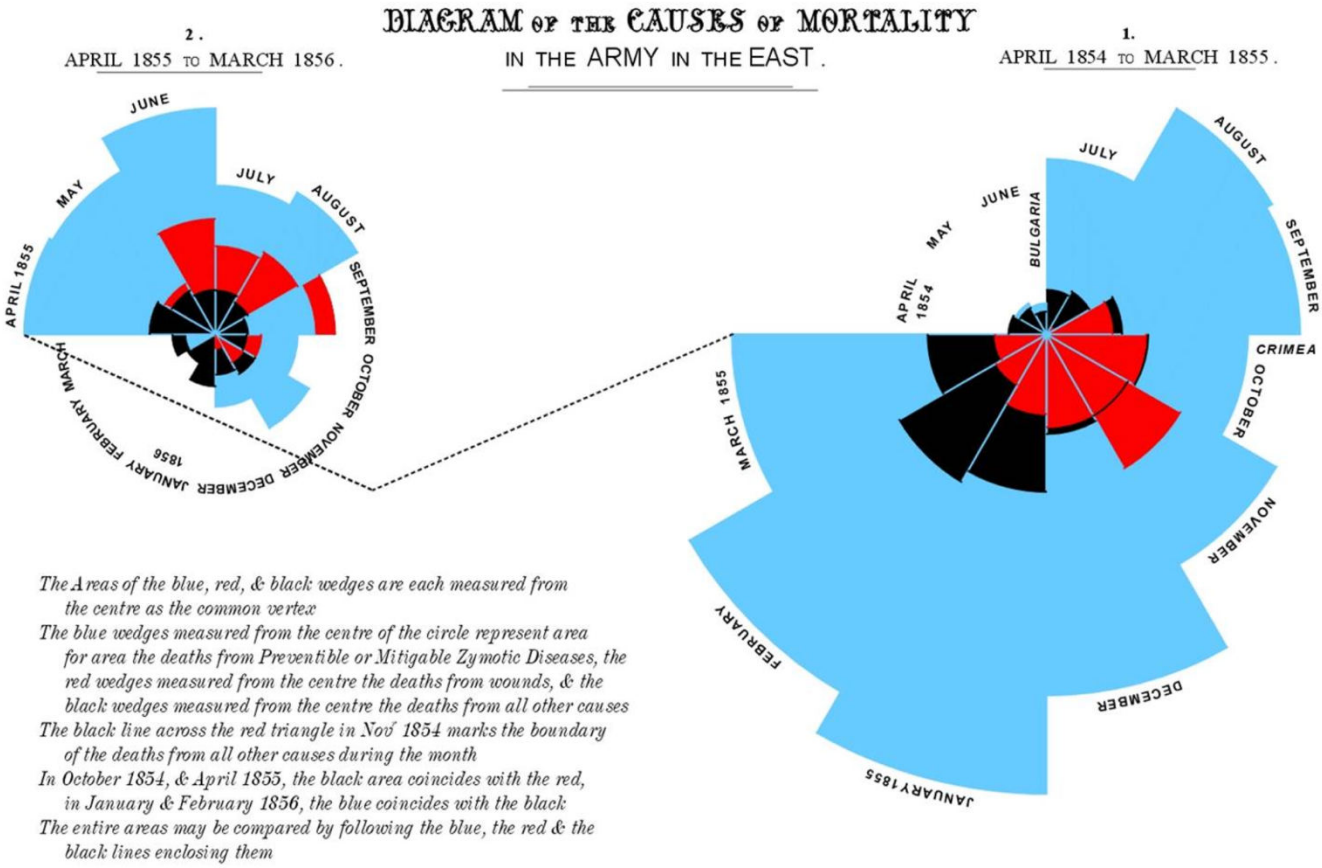


Figure 4. Causes of mortality in the army in the east (*Notes on matters affecting the health, efficiency and hospital administration of the British Army;* 1856, Florence Nightingale).

Jacques Bertillon, in 1896, further developed and refined the idea of a spatialized histogram and proposed his chart of the distribution of foreigner habitants of Paris by district (figure 5). In this map, each column represents the number of foreigners in that specific area. The graph is particularly efficient since it allows to immediately grasp where the majority of foreigners lives, without the need to go through each district one by one.

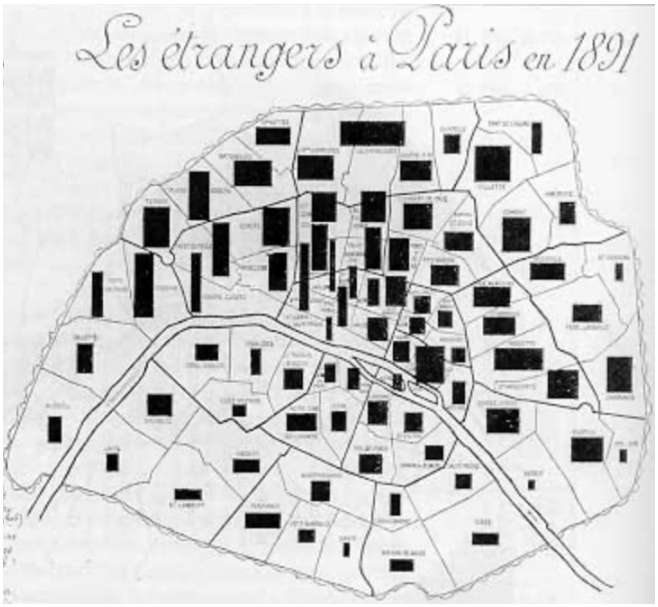


Figure 5. Frequency of foreigners in Paris in 1891 (*Cours élementaire de statistique administrative;* 1896, Jacques Bertillon).

Despite the fast development of new graphical representations over the course of 19th century, most commentators and scientists of the time thought that graphs were nothing more than an embellishment, with no useful purpose and lacking a sufficient rigor (Gray Funkhouser, 1937). If one critic might still be addressed towards those first graphs is probably to be found in the rather poor flexibility of their "syntax". In fact, all examples of statistical graphs I have just introduced were either a representation of a trend over time (Playfair, Nightingale) or distributions of occurrences of events (Snow, Bertillon). In other words, they did not depict the relationships of two purely empirical variables. This is exactly the most common purpose of another very powerful, yet simple, graphical representation: the scatterplot.

The rise of the scatterplot in the 20th century

It has been argued (Friendly & Denis, 2005) that scatterplots are the most versatile type of graphs, since they allow to discover regularities and trends among two purely empirical factors. Then, why did we have to wait until the 20th century to see scatterplots (Kurtz & Edgerton, 1939)? The most plausible reason is that people were primarily concerned about trends over time or distributions over locations, which are better displayed, respectively, through line graphs and histograms. It is thus not surprising that scatterplots' rise (except two

notable exceptions by Herschel and Galton, which do not entirely fall into the current definition of scatterplot) coincided with the rising success of the experimental sciences at the beginning of the 20th century. The first actual occurrence of the term scatterplot dates back to 1906 (Jenkinson, 1906): John Jenkinson described the relation between the symmetry of the egg and the symmetry of the embryo in the frog by plotting several observations along two axes (figure 6). All the typical elements of a scatterplot are there: observations are realized over multiple units (in this specific case, multiple eggs); the points are not connected one by one, but the general trend is left to the reader to discover (in this case, a negative linear relationship); each unit comprises two measurements, which refer to two different variables (namely: the symmetry of the egg and the symmetry of the embryo).

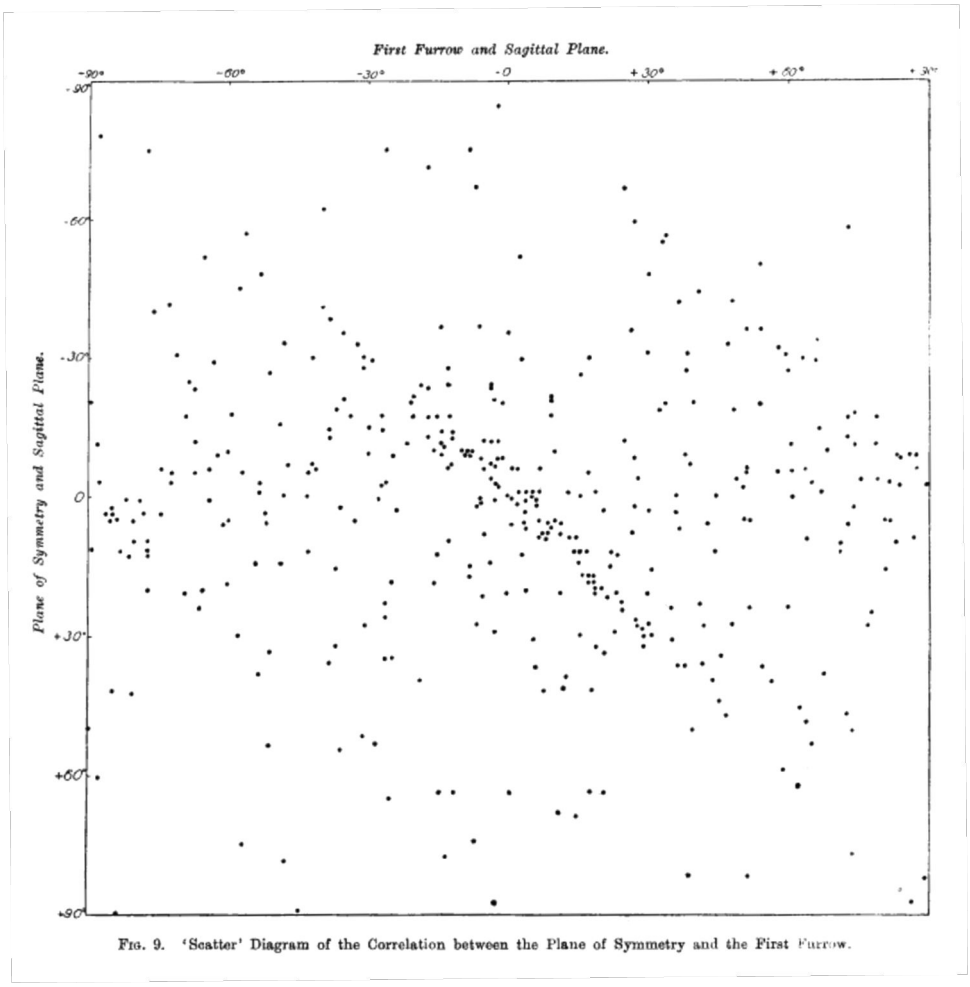


FIG. 9. 'Scatter' Diagram of the Correlation between the Plane of Symmetry and the First Furrow.

Figure 6. Scatter Diagram of the correlation between the plane of symmetry and the first furrow (*On the relation between the symmetry of the egg and the symmetry of the embryo in the frog (Rana Temporaria);* 1906, John Jenkinson).

From the beginning of the 20th century on, the use of scatterplots exploded. Since they represent multiple (often numerous) units of information, the invention and use of automatized computing and visualization methods contributed to the wide diffusion of scatterplots (and all other graphical representations). Today scatterplots have maintained their key aspects as described in the previous paragraph, but a few remarkable variations have been invented: three-dimensional scatterplots that integrate a third coordinate axis, thus allowing to concomitantly plot 3 variables of interest (figure 7); and dynamic two-dimensional scatterplots that change over the course of a video presentation, thus allowing to see the evolution of the variables' relationship over time (figure 8). These animated charts were invented by Hans Rosling a few years ago and quickly became widely appreciated and used.

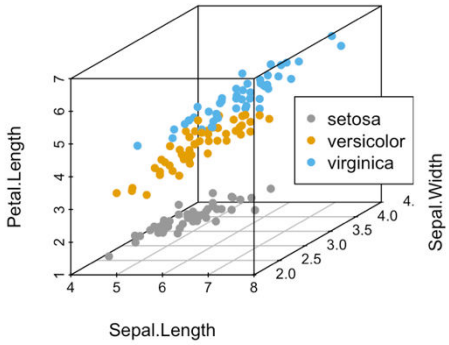


Figure 7. Example of 3D scatterplot. "Iris" dataset, *R package*.

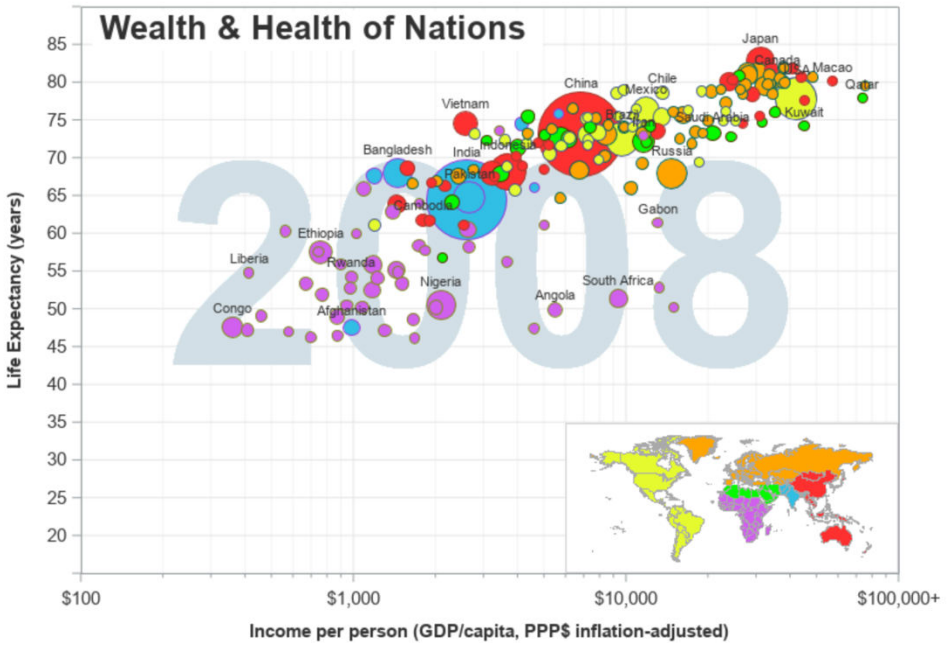


Figure 8. Snapshot of Hans Rosling bubble chart. *Gapminder.com*.

PRINCIPLES OF DATA VISUALIZATION

Graphs have quickly become the elective tool for data visualization: we see them on television, newspapers, scientific books or financial reports. Their ubiquity has been accompanied with many guidelines aimed at improving the way graphs communicate the intended message. These guidelines might be summarized as follows: first, different types of graphs are better suited for specific contents and messages; second, certain elements in graphs can be added, modified or eliminated in order to improve their readability.

Different graphs for different purposes

Every quantitative information could potentially be represented with any type of graphs. However, some pieces of information are better plotted through specific graphical representations. The following paragraphs provide a short review of the uses of the most common types of graphs: the line graph, the histogram, the boxplot and the scatterplot.

The line graph. It is often used to plot the evolution of an event or a quantity over time: in fact, a line allows to visually appreciate a trend, such as the evolution of the number of new Covid-19 cases or the change in stock prices. Line graphs are particularly efficient when attention should be focused on the rapid evolution of a value: if it drastically increases or decreases from a moment in time to the following one, the line connecting the two value positions will be proportionally longer. This aspect is particularly important since it might be intentionally used to exaggerate the message of the graph: by stretching the y axis scale, by truncating it (Correll et al., 2020), or by compressing the x axis scale, the resulting line becomes longer and steeper, thus conveying the message that the change is more drastic than it actually is. It has been shown (Beattie & Jones, 2002) that financial experts are indeed affected by the slope of a line graph when asked to make a judgment over the improving

performance of two competing companies: the actual numbers in the graphs are the same, but the stretched steeper line creates the illusion that a more impressive evolution has occurred (compared to the shallower one). If the factors are more than one but they are all measured along the same numerical metrics, multiple lines can be plotted on the same graph, particularly when the purpose of the graph creator is to make the comparison between the different conditions more salient.

The histogram. It is usually employed when the graph creator wants to emphasize the distribution of occurrences of a certain measure or event. In these cases, the x axis simply represents the different values (continuously or discretely ordered) that the factor can take and, on top of each value, a column is plotted with a height proportional to the number of occurrences of that value. In this context, histograms make immediately apparent if the distributions follow a known mathematical curve (such as the famous normal distribution) and where their measures of central tendency and dispersions are: in other words, histograms are also perfectly suitable to show how much the data are spread around a central value. But the x axis can also contain the factors of a nominal variable: in this case, the choice of their ordering is left to the graph creator, and can be based either on a contextual reason (i.e., putting closer the factors that are logically closer) or based on the number of their occurrences. It is worth noting that only in the case of a nominal variable (or of a discrete variable with a limited number of factors) the histogram represents the complete set of information available: if the variable is continuous, the graph creator is obliged to bin the data in intervals, thus losing the precise value of each information unit. It is therefore clear that different binning strategies might end up not revealing any reasonable distribution (if the range of the bins is too small) or hiding potentially interesting multiple distributions and outliers (if the range of the bins is too large; Knuth, 2019).

The pie chart. In a sense, a pie chart could be seen as a histogram plotted on a polar grid. In fact, each angle (and therefore each corresponding area of the circle) is proportionally larger for higher occurrences of a certain factor. It is, however, less flexible than the histogram since it is adapted only for nominable variables with a limited number of levels: many factors result in small slices, making it hard for the reader to identify their corresponding labels. Although pie charts might be useful when the graph creator wants to highlight some striking differences among the occurrences of certain events or measurements (e.g., by showing the higher number of men in positions of power as compared to the number of women), several studies (Kosara & Skau, 2016; Siirtola, 2019) concluded that pie charts are often not as efficient as other graph types, thus making them less and less used.

The boxplot. This is probably one of the most recently invented types of graphs (Spear, 1952; Tukey, 1977) and it is mainly used to conduct exploratory data analysis, since it provides, for each experimental factor, an indication of the lower and upper quartile (the horizontal borders of the box) and of the central tendency (the line inside the box) of the data distribution for that factor. As compared to the histogram, it is more limited: in fact, it does not allow to look at the actual distribution of the data but only at its summary values, making it hard, for example, to appreciate outliers (Godau et al., 2016; Pastore et al., 2017), unless they are added as separate points to the graph. The clear advantage of boxplots is in those situations where multiple distributions need to be compared: a separate histogram for each distribution would in fact take up much more space and it would not precisely point to the numerical values of the quartiles or the central tendency.

The scatterplot. This has been claimed to be the most used graph (Tufte, 2001) and probably the most flexible one. Indeed, it has a few unique features: it is suitable for showing the association between two variables, even when they are both independent from each

other (thus, it is not only adapt to show evolutions over time, as the line graph tends to do); it can easily represent the entire dataset available, since each information unit is indicated by a distinct point; it allows to easily make mathematical and statistical computations over it, since each point is a pair of two numerical values on a cartesian plane; lastly, it makes easy to extract, if present, the trend in the dataset and the noise around it, thus concomitantly providing indications of tendency and dispersion. Today, multiple variations of scatterplots exist: those including vertical lines to facilitate the location of the points on the x axis (Reimann et al., 2022); those with connected points (Haroz et al., 2016); those separating clusters of data by color (Wang et al., 2019); those varying the size of each datapoint depending on a third variable (Hong et al., 2021). The simplicity yet large versatility of scatterplots makes them highly suitable for a psychophysical investigation, as it will be explained at the end of the introduction.

Improving data visualizations

Independently from which type of graph one might want to use, several general suggestions have been made in order to improve graph readability and understanding. Most of these suggestions come from the seminal works of Edward Tufte (the most famous of which: Tufte, 2001). In his books, he provides a large number of guidelines to make better graphical representations of data: some of them have been confirmed to improved graph readability; some of them have not been confirmed by experimental evidence; some others (probably the majority) have never been put to test. Here I consider his most famous guidelines.

Increase the data density. Tufte argues for the need of condensing as much information as possible in one graphic. This suggestion translates into displaying the entire set of information available in a single graph. Against this norm, a given chart can be repeated

multiple times in order to convey the evolution of a factor over time or places. These graphs have been called “small multiples” and an example is provided in figure 9. Recent evidence seems to support the higher readability of this graph type when compared with more classic and “dense” visualizations (Yoghourdjian et al., 2018).

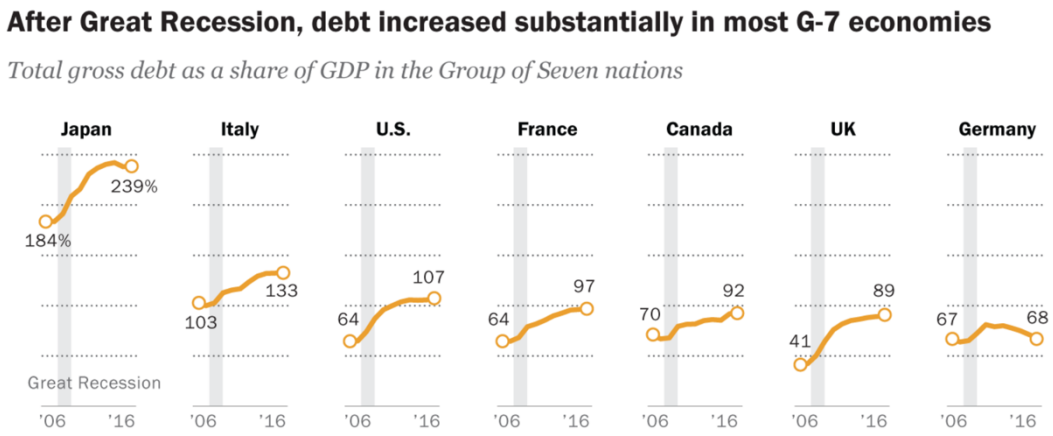


Figure 9. Example of “small-multiples” graph, depicting the increase in debt in several countries. 2017, *Pew Research Center*.

Maximize the data-ink ratio. This is probably the most famous of his guidelines: it suggests that the ratio between the quantity of data being displayed and the “ink” used to display it (i.e., the area that the visualization occupies) should be maximized. In Tufte’s words, this maximization is *the non-erasable core of a graphic*. Concretely, following this guideline means: avoiding the repetition of axes’ labels if those axes are presented multiple times; displaying the legend only once if it refers to multiple charts; if possible, labelling conditions inside the graph instead of adding a separate legend. The data-ink ratio principle has been both confirmed (e.g.: adding esthetic enhancements does not improve memory for the data values: Peña et al., 2020) and refuted (e.g., adding two axes to scatterplots increases graph understanding: Poulton, 1985), suggesting that it might largely depend on the specific graph

and content that is plotted (Franconeri et al., 2021). In fact, highly decorated graphs might push the reader to engage more into the exploration of the graph.

Minimize the lie-factor. Tufte discourages representing graphics in which the numerical metrics are not proportional or faithful to the actual numerical values of the dataset. In other words, if an effect size is small, it should be graphically represented as small. As I have already argued when discussing the line graph, manipulating the axes in order to increase the slope of lines is an example of "lie" that has been shown to affect people judgments and understanding. An example of graph with a large lie-factor is provided in figure 10.

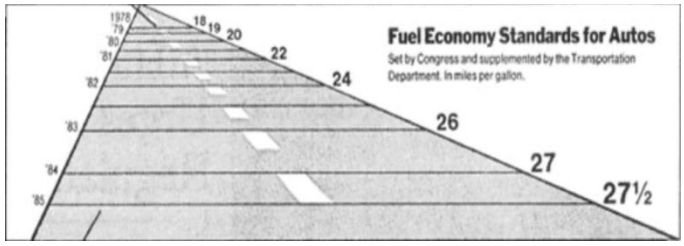


Figure 10. Example of graph with a large lie-factor. The increase in numerical magnitude is not proportional to the increase in length of the corresponding lines. 2001, Tufte.

The other guidelines are more subjective (e.g., the encouragement to make elegant graphs) and therefore, although appealing and seemingly reasonable, are difficult to be accurately tested. What seems important to point out about Tufte's work is that his suggestions pertain to the globality of the graph and do not concern the fundamental visual operations (such as detecting the difference in length of two lines or locating the relative position of two points) that are necessary to understand all types of graphs.

MASTERING GRAPHICACY: WHAT DOES IT TAKE TO UNDERSTAND A GRAPH?

Graphicacy is the human ability to read and understand a graph (Balchin & Coleman, 1966): the word clearly echoes literacy and numeracy, the two other human skills based on the explicit learning of cultural symbols (letters and numbers, respectively). Independently from

the graph type and the techniques to improve data visualization described in the previous paragraphs, mastering graphicacy requires three steps, which have been metaphorically defined as (1) reading *the* data, (2) reading *between* the data and (3) reading *beyond* the data (Curcio, 1987; Friel et al., 2001). The first step consists in extracting the data by visually inspecting the elements drawn in the graph; the second step requires inferring the mathematical relation between the variables; the third step refers to the ability to make a numerical forecast based on the graph's underlying trend. This three-steps classification is particularly useful to organize the heterogeneous corpus of studies about graph perception and comprehension. I will detail each step separately in the following paragraphs.

Data extraction: the visual processing of graphs

Before any complex understanding of the graph content, the reader has to correctly process the visual features of the graph. The first attempt to systematize them was made by the French cartographer Jacques Bertin in 1967 (Bertin, 1967). First, he defined the three elements used to signal the position of the information units on the graph: points, lines and areas, which are, respectively, the elements used in scatterplots, line graphs and histograms/pie charts. Second, and most importantly, he provided examples of the six features that those elements can take inside a graph, i.e., the way such elements can be distinguished from one another, often in order to introduce a third variable in the

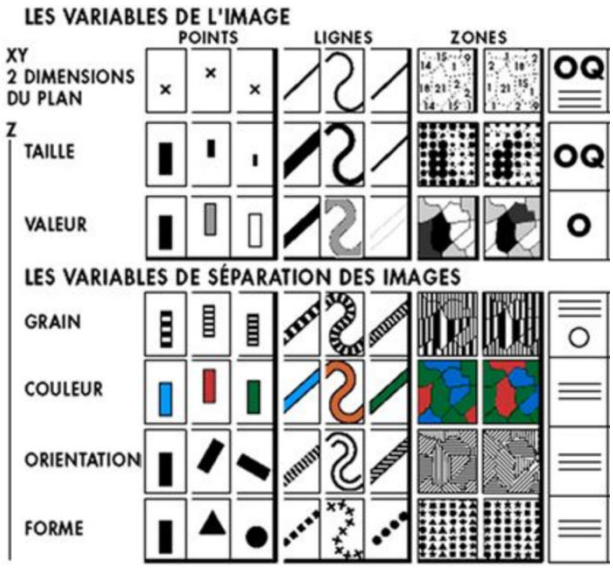


Figure 11. Combination of elements and features in graphical representations (Sémiologie graphique; 1967, Jacques Bertin).

representation. These features are: size, value, texture, color, orientation and shape. A summary of the combination of elements and features is provided in figure 11.

Bertin offered many suggestions about which combination is better suited for a given dataset: for example, he encouraged using size and position when representing quantitative variables and he saw value and texture as more adapt than color at communicating order. However, he did not provide an empirical investigation of the human abilities at perceiving these elementary visual elements. The first to do so were Cleveland and Mcgill (Cleveland et al., 1982; Cleveland & McGill, 1985, 1985): they considered Bertin's (and other) visual elements and asked human participants to judge the ratio of two values expressed through a given visual element. For example, how larger a circle is compared to another one, how wider an angle is compared to another one. Their findings (recently replicated with online testing: Heer & Bostock, 2010) allowed them to order visual elements on the basis of participants' accuracy at judging those ratios. Figure 12 shows this ordering, from visual elements harder to distinguish to those easier to tell apart.

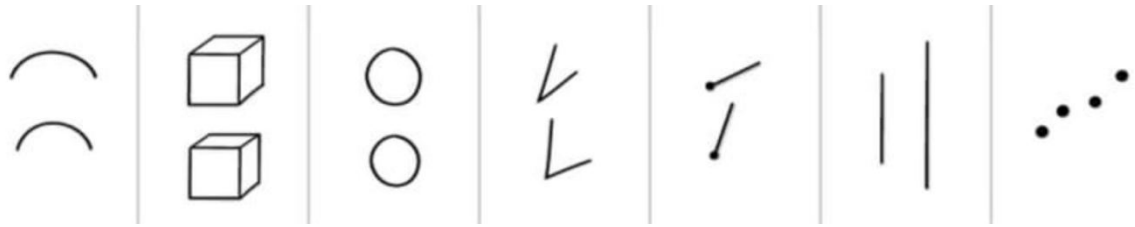


Figure 12. Visual features used in a graph, ordered from the hardest to detect to the easiest to detect, based on the literature reviewed in the text: curvature, volume, area, angle, orientation, length, position.

Color and brightness can also affect the ease of data extraction from a graph (Vanderplas et al., 2020). Indeed, most Gestalt principles of perceptual organization play a crucial role in data extraction (Kosslyn & Kosslyn, 2006): for example, data points that are closer in space are

more likely to be perceived as grouped (Ciccione & Dehaene, 2020), and vertical and horizontal lines are easier to discriminate than oblique ones (Appelle, 1972).

Other studies have also investigated the visual illusions that certain representations might induce in extracting data from a graph, thus resulting in possible confusions and misunderstandings for the reader. Franconeri and colleagues recently provided a comprehensive review of these illusions (Franconeri et al., 2021), together with a critical analysis of the existing research-based guidelines in data visualization. It is worth noting that, despite this and other recent efforts, data visualization guidelines are still often taken as granted (as it is the case for most of Tufte's work), although not being rooted in scientific evidence; and even when they are backed up by empirical proofs, they come from very distant research communities (ergonomics, graphic design, cognitive psychology, user experience…), thus making it hard to easily navigate through the different pieces of evidence.

Data understanding: inferring the mathematical function

The second step, inference or "reading between the data", refers to the ability to infer the nature of the mathematical function that relates the data in the graph. Most studies in this field have focused on a specific paradigm of "function learning", the ability to learn the functional mapping between a set of input values and a set of output values. However, these studies did not primarily focus on graphs, but mostly on the sequential presentation of pairs of input/output data. In such paradigms, there is a learning phase during which participants slowly infer (with or without corrective feedback) the nature of the relation between paired stimuli, the data being presented for instance as horizontal bars (one for the input value, and another for the output value of the function given that input). When participants were asked to generalize from new input stimuli, they often gave evidence of having correctly learned and

applied the function that linked the input and output values (Bott & Heit, 2004; Carroll, 1963; DeLosh et al., 1997), but failed at doing so for more complex non-linear functions (Kalish, 2013).

While a growing body of literature focused on the cognitive mechanisms of function learning, few studies investigated the capacity to infer functions from classic graphical stimuli (such as line plots or scatterplots drawn on a Cartesian plane). A notable exception is the work of Schulz and colleagues (2017), who found that human adults could successfully interpolate and extrapolate sophisticated functions (such as a sinusoidal function with an increasing amplitude) from a plot of their graph. Schulz and colleagues suggested that this ability reflected the existence of a compositional grammar of functions, which allows human adults to understand complex functions as the composition of a small repertoire of simpler ones. Other authors have shown that, when exposed to a noisy scatterplot, participants tend to interpolate functions with a lower polynomial degree than the real one (Little & Shiffrin, 2009) and their subjective ability to interpolate is negatively affected by increasing levels of noise (Schulz et al., 2015). Other studies showed that participants could, in a slow and reflexive manner, fit a linear function to a given scatterplot after receiving formal training on statistical regressions (Mosteller et al., 1981). Human adults may even adjust quadratic and trigonometric functions to an underlying scatterplot, once they are precisely informed about the nature of each curve (Correll & Heer, 2017).

Closer to the present research, the perceived correlation in a scatterplot has been investigated as well. When asked to judge the degree of association between two variables in a scatterplot, participants tend to underestimate it (Strahan & Hansen, 1978) and this underestimation is higher for regression slopes further from a 45° orientation (Bobko & Karren, 1979). Participants' ability to compare scatterplots with different correlation coefficients follows

Weber's law (Harrison et al., 2014; Rensink & Baldridge, 2010) and they perceive datasets as more highly correlated if the axes are scaled in order to make the underlying function's slope appear steeper (Beattie & Jones, 2002; Cleveland et al., 1982). Interestingly, it has been observed that correlation judgments might indeed be based on a small number of visual features in the graph, rather than on the correlation itself (Yang et al., 2019). Also, participants' performance at extracting the regression's slope is affected by the localization of the scatterplot: positive slopes are more easily detected if the scatterplots are presented on the right of the visualization area, and the opposite is true for negative slopes (Parrott et al., 2014), a bias consistent with the Spatial-Numerical Association of Response Codes (SNARC) effect (Dehaene et al., 1993).

Data forecasting: predicting the future

The third step, numerical forecasting or "reading beyond the data", is the capacity to make numerical predictions and forecasts based on the data presented, i.e., to extrapolate beyond the existing data range. Given its many concrete applications, it is not surprising that the majority of studies in this area have been conducted by researchers in finance, economics and politics. Studies on this topic are too diverse in methods and purposes to be reviewed here. Most pertinent to the present work is a general tendency to underestimate future data points if they are based upon a non-linear, positively accelerated function (Lawrence et al., 2006; W. Wagenaar & Sagaria, 1975). The authors speculated that, in a real-world context, most trends do not keep growing at the same steady rate and, knowing this, participants might be conservative in their predictions. Better extrapolation performance has been obtained through experimental designs involving numerical values instead of scatterplots (Lawrence & Makridakis, 1989). The noise of the dataset was also found to affect extrapolation

performance (Harvey et al., 1997), and participants seemed to add noise to their extrapolations, as if they were attempting to make their predictions more consistent with a noisy forecast (Bolger & Harvey, 1993).

THE THESIS' GOAL: INVESTIGATING THE COGNITIVE AND NEURAL BASES OF GRAPH PERCEPTION THROUGH A PSYCHOPHYSICAL APPROACH

Overall, the heterogeneous set of studies presented so far, which vary considerably in both methods and research questions, do not provide a thorough psychophysical investigation of human abilities to detect trends in noisy graphical representations, in the absence of any reference to the graph's meaning, context, or underlying function. The vast majority of these studies used very few stimuli, did not systematically explore the multiple parameters of the graphs, provided considerable training and/or background information about the underlying functions, left considerable time for subjects to inspect the data and strategize about the task, and none of them measured the time needed to extract statistical information from a graph, for instance to perform a simple trend judgment. Here, our goal was to begin to fill these gaps. We ran a series of behavioral experiments with the purpose of studying human accuracy, response times and biases in the visual extraction and inference of statistical information from noisy graphs. For simplicity, we focused on one of the simplest graphical representations, the scatterplot. This choice was made for different reasons. First, in order to minimize the effects of learning and memory, which, as pointed out by several authors (Little & Shiffrin, 2009; Lucas et al., 2015; Villagra et al., 2018), were certainly at play in the aforementioned classic studies of function learning, mostly based on long training sessions with serial presentations of input/output pairs. Indeed, one of the most remarkable properties of the scatterplot is its ability to simultaneously represent, in a single graph, a very large data set, thus allowing

participants to perform a statistical judgment on each single trial. Second, the scatterplot is perfectly suited for a fine manipulation of its appearance, for which – unlike natural scenes – we can know all statistical aspects: as I will describe in the methods' sections of the following chapters, it is easy to generate a desired number of points with gaussian noise around the equation of a line with a certain slope, thus investigating separately the influence of each of these factors (i.e., number of points, noise and slope) on participants' performance (on the usefulness of graphs as ideal perceptual stimuli, see: Szafir et al., 2016). Third, scatterplots can easily reveal non-linear trends, therefore not limiting our investigation to linear functions. They are also the elective graphical tool (Orr et al., 1991) for the visual exploration of the presence of outlier observations. Fourth, they are represented along two axes, which can in turn be experimentally manipulated, for example by varying their scale from linear to logarithmic. Fifth, they are generally used for different purposes, such as detecting a general trend, interpreting the rate of growth or extrapolating future observations. This means that we could ask participants to perform a variety of tasks on the same kind of stimuli, which is crucial for the sake of generalization of our findings. Sixth, the simplicity and versatility of scatterplots made them good stimuli candidates for a neuroimaging investigation of the neural bases of the perception of graphs. Last but not least, as briefly stated in the introduction, scatterplots are one of the most used graphical representation: it thus seemed reasonable to start our psychophysical investigation of graphs from a well-known and used existing chart.

Research questions and organization of the chapters

I applied a psychophysical approach to all three fundamental steps of graphicacy described in the previous section: data extraction, inference of the mathematical function and data forecasting. More precisely, I organized my investigation of the cognitive bases of graph

perception guided by several research questions. I here introduce those questions and the studies performed in order to answer them; each number in the list refers to the corresponding chapter in the thesis.

1) What are the precursors of graph perception? In order to answer this question, I designed a novel psychophysical task of trend judgment, which is one of the simplest statistical judgments that can be performed over a noisy graphical representation. By finely manipulating several features of the scatterplot, I was able to characterize the psychophysical characteristics of such intuitive data extraction, finding that human performance is tightly predicted by the t-value, i.e., the statistical significance of the correlation expressed in the graph (study 1). I replicated these findings with a large-scale online study and I operationalized a graphicacy index based on the performance on this simple trend judgment task, finding that it correlates with self-evaluations of statistical and mathematical knowledge (study 2). I further investigated whether these intuitive graphics abilities are available early in the development and independently from formal schooling: to do so, I asked unschooled Himba people (study 3) and 6-year-old children (study 4) to perform the same trend judgment task, obtaining similar results.
2) How accurately can humans perform a mental regression over a noisy graph? In order to answer this question, I asked participants to adjust a line over a flashed scatterplot (study 5), finding that they do not minimize the vertical distance of the points to the fit (as expected by classic OLS regression) but rather the orthogonal distance (as in Deming regression). I replicated this Deming bias with an extrapolation task in which participants were asked to forecast data outside the portion of graph presented to them (study 6), thus showing that the bias we discovered is independent from task

modality and probably due to the way our visual system extracts summary information from noise.

3) Is human mental regression resistant to outliers? In order to investigate the limits of human intuitive statistical abilities over noisy graphs, I asked participants to perform the same trend judgment and line adjustment tasks over scatterplots including outliers (study 7), either without informing them of their presence or by asking them to exclude outliers from their judgments. I found that humans are strongly attracted by outliers in their judgment, in partial contradiction with the findings from the ensemble perception literature, which suggest that, when averaging a perceptual dimension over multiple stimuli, humans automatically exclude distractors and outliers. I also found that, when asking participants to explicitly detect outliers before making a mental regression, their attraction towards deviant observations is reduced but not excluded.
4) Can humans correctly forecast data from graphs underlying non-linear functions, including the famous exponential growths? I asked participants to perform an extrapolation task from noisy scatterplots generated from non-linear trends, including piece-wise linear, sinusoid and quadratic functions (study 8); I found that their performance, generally accurate, depend on the evidence provided in the graph and is quite low for quadratics. In a large-scale online study (study 9) I found that, for exponentials, participants tend to largely underestimate their forecasts, unless the data are presented on a log scale (thus resulting in a visually linear function), in which case other biases appeared, including anchoring effects to displayed numerical values.
5) What are the neural bases of graph perception? With 3T fMRI I investigated the neural bases of the trend judgment task (study 10; the study is currently ongoing).

CHAPTER 1

THE PRECURSORS OF GRAPH PERCEPTION: HUMAN ACCURACY IN A TREND JUDGMENT TASK

"He's right, when you look at it that way, it's not so bad!"

Mark Anderson

In the studies described in this chapter we asked participants to make a simple decision about a noisy scatterplot ("does it go up or down?"), bringing classical methods of psychophysics and mental chronometry to the study of human graph perception. Specifically, our first empirical questions were the following: can human adults perform a fast judgment of the *trend* underlying a noisy scatterplot, i.e., understand whether the data is increasing or decreasing? Which factors affect participants' accuracy and response times in such a task? Do participants perform a mental computation akin to the computation that a statistician would perform to detect if a significant positive or negative trend is present? Are these abilities also found in 6-year-old children and unschooled adults? Do they correlate with mathematical and statistical knowledge?

We tested 10 adult participants in our laboratory (study 1), 3943 adult participants online, who performed the task on their computers or smartphones/tablets (study 2), 87 adult and teenager Himba participants in Namibia, who performed the task on tablets (study 3) and 27 6-year-old children of the Académie de Versailles, who also performed the task on tablets (study 4).

STUDY 1: TREND JUDGMENT ON NOISY SCATTERPLOTS

In our first experiment, we tested in our laboratory if human adults can perform a fast, intuitive judgment of whether a scatterplot of data shows an increasing or decreasing trend. We generated the graphs according to the hypotheses of classical linear regression ("ordinary least squares"): the values on the ordinate (called y_i) were a linear function of the values on the abscissa (called x_i) plus independent Gaussian noise ($y_i = \alpha x_i + \varepsilon_i$, where ε_i are random numbers independently drawn from a normal distribution centered on zero and with standard deviation σ). We varied orthogonally three parameters of the graphs: the slope of the linear trend (α); the number of points (n); and the standard deviation of the noise (σ).

This experimental design was chosen because it allowed to compare the performance of human participants with a normative model of decision making in this task. As further detailed in appendix A, classical statistical theory predicts that the optimal decision should be determined by a simple *t* test, similar to the one that statisticians use to test for the presence of a positive or negative linear trend. The theory further predicts that responses should be a sigmoidal function of the t-value, and that the response time should be a decreasing, convex upward function of the absolute deviation of the *t-value* from zero. The sole dependence of decisions on the t-value also implied that decision difficulty should vary significantly with all three of the manipulated graph parameters (n, σ and α), because all of them influence the statistical t-value: it varies positively and linearly with the slope α, positively with the number of points (as the square root of n-2), and inversely with the noise level σ. Finally, the theory predicts that the effects of these variables should be jointly subsumed by an effect of the t-value on behavior.

METHODS

Participants

10 participants were recruited (age: 23.9 $\pm$ 1.5, 4 females, 6 males). All participants had normal or corrected to normal vision, no medical history of epilepsy, were right-handed, and did not take psychoactive drugs. They all signed an informed consent and were paid 5 euros for their participation. The experimental session lasted approximately 30 minutes. The experimental procedure was approved by the local ethical committee.

Experimental design and procedure

Each participant was presented with 672 scatterplots and, for each of them, was asked to decide, as fast as possible, if the dataset was increasing or decreasing. Each scatterplot was the graphical representation of a dataset that was generated randomly, independently for each participant, using a linear equation plus noise (see below). The design was a full factorial design where we varied the number of points (*n* = 6, 18, 38 or 66), the standard deviation of the noise (σ = 0.05, 0.1, 0.15 or 0.2), and the slope of the underlying linear trend (α = -0.1875, -0.125, -0.0625, 0, +0.0625, +0.125 or +0.1875), for a total of 4x4x7=112 combinations. The values of *n* were chosen so that $\sqrt{n-2}$, which is the value that enters in the t-test for the significance of a regression, was linearly distributed ($\sqrt{n-2}$ = 2, 4, 6 or 8). The other factors were selected after piloting in order to avoid excessive difficulty as well as ceiling effects; specifically, we chose relatively high levels of noise and relatively small levels of α in order to make the task non-trivial. The 112 combinations of parameters were randomly presented to each participants in each of the 6 experimental blocks, for a total of 672 trials per participant. The participants were invited to sit on a fixed chair with their head at a distance of 50 cm from the screen. As illustrated in figure 13A, a fixation cross first appeared for 1000 ms, immediately followed by the flashing of a scatterplot for 100 ms, and then by a fixation circle

of 1 cm diameter at the center of the screen; the participants were informed that the circle marked the onset of the response window. Participants were asked to respond as fast and as accurately as possible. Half of them responded by pressing with their right index on a key (signaled with a ⇑ sticker) on the right side of the keyboard if they thought that the trend in the scatterplot was increasing; and, conversely, they pressed with their left index on a key (signaled with a ⇓ sticker) on the left side of the keyboard if they thought that the trend in the scatterplot was decreasing. The opposite response configuration was presented to the other half of the participants. Once they gave their answer, a fixation cross appeared again for 1000 ms, inviting the participants to concentrate on the center of the screen before a new stimulus appeared. As mentioned, the task was divided into 6 blocks of 112 trials; the duration of each block was ~4 minutes. After each block, the participants could take a short break and received feedback on the total number of correct responses they gave in that block. Before the beginning of the actual experiment, 25 practice trials were run under the supervision of the researcher, in order to control for the correct execution of the task.

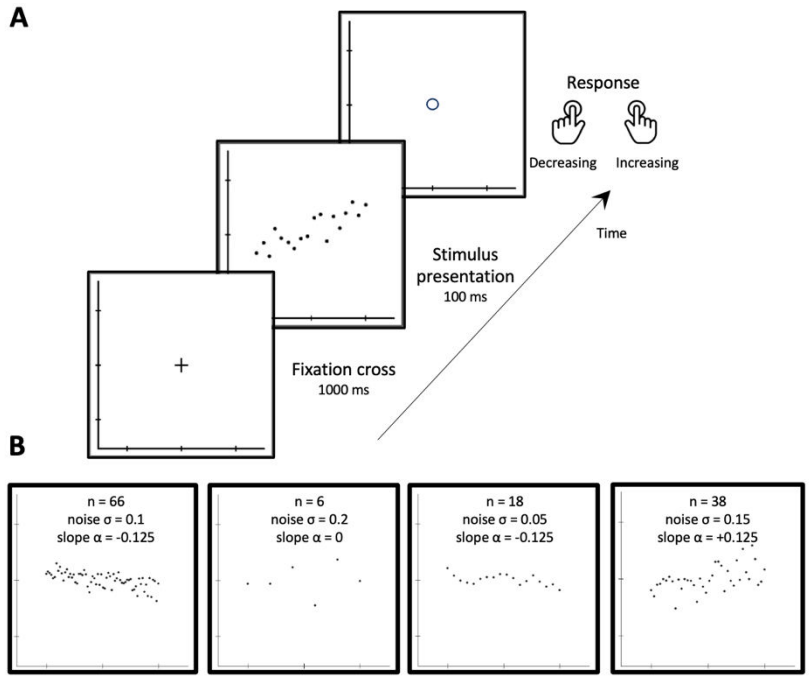


Figure 13. A, task: subjects were presented with a simple scatterplot, generated by a linear function plus noise, with a variable number of datapoints, and were asked to judge if the trend was ascending or descending. **B, examples of stimuli**.

Stimuli

Each scatterplot comprised two unlabeled lines denoting the x and y axes (which remained on screen for the duration of the experiment), each marked with three small ticks at the values 0, 0.5, and 1 (see figure 13; those numbers were arbitrary and were not shown to the participants). The dots' coordinates were calculated by a Python program as follows. First, the algorithm used the desired number of dots, *n*, to generate the x values (denoted x_i) such that they ranged from 0 to 1 and were equally spaced. Thus, for example, for all configurations having 6 points, the x values were always [0, 0.2, 0.4, 0.6, 0.8, 1]. The y coordinates were then determined according to the following equation: $y_i = \alpha x_i + \varepsilon_i$, where α is the prescribed slope and the ε_i are random numbers independently drawn from a normal distribution centered on zero and with standard deviation σ. If, occasionally, a point took a particularly high or small y coordinate (y<-0.27 or y>1.27), which would have exceeded the boundaries of the y axis, the algorithm was reinitialized for that particular trial. The noise terms ε_I were generated independently for each trial and each participant. Due to this noise term, the actual linear regression line, as calculated from each dataset, could pass slightly above or below the center of the screen. To sidestep this issue, the coordinates were adjusted vertically by subtraction of the mean to ensure that the underlying regression actually passed through the exact center of the screen (i.e., through the point P having coordinates x = 0.5 and y = 0.5). The x and y coordinates were then rescaled to the coordinates of the computer screen used in the experiment, and each data point was represented by a 2-mm white dot centered at the appropriate location (~0.23° of visual angle given the distance of 50 cm from the screen). Figure 13B shows four examples of scatterplots derived from datasets with different parameter values. As we can see, higher values of α correspond to higher inclinations of the graph, and higher values of σ result in noisier scatterplots.

RESULTS

Performance depends on the t-value of the scatterplot

We first looked at the proportion of "increasing" responses as a function of the prescribed slope α and either the prescribed noise σ or the number of points (figure 14). A repeated-measures ANOVA on the fraction of "increasing" responses confirmed a main effect of the prescribed slope ($F[6, 54] = 466.51$, partial $\eta^2 = .98$, $p < .0001$), and its interaction with noise ($F[18,162] = 12.62$, partial $\eta^2 = .58$, $p < .0001$) and with the number of points ($F[18,162] = 4.97$, partial $\eta^2 = .36$, $p < .0001$). In other words, the smaller the slope, the higher the influence of noise and number of points on the trend detection task.

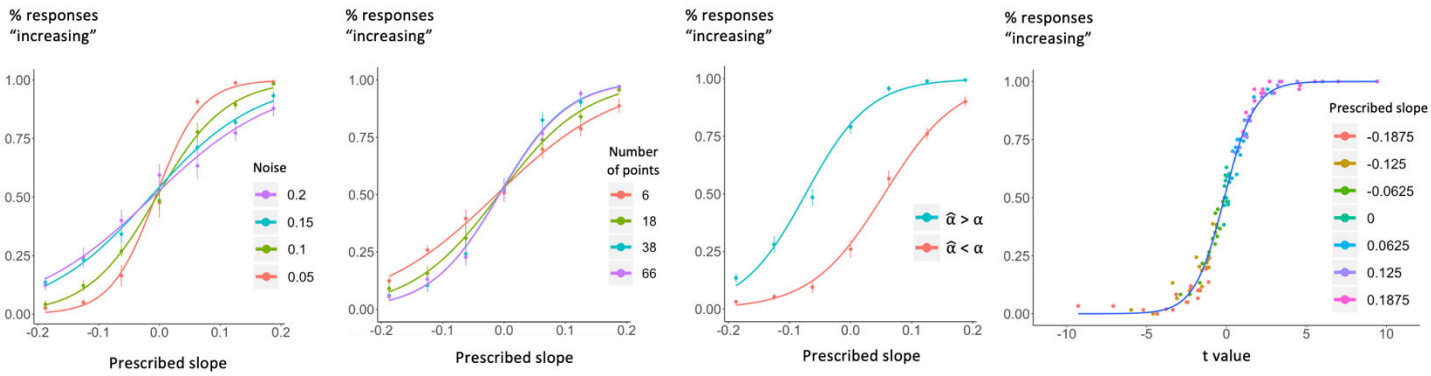


Figure 14. Accuracy of human subjects in judging whether a noisy scatterplot is increasing or decreasing. The percentage of "increasing" responses is affected by the prescribed slope of the graph (α), the noise in the graph (σ) and the number of points (*n*). The third plot shows that subjects' responses depended not only on the prescribed slope α, but on the actual slope $\hat{\alpha}$ after the addition of noise. The fourth plot shows that all the effects of noise, number of points and slope could be subsumed by an influence of the t-value associated with the Pearson coefficient of correlation, as if participants performed a mental linear regression. In this graph, each dot represents the mean, across trials and subjects, of all data for one of the 112 experimental conditions determined by each combination of α, σ and *n*.

We also conducted an analysis of each participant's sensitivity to prescribed slope, as a function of the noise σ and number of points *n*. We fitted a logistic regression to the percentage of "increasing" responses as a function of the prescribed slope, separately for each participant, noise level, and number of points, and used the slope of that logistic function

as an indicator of sensitivity. In a few cases (11 out of 160 combinations), the logistic regression was not a meaningful indicator of performance since the data were better modeled through a step function (meaning that perfect answers were always given, thus resulting in a slope approaching to infinity); these cases were substituted with the maximum observed value. We submitted the resulting sensitivity values to a repeated-measures omnibus ANOVA and found a significant main effect of noise ($F[1.61, 14.45] = 82.72$, partial $\eta^2 = .90$, $p < .0001$) and a significant main effect of the number of points ($F[2.29, 20.60] = 19.33$, partial $\eta^2 = .68$, $p < .0001$). The direction and monotonicity of these effects indicated that, as expected, participants' decisions became increasingly sensitive as the datasets had a smaller noise level and a higher number of data points.

Because of the randomness in the stimulus generation process, on any given trial the actual slope of the linear regression line $\hat{\alpha}$ could differ from its prescribed value α. We wondered if participants were sensitive enough to detect such variations in the actual slope of the graph. To this end, we looked at the fraction of "increasing" responses as a function of the prescribed slope α, while separating the trials based on whether the actual slope $\hat{\alpha}$ was above or below α (figure 14, third plot from the left). A repeated measures omnibus ANOVA revealed a significant effect of the prescribed slope ($F[2.83, 25.51] = 549.13$, partial $\eta^2 = .98$, $p < .0001$), of the direction of the actual slope ($F[1, 9] = 3109.05$, partial $\eta^2 > .99$, $p < .0001$) and an interaction of the two factors ($F[2.37, 21.32] = 39.62$, partial $\eta^2 = .81$, $p < .0001$), meaning that participants were strongly influenced by the actual slope $\hat{\alpha}$. We confirmed this effect by focusing on scatterplots with a prescribed slope of zero, and with an actual slope extremely close to zero (between -0.03 and +0.03), and examined if participants were still able to extract the correct trend for those very hard trials. Indeed, the number of "increasing" responses was

significantly larger for positive slopes (95/156 = 61 %) than for negative slopes (68/170 = 40 %; $\chi^2 = 14.21$, df = 1, p<.001), indicating a significant sensitivity even within this limited range. The results so far indicate that participants' judgements were highly sensitive to the slope, the noise, and the number of points in a graph. We next tested the prediction of the "mental regression" hypothesis, according to which all of these effects may be subsumed by a single equation, the t-value that a statistician would compute to judge whether a significant trend is present in the data. For each graph, we computed the Student t-value associated to its Pearson coefficient of correlation and replotted the percentage of "increasing" responses as a function of that t-value (figure 14, right). As we can see, participants' mean performance was a sigmoid function of *t*. We compared the logistic regression of participants' responses as a function of either the actual slope ($\widehat{\alpha}$) or the t-value. A simple model comparison based on the Akaike Information Criterion (AIC) values revealed that participants' responses were significantly better predicted by the t-value (AIC for actual slope as predictor: 4138; AIC for t-value as predictor: 4086.7; $\Delta_{AIC} = 51.3$, $p < 10^{-16}$). Furthermore, we replicated the above sensitivity analysis once the data were accounted for by the t-value, and verified that the sensitivity values, once computed as a logistic function of *t*, were no longer affected either by σ or by *n* (respectively F[1.4, 12.63] = 1.27, partial $\eta^2 = .12$, p =.3 and F[1.61, 14.53] = 1.18, partial $\eta^2 = .12$, p=.32). In other words, the entire behavior was captured by a single value, the t-value (figure 14, fourth plot from the left).

One might argue that the Pearson coefficient of correlation *r* may also provide a good model of human behavior. Indeed, *r* is a measure of the strength of a linear trend that jointly summarizes the effects of the slope α and the noise σ – but crucially, not the number of points *n*. To investigate whether *r* alone sufficed to account for behavior, we performed the same sensitivity analysis as above, as a logistic curve either as a function of *r* or of *t*. Sensitivity

values, when computed as a logistic function of *r*, were still significantly affected by the number of points ($F[2.49, 22.39] = 10.64$, partial $\eta^2 = .54$, $p < .001$), whereas when computed as a logistic function of the t-value, as already noted, the effect of number of points disappeared.

Response times are also predicted by the t-value of the scatterplot

We then looked at the response times as a function of the prescribed slope and either the prescribed noise (figure 15, left) or the number of points (figure 15, middle). We conducted a repeated-measures omnibus ANOVA with median response times per condition as dependent variable, and prescribed slope, noise and number of points as within-participants factors. We found a main effect of slope ($F[6, 54] = 21.06$, partial $\eta^2 = .7$, $p < .0001$), a main effect of noise ($F[3, 27] = 20.07$, partial $\eta^2 = .69$, $p < .0001$), and an interaction of noise and slope ($F[18, 162] = 3.05$, partial $\eta^2 = .25$, $p < .0001$). As we can see from figure 15 (left), higher slope values and smaller noise values led to faster responses. As concerns the number of points, although there

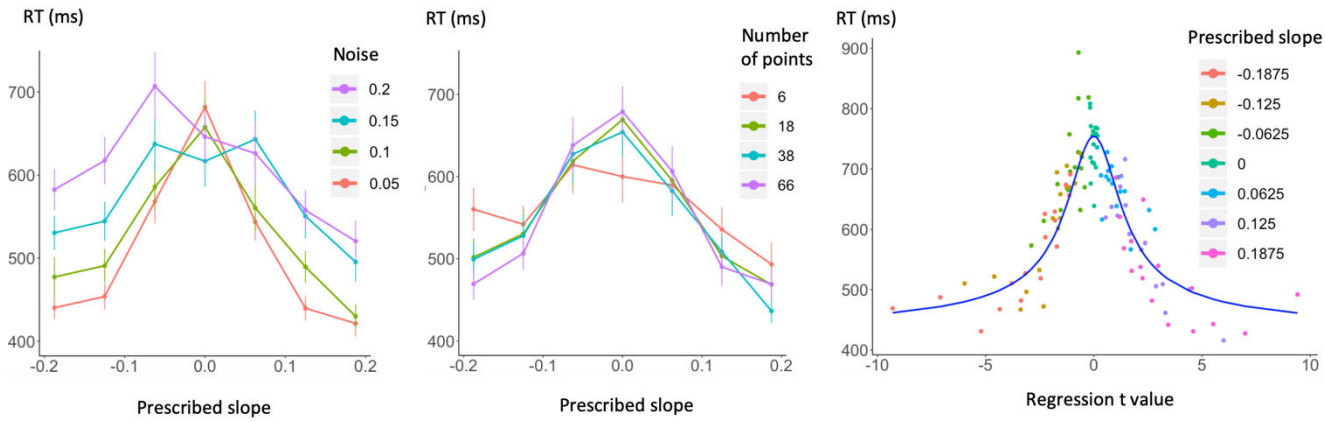


Figure 15. Response times in study 1 (trend judgment). Mean response times are shown as a function of the prescribed slope (α) and either the noise (σ, left) or the number of points (*n*, middle). Error bars indicate one standard error of the mean across subjects. The plot on the right shows the response times as a function of the t-value associated to the Pearson coefficient of correlation. The blue line indicates the response times predicted by a simple accumulation-of-evidence model (Gold and Shadlen, 2002).

was no main effect of this variable (F[3, 27] = 0.47, partial η^2 = .05, p =.7), it entered into significant interactions with both slope (F[18,162] = 1.88, partial η^2 = .17, p = .02) and noise (F[9,81] = 2.7, partial η^2 = .23, p <.01). Once again, as shown in figure 15 (right), those effects on median response times were well summarized by a single function of the t-value associated to the regression: RTs varied as a symmetrical, convex-upward, monotonously decreasing function of the distance of the t-value from zero.

Could the shape of this RT effect be predicted by the mental Pearson model? Following Gold & Shadlen (2002), we assumed that participants based their decisions on a noisy accumulation of evidence towards a fixed decision bound. Given our theoretical assumptions (see appendix A), we assumed that the noisy samples upon which the decision was based arose from a sampling of the regression t-value. Under those assumptions, the probability P of responding "increasing" and the mean RTs should follow the following joint equations (see Gold & Shadlen, 2002):

$$\mathrm{P(t)} = \frac{1}{1 + e^{-2B\mathrm{t}}}$$

and

$$\mathrm{RT(t)} = \frac{B}{|\mathrm{t}|} \tanh\left(B|\mathrm{t}|\right)$$

where B is a free parameter that corresponds to the slope of the psychometric function. We first fitted B using the performance data presented in figure 14 (right), then plugged this value into the second equation to obtain the shape of the predicted RT as a function of the t-value in our stimuli. We thus obtained dimensionless predicted RTs for each experimental condition, which we fitted to the data using a 2-parameter linear regression where the dependent variable was the across-participants mean RT in each of the 112 experimental conditions. The model provided a very good fit to the participants' RTs (r^2= 0.76; regression

slope = 286 ± 15 ms, $t(110) = 18.97$, $p < 10^{-16}$; intercept (non-decision time) = 429 ± 11 ms, $t(110) = 37.43$, $p < 10^{-16}$). The corresponding curve is shown in blue in figure 15 (right).

DISCUSSION

The results from the first experiment reveal that participants are able to quickly extract the linear trend of a scatterplot, without requiring any sophisticated training or long exposures to the stimulus. The fast presentation time (100 ms) and the short response times we observed (below 900 ms on average, see figure 15) imply that participants did not have time to perform complex calculations. Rather, they must have relied on an intuitive yet accurate estimate of the correlation. In fact, even on trials with a prescribed slope equal to 0, participants' performance remained above chance level, indicating a fine sensitivity to random variations in the graphs.

As expected, all three parameters of slope, noise and number of points significantly affected participants' accuracy, making it lower for shallower slopes, higher noise levels and smaller datasets. Similar effects were observed on RTs, and all those effects were subsumed by the t-value predicted by Pearson coefficient of correlation. By applying the model of Gold and Shadlen (2002) to our data, we found that participants' decisions followed the prediction of a classic accumulation-of-evidence decision model; in our data, the decision variable was the strength of the t-value associated with the Pearson correlation coefficient. This finding suggests that, before giving an answer, participants were accumulating evidence on the dataset's trend, and that this decision process approximated a statistical regression procedure. Indeed, participants' performance was better modeled as a function of the t-value rather than as a function of the prescribed slope. Thus, when detecting a scatterplot's

tendency, human adults do not solely rely on the slope of the linear regression, but extract an approximate summary statistic.

It is noteworthy that the mean RTs did not increase with the number of points *n*. On the contrary, RTs either stayed roughly constant or even decreased with *n* for large values of the slope α (see figure 15). Thus, participants did not treat the data points serially, as would have been unavoidable if the data were presented through numbers (such as in a tabular form), but took advantage of the graphic presentation to process them in parallel. We conclude that the human visual system affords a parallel form of approximate regression. Note that the coefficient of correlation formula involves only variances (in x and y) and covariances, all of which are sums or, equivalently, averages over values provided by each data point – and there is considerable prior evidence that the visual system can compute averages of various features in parallel across the items in a set ("ensemble perception"; see e.g. Chong & Treisman, 2003, 2005; Van Opstal et al., 2011). The findings of study 1 seemed to extend this concept to the case of statistical trend perception.

The present results go beyond previous studies (Cleveland et al., 1982; Lane et al., 1985; Rensink & Baldridge, 2010) which showed that, when participants are asked to judge the strength of an association presented in a scatterplot, their judgement is more accurate for higher levels of the Pearson correlation coefficient *r*, although still affected by the variance of the dataset (Lane et al., 1985). As pointed out by Surber (Surber, 1986), however, the interpretation of subjective ratings of correlation is complex and may be hard to relate to objective statistics. In agreement with this observation, Cleveland and colleagues (Cleveland et al., 1982) found that the perceived correlation was particularly small for values of $r < .5$, suggesting that only high levels of r ($> .95$) might relate to the actual *r* value. Also, prior beliefs have been recently shown to bias correlational judgments (Xiong et al., 2022). For these

reasons, in our study, we decided to avoid subjective measures of perceived correlation and asked for a more direct categorical trend judgement (increasing or decreasing). This forced-choice task provides solid evidence that participants could perform a "mental regression" with a high sensitivity, even when the slope of the linear trend in the data was shallow and the noise was high.

STUDIES 2, 3, AND 4: TREND JUDGMENT ACROSS AGE, EDUCATION, AND CULTURE

Humans often exhibit a surprising intuitive grasp of the core concepts of mathematics, physics or statistics. These intuitive abilities, which emerge in the absence of formal education, are likely to rely on a system of core implicit knowledge about the fundamental properties of the environment in which humans evolved (Spelke & Kinzler, 2007). A solid body of research shows, for example, that humans can accurately and quickly grasp the approximate numerosity of sets of objects (Dehaene, 2011) and perform approximate calculations even in the absence of formal mathematical education (Pica et al., 2004). Euclidean and even non-Euclidian geometrical intuitions of space are present in remote Amazon populations without access to formal education (Dehaene et al., 2006; Izard et al., 2011). In concrete settings, humans may also excel in intuitive physics: their misconceptions concerning the behavior of moving objects (McCloskey, 1983) disappear when questions are framed in familiar and real-life contexts (Kubricht et al., 2017). Humans are also remarkably good at performing intuitive statistical estimations in a variety of tasks (Nisbett & Krantz, 1983) and they seem to be endowed with these abilities from early on in their development (Xu & Garcia, 2008). Indeed, many quantitative assessments of intuitive mathematics and physics have been proposed, and they often predict the subsequent development of higher-level cognitive abilities (Baron-Cohen et al., 2001; Halberda et al., 2008; Perez & Feigenson, 2021; Piazza et al., 2010; Riener

et al., 2005), suggesting a strong link between basic intuitions and the formal mastery of complex concepts.

Whether these core intuitions extend to graphical representations is still an open question. In fact, despite the widespread use of charts and plots in our everyday life, no quantitative assessment of intuitive graphics skills has been proposed. Here, we show how the graph-based trend judgment task presented in study 1 offers a tool to quantify intuitive graphics skills. As I showed above, when facing a bivariate visual representation such as a noisy scatterplot (figure 1A), human adults can detect whether the curve is increasing or decreasing, regardless of the number of dots, noise level or slope of the graph. Their performance is simply subsumed by the *t*-value that a statistician would calculate to determine the significance of the trend in the data. In other words, the percentage of "increasing" responses is a sigmoid function of the *t*-value of the scatterplot. Here, we show that this task can be used to provide a quantitative assessment of graphicacy. As in any two-alternative forced choice task, the slope of the psychometric function (figure 14, right plot) provides a measure of a participant's sensitivity to detect variations in the stimulus: the steeper the function, the higher the participant's precision. In studies 2, 3, and 4, we demonstrate that this simple psychophysical task is reliable and provides a simple way to quantitatively assess how good a single individual is in such a trend judgment. We also investigated the emergence and distribution of trend judgment skills across people of different ages, education levels and cultures.

First (study 2), we tested intuitive graphics in a large-scale online sample of educated adults from all over the world, from which we obtained information about several demographical aspects including age, sex and education level, together with self-reports of mathematical and statistical understanding. Testing intuitive graphics on such a large and diverse population

offered insights into the predictors of these skills; also, it provided a large-scale replication of the psychophysical results obtained in the controlled laboratory environment of study 1, thus contributing to the growing but still scarce body of research on psychophysical measurements outside the lab (de Leeuw & Motz, 2016; Halberda et al., 2012; Semmelmann & Weigelt, 2017). Second (study 3), we explored whether graphicacy emerges as a result of graph exposure or whether some premises of graphicacy are universally available, even in the absence of formal education. To investigate this, we tested Himba participants, a Namibian people with no or little formal education, who are not exposed to any form of graphical representations. This sample of participants allowed us to test for the generalizability of such skills in non-western and unindustrialized societies, as previously been done for other intuitive skills (Spelke & Kinzler, 2007), including the perception of number (Pica et al., 2004) and geometry (Dehaene et al., 2006; Izard et al., 2011; Sablé-Meyer et al., 2021).

In addition (study 4), we tested graphicacy in French 6-year-old 1st-graders who had not yet encountered any graphical representation in their school curriculum. In this manner, we asked whether the ability to compute intuitive visual statistics from graphical representations arises early on in development, as should be the case if it relies on a core skill of human cognition, similar to number sense or shape perception. As proposed by the cultural recycling hypothesis (Dehaene & Cohen, 2007), the latter two evolutionary old cognitive functions sustain culturally learned skills (respectively, arithmetic and reading abilities). Similarly, humans' ability to read and interpret complex graphs might be sustained by more fundamental cognitive functions available early on in development and irrespectively of formal education, such as the ability to recognize the orientation of objects. As previously discussed, our findings from study 1 point indeed to that direction: simple trend judgments performed on noisy scatterplots are based on metrics close to their principal axis', thus

suggesting that the ability to identify objects' orientation might be a core skill at the root of graph perception and understanding. Showing that graphical intuitions are available to 1st-graders would constitute a further piece of evidence of its "core" nature.

METHODS

Experimental procedure and participants

Online participants. The online test was advertised and shared on social networks (mainly through Twitter). It could be performed either on computers or on tactile devices. Participants had to read and accept a written consent and to declare to be at least 18 years old before taking part in the experiment, in compliance with the local Ethical Committee that approved our research. Data collection for the purpose of the study started on the 15th of January and ended on the 15th of March 2021, as planned ahead of the experiment. The link to the test was still running after that date, but the data were not included in the current work.

Before taking the test, all participants answered a demographic questionnaire consisting in a series of single-answer questions about: country of origin, age, gender, number of previous participations in the task (if any) and the highest level of education attained. If participants declared to have completed a university degree, they were asked to choose the closest field of the degree within a list and their average grade in mathematics during their school and university years. Using a Likert scale (ranging from 1 to 10, with intermediate numbers not shown), all participants had to rate their subjective self-evaluation of their: familiarity with graphs, ability to read scatterplots, knowledge of statistics, current skills in mathematics, and current skills in their first language in terms of spelling, grammar and communication. Once the demographic questionnaire was completed, participants started the experiment (smartphone users were asked to rotate their phone horizontally: otherwise, the task would

not start; accidentally orientating the phone vertically during the task lead it to pause the experiment). The instructions and the questionnaire were available in six languages: English, French, Italian, Spanish, Portuguese and Chinese. 3943 subjects participated and completed the online experiment (the ones that did not complete the task were not included in the data analysis). 2409 of them declared being women, 1294 men, 183 non-binary, 20 "other" than the previous ones, and 137 preferred not to answer. The average age was 28.8 $\pm$ 9.6 years.

Himba. 87 Himba participants (39 women and 48 men) were recruited in small villages in the Kunene region, Northern Namibia. Most Himba do not know their age. Participants' age, 21.1 $\pm$ 9.4 years, was evaluated by local research assistants who were bilingual Namibians (in Otjiherero and English) and instructed each participant about how to perform the task on a tablet using the native language of the participant (Otjiherero). Before the experiment, each participant was provided with four examples of stimuli and the expected correct answers. Each participant indicated whether they had received any type of formal schooling. Rudimentary mobile schools (using black board and chalk) exist in the Kunene region and 12 participants declared having received at least one year of such form of schooling.

Children. 27 French 1st graders (6 $\pm$ 0.6 years) took part in the experiment (approved by the local ethical committee under the reference CER-Paris-Saclay-2021-046). 13 of them declared being girls, 14 being boys. They all completed the experimental tasks. Each child was accompanied by an experimenter to a silent room and invited to sit on a chair facing a table. Before starting the actual experiment, they performed three short behavioral tests: a one-minute reading task consisting in a series of French words of increasing difficulty; a one-minute counting task of sets of points of discontinuously increasing numerosity; a one-minute counting task of those same sets of points but organized in groups (e.g., 4 groups of 3). The first task provided a number of correctly read items in one minute, which was used as a proxy

of reading abilities. The difference in correctly enumerated items between the second and the third task provided an implicit measure of the mastery of arithmetic operations, because grouped items can be enumerated faster if children know how to perform mental arithmetic (Ciccione & Dehaene, 2020; Starkey & McCandliss, 2014). The main experimental task was performed on a tablet and, immediately before it, each child was provided with four examples of stimuli and their expected correct answers.

Experimental task

The task consisted in the rapid presentation (100 ms) of scatterplots (see figure 16A for a few examples). Participants performed a trend judgment task, identical to that described in study 1: they had to judge, as fast and accurately as possible, the trend of the scatterplot (increasing or decreasing), by pressing one of two separate keys on their computer keyboard or, if they played on a smartphone/tablet, by touching an upwards or a downwards arrow. For the online experiment, the response configuration of the keys and the arrows was randomly determined at the beginning of the experiment for each subject, in order to control for possible preferential response sides; also, each correct response was rewarded with a certain number of points, inversely proportional to the response time, in order to push participants to be both accurate and fast. To maintain a high level of attention in the task, consecutive correct responses were rewarded with increasingly higher points. Also, a pleasant sound followed each correct trial and an unpleasant sound followed each incorrect trial. For children and Himba participants, a smiling green face or a red unsmiling face was displayed instead of the numerical score. A fixation cross was presented for 1000 ms before the following trial appeared. The experimental session lasted around 6 minutes. Online and Himba participants had the opportunity to realize another run or to stop. Online participants could also check their percentage of correct responses and their ranking among other participants in the

world. For data analysis, we rejected any answer that was given after more than 5 seconds from the stimulus onset (0.75% of trials for online participants; 9.39% for children; 0.91% for Himbas).

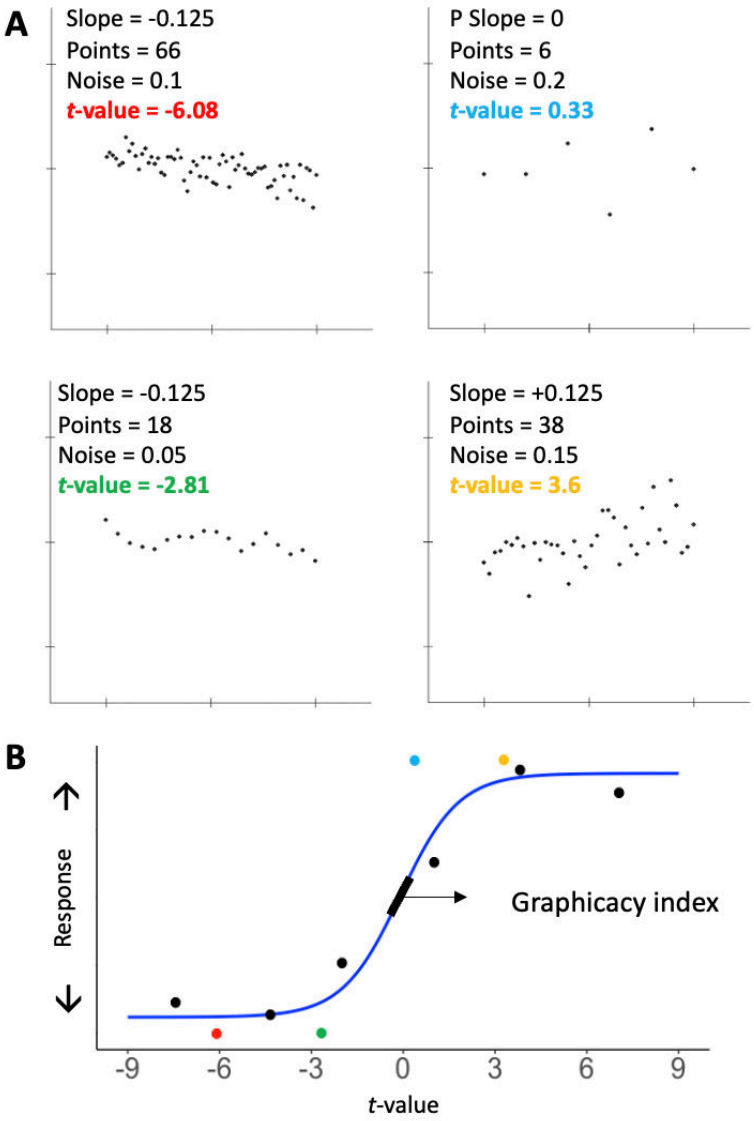


Figure 16. A simple test of graphicacy. A: four examples of stimuli shown to one participant in the trend judgment task, where the participant was asked if the graph was increasing or decreasing. The actual stimuli were white dots on a black background. Each scatterplot was created according to the combinations of different parameters: slope of the generative function, number of points, and noise level. Each thus had a certain *t*-value, corresponding to the *t* statistic used to calculate the significance of the trend in the dataset. **B**: Responses given by a representative participant are plotted as a function of the *t*-value of each scatterplot. Color dots show the data for the four example trials in panel A. For visualization purposes, the black dots show averages over bins of t-values. We fitted the data with a psychometric function (blue curve). The slope of the sigmoid, indicated by a black bar, evaluates the subject's sensitivity in the trend judgment task, and was used as a "graphicacy index", a proxy of the participant's intuitive graphics skills.

Stimuli

Both the experimental task and the stimulus generation algorithm were identical to the ones already used in the laboratory study version of the task (study 1). Figure 16 A shows four examples of stimuli. For each scatterplot, the *t*-value associated to its Pearson coefficient of correlation was calculated. Figure 16B shows examples of responses for one subject: each

answer is plotted as a function of the *t*-value of the corresponding trial. For each subject, we fitted a classic psychometric function to the data (shown in blue in figure 16B) and we extracted its slope, which provided a measure of precision at performing the trend judgment task. The first 12 trials for each subject were considered as practice trials and thus excluded from the computation of such index. Also, a minority of subjects that participated in the online experiment had a very large sensitivity index, meaning that their performance was close to perfect (in fact, it was better modelled by a step function rather than by a sigmoid one). To avoid excessive variability, all sensitivities higher than 5 (.03% of all participants) were capped at 5.

RESULTS

Study 2: performance in the trend judgment task is predicted by the *t*-value of the scatterplot

First, we looked at the percentage of "increasing" responses as a function of the prescribed slope (i.e., how steep is the scatterplot), the noise level and the number of dots (figure 17). We replicated results from study 1, finding that the proportion of responses "increasing" was affected by all the above parameters (figure 17A, left and middle plots). In an ANOVA on the proportion of "increasing" responses as a function of the prescribed slope, the noise level and the number of points, all factors had a significant main effect, and the prescribed slope significantly interacted with both the noise and the number of points (all $p < .001$). These findings confirm what is clearly visible in figure 17A: the smaller the slope of the graph, the higher the influence of the noise level and the number of points on the trend judgment task. No interaction effect was found between the noise and the number of points ($F[8.5, 19667.8] = .81$, $p = .6$), suggesting that the two factors independently affected human trend judgments.

All of these effects, as in study 1, were subsumed by an effect of the *t*-value associated to the Pearson coefficient of correlation (figure 17A, right plot). Accordingly, we computed a multiple logistic regression on "increasing" responses as a function of the *t*-value, the number of points and the noise level (averaged across the 112 combinations of the experimental conditions and across all subjects) and we found that only the *t*-value was a significant predictor of participants' responses ($\beta_{t\text{-value}}$= .77, p < .0001; $\beta_{\text{number of points}}$= .002, p = .83; β_{noise}= -.48, p = .91).

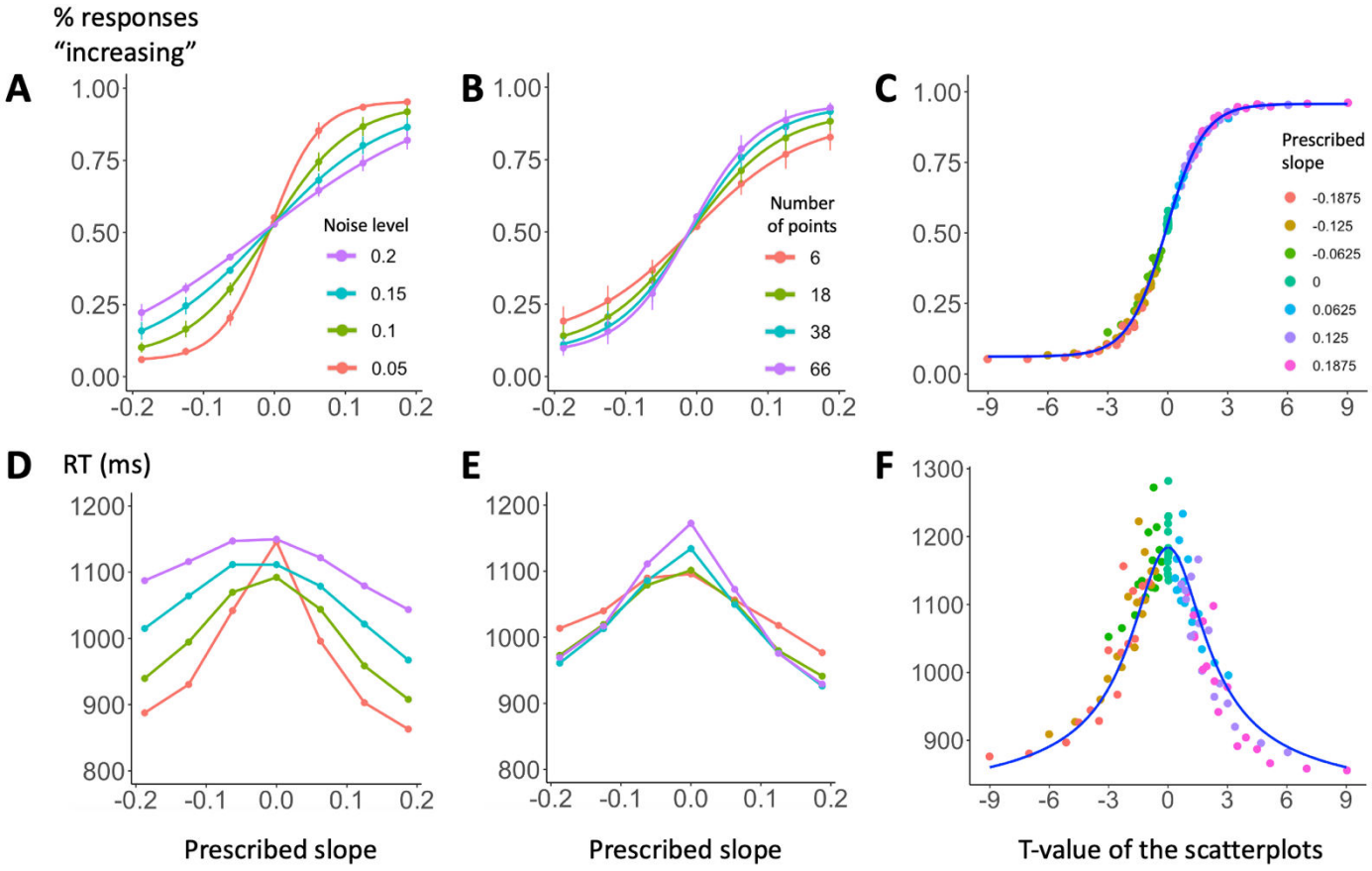


Figure 17. Psychophysics of graph perception (N=3943). The top row shows the percentage of responses "increasing" as a function of prescribed slope and noise level (**A**), prescribed slope and number of points (**B**), and *t*-value associated to the Pearson correlation coefficient of the scatterplot (**C**). Panel C shows that participants' performance can be subsumed by the t-value. **D-E-F**, equivalent plots for response time. The blue line in plot F shows the prediction of a simple accumulation of evidence model (Gold and Shadlen, 2002).

As far as response times for correct answers are concerned (figure 17D and 17E), we submitted them to separate linear regressions as a function of the three main experimental factors (the absolute value of the main slope, the noise level and the number of points) and with the subjects as random factors. We found that, while both the prescribed slope and the noise level significantly predicted the response times (β_{slope} = -946.5, $p < .0001$; β_{noise} = 920.3, $p < .0001$), this was not the case for the number of points ($\beta_{number\ of\ points}$ = .06, $p > .05$), thus suggesting a parallel processing of all items in the set. The simple model of noisy evidence accumulation already used in study 1 (Gold & Shadlen, 2002) correctly predicts response times on all trials (blue line in figure 17F, right plot), based on the responses given by the subjects.

Study 2: The trend judgement task provides a reliable index of graphicacy and its variations across individuals

We modeled participants' responses as a sigmoid function of the *t*-value of each stimulus they saw (figure 16B). We postulate that this function provides a measure of their intuitive graphics skills, which we called the "graphicacy index". Figure 18A shows the broad distribution of this index across the large sample we collected online (median value = 1.24). For the vast majority of participants (97.9%), the regression was significant and with a positive index, thus providing a reliable estimate. However, graphicacy varied considerably, with 95% of the distribution falling between 0.19 and 3.22.

To evaluate training effects during the course of an experimental run, we computed (for the 3419 participants that performed only one experimental run) the index separately on the first 50 trials and on the following 50 trials. Although the increase was significant (Wilcoxon signed rank test, $p < .0001$), it was small, passing from a median of 1.28 to 1.35 – and more crucially, there was a significant correlation between the two values ($r(3417) = .38$, $p < .0001$), thus

showing a relative stability of the graphicacy index. To further evaluate whether the graphicacy index remained stable over time, we computed the orthogonal linear regression between the above two index measurements and we found that the regression was close to one (1.02; 95% confidence interval = [.94, 1.1]), thus suggesting that, on average, the index did not increase nor decrease over time.

We then restricted the analysis to those subjects (n = 387) that completed more than one block of trials and analyzed the correlation between their graphicacy index in the first experimental run and in the second one: again, the two measures correlated ($r(385) = .49$, $p < .0001$) and the orthogonal linear regression between the two was still close to one (1.16; 95% confidence interval = [.95, 1.36]).

Overall, these results suggest that our measure of intuitive graphics skills is stable, at least in the absence of long training, and can be reasonably estimated in a 6-minute on-line test. It is likely that, at the individual level, a longer testing session would provide an even more reliable graphicacy index.

Study 2: Graphicacy correlates with statistical knowledge and academic field

We then tested whether graphicacy correlated with participants' self-evaluation of statistical knowledge. For the following analyses, in order to avoid any (although modest) effect of training described above, we included, for each subject, only their first block of responses.

We found a significant correlation (figure 18B; $r = .21$, $df = 3092$, $p < .0001$) between participants' graphicacy index and their self-reported statistical knowledge. Was this correlation specific to statistical knowledge? Among the subjects included in the analysis, a large majority (N = 2030) also answered a self-evaluation question on their first language skills, always using a scale from 1 to 10. We performed a multiple linear regression on the graphicacy index as a function of statistical knowledge and language skills, finding that the

former was a significant predictor (β = .06, p < .0001) but the latter was not (β = -.004, p = .65), thus suggesting that participants' ability to perform the task was not simply predicted by general personal skills (or self-confidence).

Figure 18C shows the graphicacy index as a function of the academic field in which graduate participants obtained their title: it was considerably higher for graduates in engineering, statistics and science (n = 1576, mean = 1.5) than for graduates in other disciplines (n = 1323, mean = 1.26; t(2892.8) = 8.81, p <.0001). In graduate subjects, the graphicacy index also significantly correlated with their reported average grade in mathematics (r = .04, df = 3028, p < .05).

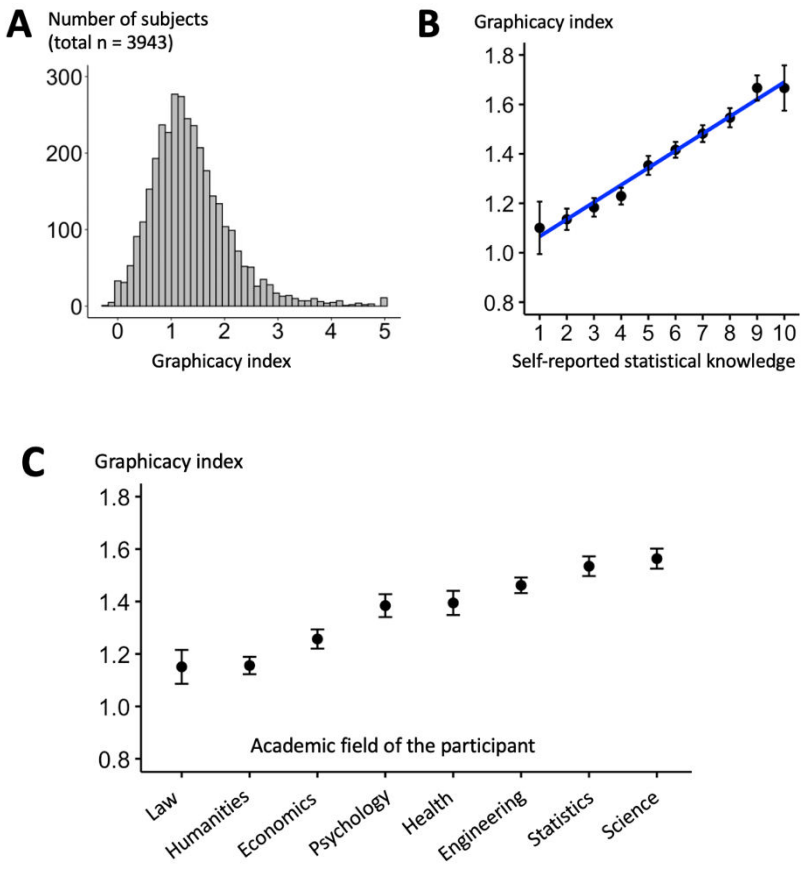


Figure 18. Inter-individual variability in graphicacy. A: Distribution of the graphicacy index across participants. **B**: Graphicacy increases as a function of participants' self-reported statistical knowledge, collected before the experimental task was introduced. **C**: Graphicacy in participants that obtained at least a bachelor degree varies as a function of the academic field in which they graduated (F(7,2891)= 15.57, p<.001; note that data were ordered according to each group's mean graphicacy index).

Study 3: performance in the trend judgment task for Himba participants is predicted by the *t*-value of the scatterplot, independently of age and educational level

Figure 19 (top plot) shows that the performance in the trend judgment task for Himba participants was well predicted by the *t*-value of the scatterplot. To statistically test for this, we computed a multiple logistic regression of responses "increasing" as a function of the *t*-value, the number of points and the noise level (averaged across the 112 combinations of the experimental conditions and across all subjects), and we found that, again, the *t*-value was the only significant predictor of participants' responses ($\beta = .27$, $p < .01$), while the noise ($\beta = .37$, $p = .92$) and the number of points ($\beta = .001$, $p = .86$) were not. The same findings held when we separated our data in three separate groups (figure 19, top right): teenagers (i.e., participants younger than 18 years old, N=36), unschooled adults (i.e., participants who did not receive any formal education, N=39), and schooled adults (i.e., participants who attended mobile schools during at least one year, N=12). For all these subgroups, responses were entirely accounted for by the *t*-value of the stimulus (all β with $p < .01$). The median graphicacy index for Himbas was of .32.

Study 4: Performance in the trend judgment task for 6-years-old children is predicted by the *t*-value of the scatterplot

The results described so far were also replicated in a group of 27 6-years-old children attending their first grade of primary school (figure 19, bottom). Although children's responses were noisier and never reached perfect performance (as is clear from the boundaries of the sigmoid function in figure 19, right plot), their responses were again significantly predicted by the *t*-value of the scatterplot, which alone accounted for children's performance: in fact, it was a significant predictor of their responses ($\beta = .17$, $p < .05$), whereas the noise ($\beta = 1.97$, $p = .58$) and the number of points ($\beta = -.0003$, $p = .98$) were not.

We then calculated, for each child, their intuitive graphics skills and correlated them with the two measures described in the methods' section: groupitizing advantage (an implicit measure of their arithmetic abilities) and their number of correctly read words in one minute (a proxy of their reading skills). Both correlations were significant (respectively: r = .51, df = 25, p < .01, and r = .46, df = 25, p < .05). It is worth noting that the performances in the implicit arithmetic task and in the reading one were also highly correlated (r = .7, df = 25, p < .0001). The median graphicacy index for children was of .12.

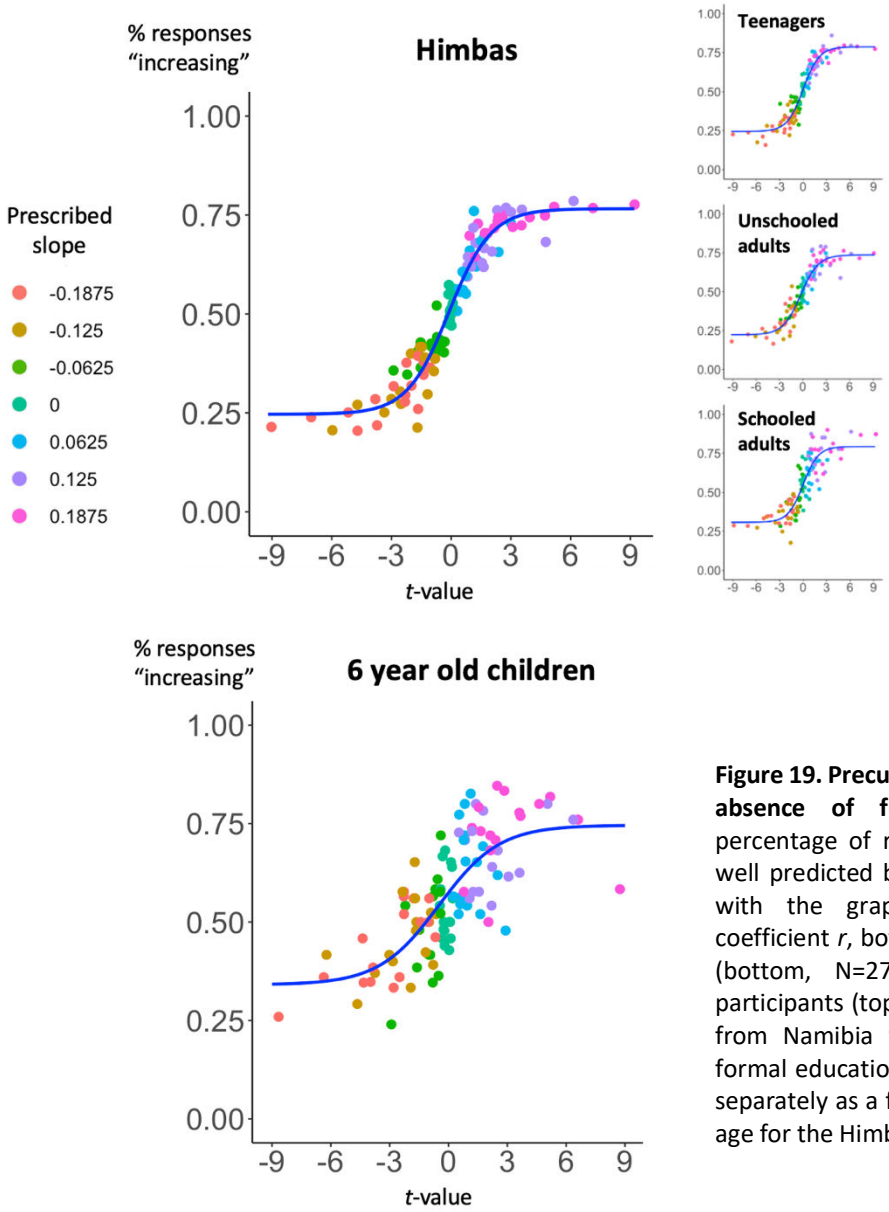


Figure 19. Precursors of graphicacy in the absence of formal education. The percentage of responses "increasing" is well predicted by the *t*-value associated with the graph Pearson correlation coefficient *r*, both in 6-years-old children (bottom, N=27) and in the Himba participants (top, N=87), an ethnic group from Namibia with reduced access to formal education. Insets show the effect separately as a function of schooling and age for the Himba people.

DISCUSSION

In study 2, we investigated in a large sample the human ability to perform a trend judgment task on a noisy graph (i.e., “Does this graph go up or down?”). Analyzing the responses of 3943 participants that performed the task online on computers or tactile devices, we found that their accuracy was affected by all three manipulated factors, namely, the steepness of the graph, its noise level and the number of points. In terms of response times, there was a significant effect of steepness and noise but not of the number of points. As already suggested by study 1, fast intuitive statistical judgments on graphs with Gaussian noise thus seem to operate similarly to ensemble perception, the human ability to rapidly extract the “average” of visually displayed items, without focusing on each particular element in the set (Cui & Liu, 2021; Szafir et al., 2016; Whitney & Yamanashi Leib, 2018).

Crucially, participants’ responses were entirely predicted by the *t*-value associated to the Pearson coefficient of correlation of the graph, showing that humans’ trend judgments approach those of an optimal statistical model (Peterson & Beach, 1967). Thus, humans are not “naïve” intuitive statisticians who wrongly assume that a sample of information derived from a restricted number of items is representative of the entire population (Fiedler, 2000; Juslin et al., 2007). In agreement with this view, it was shown that people do not include the sample size in their variability estimations (Kareev et al., 2002). On the contrary, our studies demonstrate that, for datasets represented in a bivariate graphical format, people correctly assess both variability and sample size (as proved by their reliance on the *t*-value) when performing a trend judgment. In other words, at least at a perceptual level, humans are not naïve in their statistical estimates but seem to take into account all the parameters of the dataset.

The replication of study 1, which was conducted in a controlled laboratory environment, is important both empirically and methodologically. It confirms that, unlike popular opinion among many scientists, psychophysical studies on accuracy and response times do not need to be confined to a controlled setting and can be successfully performed online. This clearly reduces research times and costs, especially when participants, such as in the present online study, were asked to participate on a purely voluntary basis and with no reward beside that of personal enjoyment.

The first drive of study 2 was to introduce a quantitative measure of intuitive graphics skills (the graphicacy index). We operationalized it as the slope of the psychometric function of responses "increasing" (Klein, 2001) and we found that such measure, which was highly variable in the general population, was predicted by participants' self-evaluation of statistical knowledge (but, crucially, not by their self-evaluation of first language skills). This suggests that numerical cognition might influence the development of intuitive graphics skills, similarly to the positive effect of mathematical understanding on the intuitive number sense (Piazza et al., 2013). Whether a better grasp of numerical concepts strengthen graph-based statistical judgments (and/or vice versa) remains an open question that could be better addressed in the future through a finer assessment of participants' numerical skills. This is particularly necessary in the case of children: in our sample, we found a strong correlation between intuitive graphics skills and both arithmetic and reading performance, which does not allow to conclude for a specific correlation of intuitive graphics with numerical cognition. Interestingly, however, a relation between complex graph understanding and numerical cognition does indeed seem to exist (Ludewig et al., 2020). While the link between intuitive statistics' skills and complex statistical graph understanding remains to be shown, we believe that our trend judgment task could be an adequate assessment tool for the former, being

simple and fast (less than 10 minutes) to perform. The test is publicly available online and can be freely run by all researchers interested in investigating the correlation between their participants' intuitive graphics skills and other abilities (https://neurospin-data.cea.fr/exp/lorenzo-ciccione/graphicacy-index/). Also, future research could determine if a long training on the task would be able to improve higher level graph understanding, in the same way that training the intuitive number sense has been shown to increase mathematical proficiency (Park & Brannon, 2013).

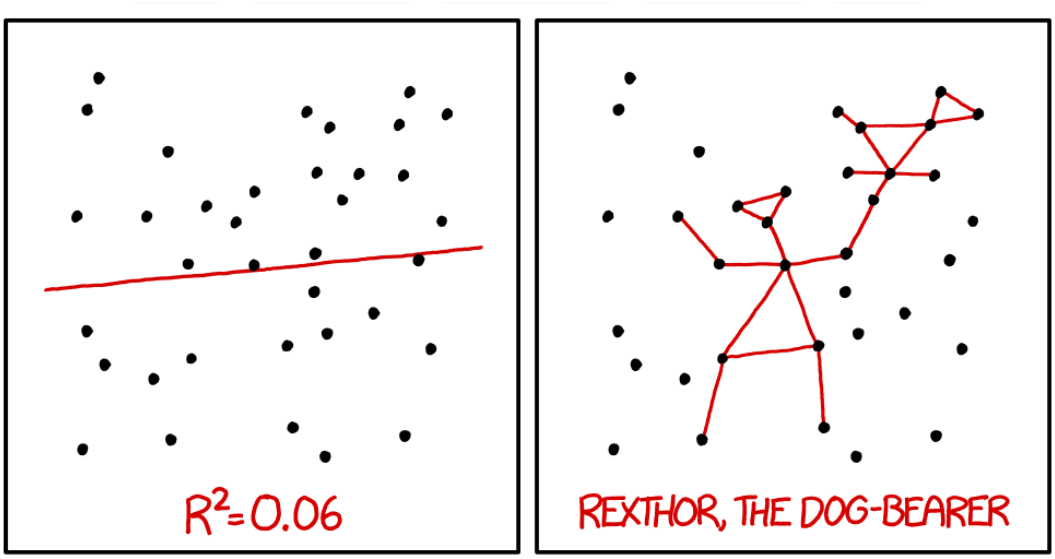
The core drive of study 3 and study 4 was to investigate whether intuitive graphics skills are available uniquely to individuals previously exposed to graphical representations, or whether they could be found in the absence of any such exposure. Children have been recently found to be able to discriminate the visual items of graphs, such as the location of datapoints and their size (Panavas et al., 2022), but no study had yet specifically tested whether they are also able to perform statistical judgments on noisy graphs. We show here that French 6-years-old children (unexposed to graphical representations) and uneducated Himba individuals (living in an unindustrialized remote society, in Northern Namibia, where there is no form of 2D visual representations, including graphical representations) base their intuitive decisions on the *t*-value of the scatterplot. This shows that intuitive graphics skills are universally available and emerge early on in development, irrespectively of previous exposure to graphical representations. The present finding echoes previous data supporting the existence of a universal understanding of quantities (Dehaene, Stanislas, 2011), geometrical shapes (Sablé-Meyer et al., 2021), probabilities (Xu & Garcia, 2008), physics (Atran, 1998), and human psychology (Bjorklund, 2014). However, we found that the graphicacy index was much lower in Himba and children than in educated adults. Taken together, our results suggest that intuitive graphics might be refined with exposure to statistics and graphical representations

but, crucially, they do not vary qualitatively with age and education. Recent evidence in numerical cognition (Piazza et al., 2018) suggests that the precision of numerical estimations increases with education through an improved ability to focus on relevant information in the task (thus discarding non numerical features). Future studies may investigate whether a progressive refinement of the filtering of irrelevant information (e.g., outlier datapoints or large noise) is also responsible for the relationship between the graphicacy index and education.

In sum, by investigating the premises of human intuitive graph perception, our study laid the foundations of a quantitative assessment of human graphicacy; such assessment would be essential in building effective and early educational interventions that might in return strengthen the comprehension of the complex graphs that humans are more and more routinely confronted with.

CHAPTER 2

MENTAL REGRESSION: HUMAN ACCURACY AND BIAS IN PERFORMING LINEAR REGRESSION AND EXTRAPOLATION



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

XKCD comics

The findings presented so far indicate that human performance in a trend judgment task approaches an optimal statistical model that would compute the t-value of the graph in order to determine its trend. But what is the actual regression line that participants would intuitively derive from a graph? Does it also approach classic statistical models? In the two studies described in this second chapter I answered these questions by investigating the actual slope of the mental regression performed by participants over a noisy scatterplot. We did so by either asking them to adjust a line (line adjustment task, study 5) or by asking them to extrapolate a point outside the shown scatterplot (extrapolation task, study 6). Both studies were performed in the laboratory and comprised 10 participants each.

STUDY 5: LINE ADJUSTMENT

In study 5, we further probed whether human participants act as good statisticians, and which procedure they use to approximate linear regression. Whereas study 1, 2, 3, and 4 solely asked participants to decide whether an ascending or descending trend was present, we now asked participants to report the slope of the best-fitting regression line. They did so by using a trackpad to adjust the tilt of a line on screen. What predictions can we derive for this task? If participants used a mental process equivalent to the classic simple linear regression (also called "ordinary least squares" method, OLS), then their average slope estimate should be centered on the prescribed slope used to generate the graph, and should not be influenced by the number of data points (n) nor by the noise level (σ). The reason is that the OLS estimate of slope is what statisticians call an "unbiased estimator", i.e. an estimate whose expected value is equal to the prescribed slope (McElroy, 1967; Puntanen & Styan, 1989). In this study, we tested whether participants' slope estimates follow this law.

METHODS

Participants

10 participants were recruited for the experiment (age: 25.2 $\pm$ 1.2, 4 females, 6 males; the inclusion criteria were the same of study 1). They were paid 10 euros for their participation. The experimental session lasted approximately 45 minutes. The experimental procedure was approved by the local ethical committee. One participant was excluded by the analyses because he failed to perform the task (he did not adjust the line for more than half of trials).

Procedure

Stimuli and procedure were identical to study 1, except that immediately after the scatterplot (presented for 100 ms), a blank screen appeared for 100 ms and then an adjustable line was

shown in the middle of the screen (see figure 20). The line was initially horizontal, but participants were asked to adjust it as accurately as possible by moving their right index on the computer trackpad. The center of the line was kept fixed (since, as in study 1, the OLS regression line of the scatterplot always passed through the exact center of the graph), so that moving the finger up or down the trackpad resulted in a rotation of the line around its center, whose angle was proportional to finger displacement; moving the finger up tilted the line in the counterclockwise direction, whereas moving the finger down tilted it in the clockwise direction. The participants were informed that we would measure the accuracy of their fit and, for this reason, they were invited to take their time to perform the task. When the adjustment was completed, they simply had to press the trackpad in order to confirm their answer and move to the next trial, which was, as in study 1, preceded by a 1s fixation cross.

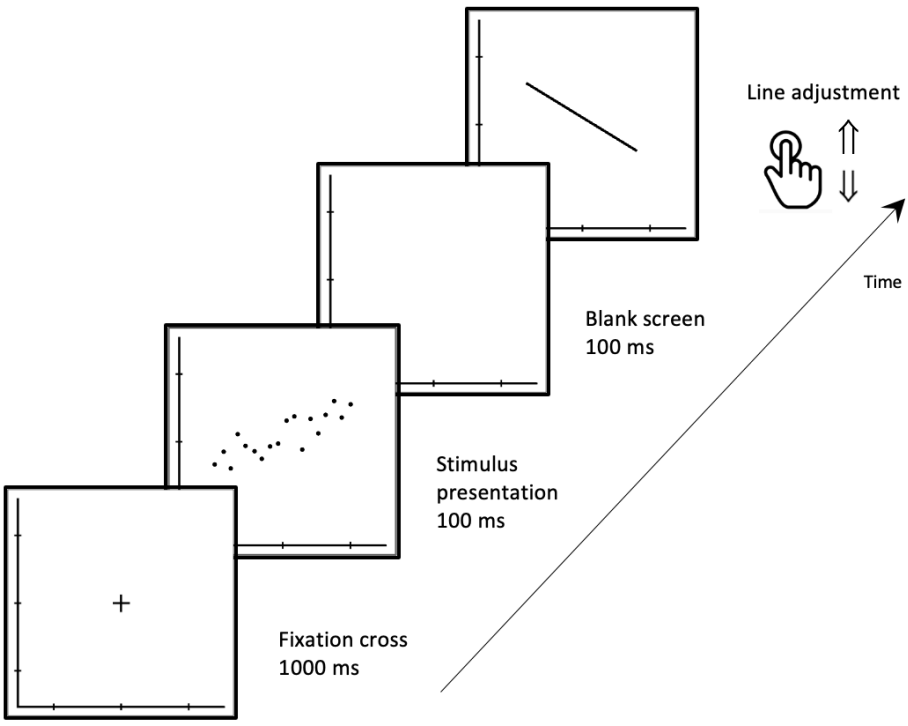


Figure 20. Experimental design for study 5 (line adjustment). Subjects were presented with a simple scatterplot, generated by a linear function plus noise, with a variable number of data points. Immediately after, they were asked to adjust a line by moving their finger on a trackpad, in order to provide an estimation of the regression line underlying the noisy scatterplot.

Exactly as for study 1, the task was divided into 6 blocks, each comprising all 112 conditions; the duration of each block was now ~6 minutes. After each block, the participants could take a short break and received feedback on the total number of correct responses they gave (for this feedback, a response was considered as correct if its slope had the correct sign, i.e. positive or negative). Before the beginning of the actual experiment, 25 practice trials were conducted under the supervision of the researcher, in order to control for the correct execution of the task.

RESULTS

Increasing and decreasing judgments: replication of study 1

To see if we could replicate the findings of experiment 1, we first categorized the participants' responses as increasing or decreasing (based on the slope of their regression lines) and examined the proportion of "increasing" responses as a function of the prescribed slope and either the prescribed noise (figure 21, bottom left) or the number of points (figure 21, bottom middle). A repeated measures ANOVA on the percentage of "increasing" responses confirmed the statistical significance of the data presented in the figures, which closely paralleled the findings from study 1: we again found a main effect of the prescribed slope ($F[6, 48] = 508.51$, partial $\eta^2 = .98$, $p < .0001$) and its interaction with noise level ($F[18,144] = 12.36$, partial $\eta^2 = .61$, $p < .0001$) and number of points ($F[18,144] = 8.88$, partial $\eta^2 = .53$, $p < .0001$). Once again, the closer the slope was to zero, the higher the influence of the noise and of the number of points (figure 21). As in study 1, we conducted an analysis of each participant's sensitivity as a function of the noise σ and number of points *n*. We fitted a logistic regression to the percentage of "increasing" responses (as a function of the prescribed slope), separately for each participant, noise level, and number of points, and used the slope of the logistic function

as an indicator of sensitivity. In the few cases (5 out of 144 combinations) where the logistic regression was not meaningful, since the data were better modeled by a step function, the values were substituted with the maximum observed value. We submitted the resulting sensitivity values to a repeated-measures omnibus ANOVA and found a significant main effect of noise ($F[3, 24] = 96.65$, partial $\eta^2 = .92$, $p < .0001$) and a significant main effect of the number of points ($F[3, 24] = 24.45$, partial $\eta^2 = .75$, $p < .0001$). As we can see from figure 21 (bottom left and middle), participants were significantly more sensitive to datasets having a smaller noise and a higher number of points, closely replicating the results from study 1. We also looked again at the fraction of "increasing" responses as a function of the prescribed slope and the direction of the actual slope compared to the prescribed one (i.e., above or below it): a repeated-measures omnibus ANOVA revealed a significant effect of the prescribed slope ($F[2.03, 22.44] = 563.86$, partial $\eta^2 = .99$, $p < .0001$), of the direction of the actual slope ($F[1, 8] = 437.65$, partial $\eta^2 = .98$, $p < .0001$) and an interaction of the two factors ($F[2.74, 21.96] = 29.79$, partial $\eta^2 = .79$, $p < .0001$), meaning that participants, once again, were able to base their judgement on the actual value $\hat{\alpha}$ rather than the prescribed slope α.

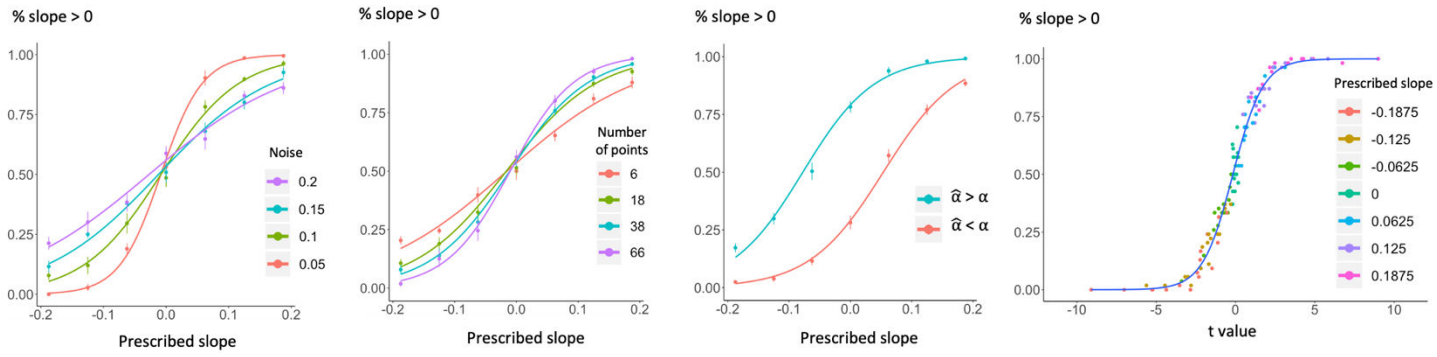


Figure 21. Accuracy of human subjects in adjusting a line over a noisy scatterplot. The percentage of lines adjusted with a positive slope (which is analogous to responses "increasing" in study 2) is affected by the prescribed slope of the graph (α), the noise in the graph (σ) and the number of points (n). The third plot shows that subjects' responses depended not only on the prescribed slope α, but on the actual slope $\hat{\alpha}$ after the addition of noise. The fourth plot shows that all the effects of noise, number of points and slope could be subsumed by an influence of the t-value associated with the Pearson coefficient of correlation, as if participants performed a mental linear regression. In this graph, each dot represents the mean, across trials and subjects, of all data for one of the 112 experimental conditions determined by each combination of α, σ and n. Data perfectly mimic those presented in figure 14.

Exactly as for study 1, for each graph, we computed the Student t-value associated to its Pearson coefficient of correlation and replotted the percentage of "increasing" responses as a function of that t-value (figure 21, bottom right). As we can see, participants' mean performance was once again a sigmoid function of *t*. We compared the logistic regression of participants' responses as a function of either the actual value of the slope ($\hat{\alpha}$) or *t*. A simple model comparison based on the Akaike Information Criterion (AIC) values revealed that participants' responses were again significantly better predicted by the t-value (AIC for actual slope as predictor: 4815; AIC for t-value as predictor: 4574; Δ_{AIC} = 241, $p < 10^{-16}$). We also replicated the above sensitivity analysis once the data were accounted for by the t-value, and verified that neither σ nor n played a significant role (respectively: F[3, 24] = 2.69, partial η^2 = .25, p =.07 and F[3, 24] = 1.04, partial η^2 = .12, p=.39), confirming the results from study 1: the entire behavior was captured by a single value, the t-value (figure 21 bottom, fourth plot from the left).

Slope estimation: participants minimize orthogonal distance from the fit

Figure 22 (top) shows the slope estimates, averaged across participants, as a function of the prescribed slope and either the noise level (left) or the number of points (right). We conducted a repeated measures ANOVA on participants' median estimated slopes. As expected, we found a main effect of slope (F[6, 48] = 91.6, partial η^2 = .92, p <0.001): as the prescribed slope increased continuously across 7 levels, so did the participants' estimates. However, the values that they reported were always in excess of the ideal slopes, both in the positive and in the negative direction (see figure 22, dashed line). Furthermore, this tendency to exaggerate the linear trends increased with noise level, and also with the number of points, as attested by significant interactions of prescribed slope and noise level (F[18, 144] = 3.56, partial η^2 = .31, p <0.001), and prescribed slope and number of points (F[18, 144] = 9.63,

partial η^2 = .55, p <0.001), as well as a triple interaction of slope, number of points and noise (F[54, 432] = 2.07, partial η^2 = .21, p <0.001). The nature of this bias can be described as follows. First, participants always overestimated positive slopes, and did so with a bias that increases with noise level and number of points (ANOVA restricted to positive slopes: main effect of noise, F[3,24] = 6.43, partial η^2 = .45, p <0.01; main effect of number of points, F[3,24] = 12.48, partial η^2 = .61, p<.0001). Second, conversely, participants always underestimated negative slopes, again increasingly so for larger noise levels and numbers of points (ANOVA restricted to negative slopes: main effect of noise, F[3,24] = 3.48, partial η^2 = .30, p =0.03; main effect of number of points, F[3,24] = 20.59, partial η^2 = .72, p<.0001).

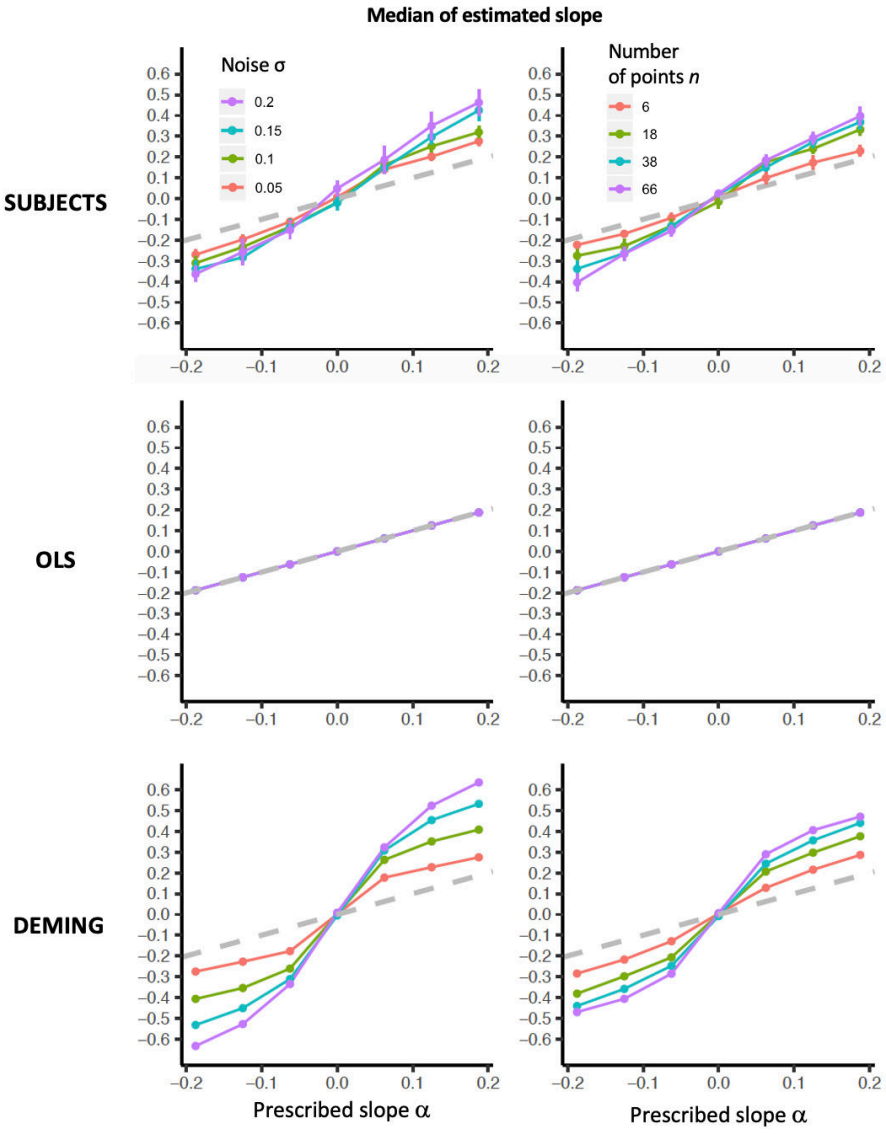


Figure 22. Human regression slopes and theoretical predictions. Slopes reported by the participants (top) and predicted by the OLS and Deming regression models (middle and bottom). Values are plotted as a function of the prescribed slope α, noise σ and number of points *n*. The dashed line represents the ground truth, i.e. the prescribed slope α. Subjects' median estimated slopes show a bias similar to Deming predictions (figure 23).

DISCUSSION

The significant dependency of participants' estimated slope on both noise and number of points violates the predictions of simple linear regression (OLS). Since OLS slopes are unbiased statistical estimators of the true underlying slope, OLS predicted no effect of either noise or number of points on the slope estimates (figure 22, middle). Those predictions were clearly violated in the data. Note in particular that the more data points were present, the more the participants' slope estimates were biased towards exceedingly extreme values. This finding may seem paradoxical, given that in OLS (red line in figure 23), a larger number of data points implies that the regression can be estimated with greater precision – and such an effect was indeed found in participants' proportion of "increasing" responses both in the trend judgment task and in the line adjustment one.

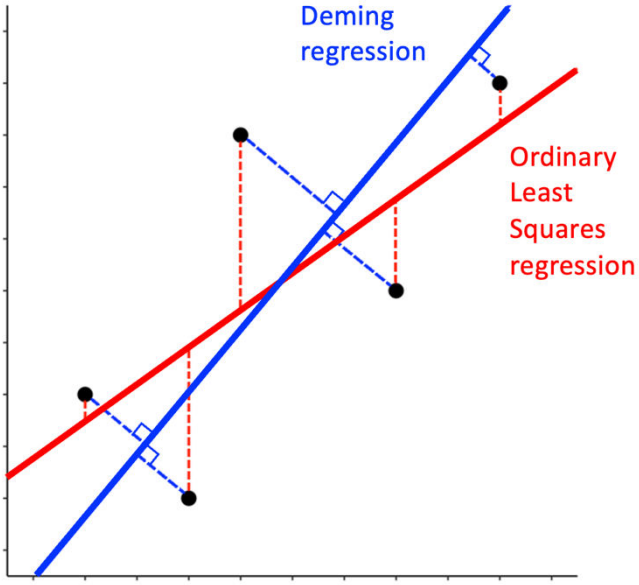


Figure 23. Illustration of the difference between ordinary least squares (OLS) and Deming regression. OLS regression (red line) minimizes the sum of the squares of the vertical distances of the points to the line. It is appropriate when the x values are fixed and there is noise only in the dependent variable (y axis). Deming regression (blue line) minimizes the sum of the squares of the orthogonal distances of the points to the line (assuming equal variance on the x and y measures). It is appropriate when the measurement is noisy on both the x and y axes.

How can we explain the participants' behavior? A key observation is that simple linear regression, based on ordinary least squares, is not the only procedure that can be used to estimate the slope of a graph. Indeed, it is not even always the optimal one. If there is variance in both the x and the y measurements, statisticians recommend the use of another procedure

termed “Deming regression” (Deming, 1943; Linnet, 1998; Martin, 2000). This procedure belongs to the category of “errors-in-variables models” which optimally account for the fact that there can be measurement errors on both x and y axes. The gist of Deming regression is illustrated in figure 23 (blue line). Essentially, it can be conceived as an “orthogonal regression”, seeking the line that minimizes the sum of square distances to the data *simultaneously in both the x and y dimensions* (strictly speaking, this is true only when assuming equal noise on x and y; otherwise one of the axes must first be scaled). This is the appropriate thing to do if x itself is the result of a noisy measurement. Classical regression (OLS), on the other hand, minimizes the sum of square distances only along the y axis; this is the proper thing to do if x is a fixed experimental factor, and only y is a noisy measure.

OLS regression is clearly appropriate for our graphs, which were created by keeping x fixed and generating y using a linear equation plus noise. However, participants, unaware of this fact, might have applied a procedure akin to Deming regression, perhaps because they treated the x and y coordinates of the dots as equivalent and therefore both potentially subject to noise. Indeed, Deming regression presents the unique advantage of yielding an identical regression line whether y is regressed on x, or x is regressed on y. This is not true of OLS regression, which treats x and y asymmetrically. Note, however, that the correlation coefficient *r*, and therefore the *t* test, are symmetrical in x and y, and are appropriate measures of the presence and strength of a linear trend for both OLS and Deming regression.

Figure 22 (bottom) shows the slopes predicted by Deming regression (calculated over 50000 stimuli generated through the identical algorithm used to generate the stimuli presented to participants). Remarkably, the Deming model made predictions strictly parallel to what we observed in our data: the median slope increased with both the noise level and with the number of data points. For Deming regression, one must provide not only the x and y values,

but also the ratio of the variances of the errors on x and on y. Here, we used the ratio of the empirical variances in the graphs, but qualitatively similar predictions were obtained if we assumed a fixed ratio of 1.

To summarize, study 5, using a line adjustment procedure, closely replicated the results of study 1 with binary trend judgement (increasing or decreasing). This parallelism suggests that, when asked to quickly extract the tendency of a scatterplot through a simple binary choice, humans can be as accurate as when precisely adjusting a regression line. Remarkably, both fast (study 1) and slow (study 5) judgments were affected by the same stimulus parameters, namely the slope, the noise and the number of points. Crucially, behavior was again subsumed by the t-value associated to the Pearson coefficient of correlation. This finding offers a methodological guidance for future experiments in graph perception: fast binary choices might be as informative as slow line adjustments when investigating the human perception of positive or negative trends in scatterplots.

However, slope adjustment also revealed a result that was inaccessible to binary judgments: humans are biased in their estimations of linear trends. They overestimate positive slopes, underestimate negative ones, and those biases increase with noise and with number of points. These findings refute the hypothesis that human adults compute a traditional OLS regression, and instead suggest that participants might use Deming regression when fitting a line to a noisy scatterplot. Simulations showed that Deming regression, far from being unbiased, leads to exactly the same qualitative biases as observed in humans.

Deming regression feels reasonable because it essentially consists in finding a line that minimizes the Euclidean distances to all points, thus treating the cloud of dots as a 2-dimensional shape, without distinguishing the x and y measurements (as OLS does). Thus, Deming regression, yields the same line whether y is regressed on x or vice-versa, unlike OLS.

Deming regression might have been induced by the stimuli we used, which were square graphs with identical layouts for the x and y axes, thus perhaps encouraging participants to treat the x and y axes as two noisy measurements. However, note that the x values were always equally spaced discrete samples, a fact that was particularly obvious for small numbers of points (see figure 1, n = 6); yet even in this case, the Deming-like bias was present. Thus, our findings suggest that human participants fail to apply the most standard regression procedure, ordinary least squares, and exhibit a strong bias. But which consequences might this bias have on human decisions and activities?

Indeed, regardless of its ultimate cause, the fact that human adults compute a Deming rather than an OLS regression may have considerable implications in real-life uses of graphical representations, such as in finance, where stock markets' noisy graphs are often used by investors to make quick selling or buying decisions. Biases in economic and financial behavior have typically been reduced to various cognitive biases (Kahneman, 2003; Ricciardi & Simon, 2000), such as confirmation bias (Nickerson, 1998) or loss aversion (Kahneman et al., 1991). Our findings suggest that such biases, although certainly at play, might not be the only factors influencing financial behavior. The Deming bias implies that investors could be more likely to keep investing in stocks showing an uprising trend (or selling stocks revealing a negative trend), because they perceive the trend as steeper than it actually is. Indeed, finance experts strongly rely on the slope information when looking at a graph (Beattie & Jones, 2002), and it is known that data series presented in graphical rather than tabular forms generally lead to a worse encoding of the actual trend in a dataset (DeLosh et al., 1997; Lawrence & Makridakis, 1989). Further studies of graph perception could be performed with specific populations such as finance experts in order to disentangle the biases of geometrical origin from other reasoning and cognitive biases, and to measure the practical import of the present findings.

STUDY 6: LINEAR EXTRAPOLATION

The studies described so far indicate that human adults can categorize a linear trend as ascending or descending, and approximate its slope. In study 6, we examined if their intuitive statistical skills also allowed them to perform a third task: linear extrapolation. We refer to extrapolation as an estimation that is made beyond the original observation range, assuming that the trend underlying the scatterplot will continue to be the same. To test it, we engaged participants in an extrapolation task, in which they adjusted a point vertically to place it on their best estimation of the regression line fitting the data points. This instruction aimed to minimize the tendency of participants to add noise to their extrapolations in order to match the noise in the graph, a phenomenon already described by Bolger and Harvey (1993).

Our predictions were simple: if participants relied on Deming regression, then they should produce exaggerated estimates (deviating too far either upwards or downwards, depending on whether the main slope is positive or negative), and this bias should be all the more pronounced that the noise in the scatterplot is high.

METHODS

Participants

10 participants were recruited for the experiment (age: 24.6 $\pm$ 1.8, 5 females, 5 males). All participants met the same inclusion criteria as in study 1 and 5. They were paid 5 euros for their participation. The experiment lasted approximately 30 minutes and was approved by the local ethical committee.

Experimental design and stimuli

The stimuli were generated according to the same algorithm of previous studies. Five levels of prescribed slopes were used to generate the scatterplots (-0.48, -0.24, 0, +0.24, +0.48). The

number of points was kept fixed at 18. The noise levels were the same used in experiments 1 and 2 (0.05, 0.1, 0.15, 0.2). The scatterplot was now confined to the left part of the screen, while the right one served as the extrapolation area (see figure 24A). The location of the scatterplot was vertically jittered by a random amount in order to induce participants to avoid responding at the same location; the jitter was later corrected for in our analyses. We included a considerable margin (12.5% of the screen) above and below the locations of the correct answers, where no expected correct answers could fall into; this was done in order to allow participants to give a free and unconstrained response, even if considerably higher or lower than the correct one.

Procedure

The procedure closely followed the previous studies, except that each scatterplot was now presented on the left side of the screen and for a long duration (until the response). On the right side, a single point was shown at one of two possible x coordinates (either x = 1.3 or x = 1.6, which we refer to as "probed positions") and at a y coordinate corresponding to the middle of the y axis (Figure 24A). Participants were asked to vertically adjust the point as accurately as possible by moving their right index on the computer trackpad. Once they were satisfied with the given response, they confirmed it by pressing the trackpad. Participants were explicitly asked to give an intuitive answer and to locate the point on their best estimation of the regression line of the scatterplot. The task was divided into 7 blocks, each comprising one trial for each of the 40 conditions (5 slopes, 4 noise levels, two probed positions), for a total of 280 trials. The duration of each block was ~4 minutes. After each block, the participants could take a short break. Before the beginning of the actual experiment, 25 practice trials were conducted under the supervision of the researcher, in order to control for the correct execution of the task.

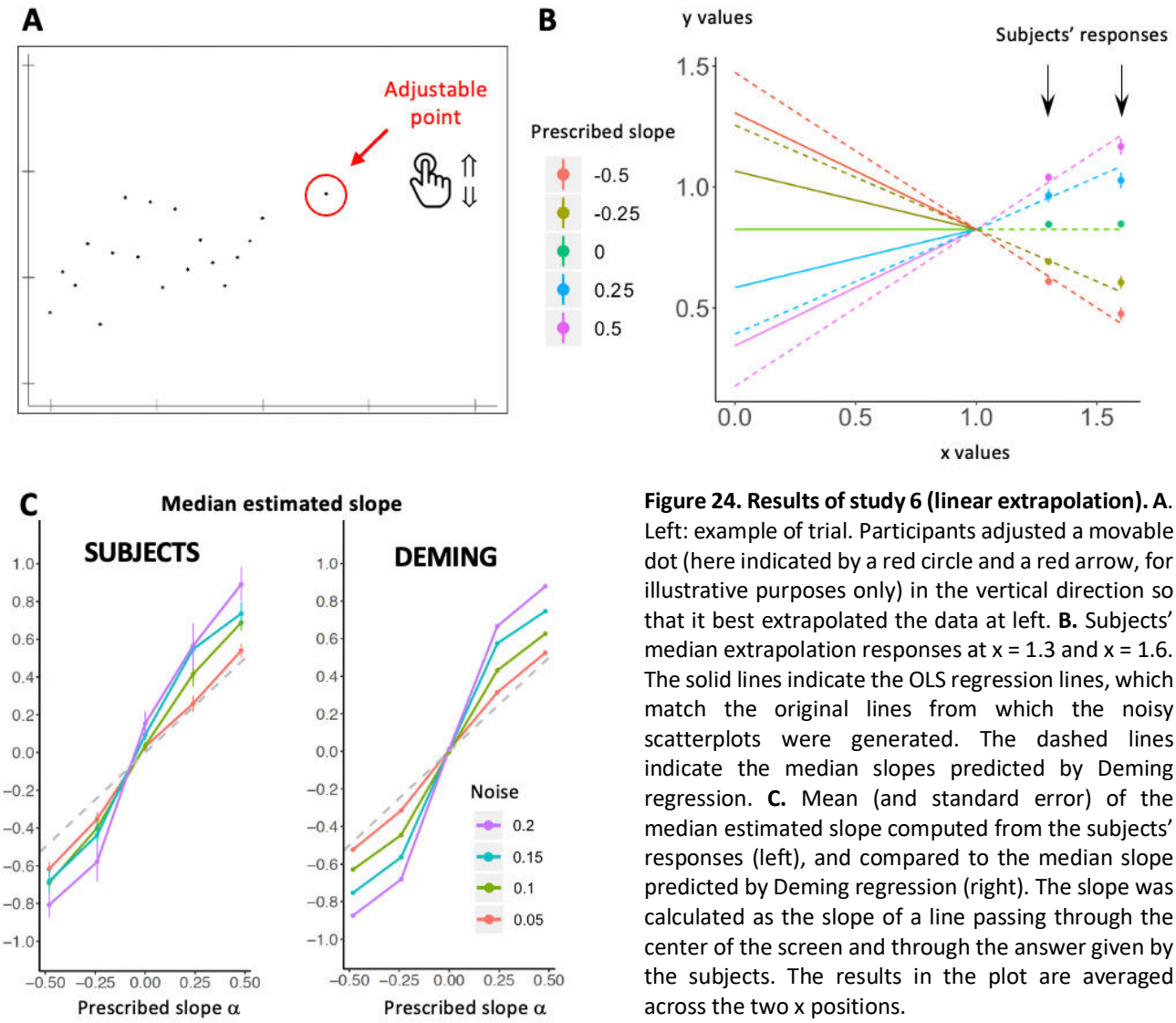


Figure 24. Results of study 6 (linear extrapolation). A. Left: example of trial. Participants adjusted a movable dot (here indicated by a red circle and a red arrow, for illustrative purposes only) in the vertical direction so that it best extrapolated the data at left. **B.** Subjects' median extrapolation responses at x = 1.3 and x = 1.6. The solid lines indicate the OLS regression lines, which match the original lines from which the noisy scatterplots were generated. The dashed lines indicate the median slopes predicted by Deming regression. **C.** Mean (and standard error) of the median estimated slope computed from the subjects' responses (left), and compared to the median slope predicted by Deming regression (right). The slope was calculated as the slope of a line passing through the center of the screen and through the answer given by the subjects. The results in the plot are averaged across the two x positions.

RESULTS

Location of the extrapolated point

For each of the 40 conditions and each participant, we computed the median response location on the y axis. We conducted a repeated measures ANOVA on those median extrapolation values with prescribed slope, noise and probed position as within-participants factors, and we found a significant effect of the slope ($F(4, 36) = 95.15$, partial $\eta^2 = .91$, $p < .0001$); no main effect of the probed position was found ($F(1,9) = 0.01$, partial $\eta^2 = .001$, $p = .92$), although it entered into a significant interaction with the slope ($F(4, 36) = 36.30$, Partial

η^2 = .80, p < .0001). As shown in figure 24B, those findings reflect the fact that participants adapted their extrapolation responses to the prescribed slope and the probed position. However, as in study 5, an interaction of prescribed slope and noise (F(12, 108) = 4.24, partial η^2 = .32, p <.0001), as well as a triple interaction of slope, probed position and noise (F(12, 108) = 2.46, partial η^2 = .21, p <.01) indicated that, as the noise increased, the extrapolation responses became increasingly biased towards exaggerated values, as predicted by Deming regression. Specifically, as in study 5, participants always overestimated positive slopes, and did so with a bias that increases with noise level (ANOVA restricted to positive slopes: main effect of the noise, F[1.68, 14.85] = 6.60, partial η^2 = .42, p =.01). Conversely, participants always underestimated negative slopes, again increasingly so for larger noise levels (ANOVA restricted to negative slopes: main effect of noise level, F[1.81,16.30] = 4.17, partial η^2 = .32, p =.04).

Next, we directly compared participants' extrapolation responses with those expected under Deming regression. The solid lines in figure 24B show the functions from which the scatterplots were generated (i.e., the OLS regressions of the dataset), whereas the dashed lines show the Deming regressions. As already described, Deming regression results in steeper slopes than OLS predictions. Relative to those lines, we can see that the participants' responses lay close to Deming predictions, although at x = 1.6 they tend to be slightly lower. To quantitatively test for the resemblance of participants' extrapolation responses to Deming predictions, we calculated the median slope associated with each extrapolated point (calculated as the slope of the line passing through the center and the given point). Figure 24C shows the remarkable similarity of participants' extrapolations with Deming predictions. We conducted a repeated measures omnibus ANOVA on participants' median slopes and found a significant main effect of the prescribed slope (F[4, 36] = 79.01, partial η^2 = .9, p

<.0001), a significant main effect of noise (F[3, 27] = 3.17, partial η^2 = .26, p =.04) and an interaction effect of the slope with the noise (F[12, 108] = 4.36, partial η^2 = .33, p <0.0001). Although there was no main effect of the probed position (F[1, 9] = 0.5, partial η^2 = .05, p =.5), it entered into a significant interaction with the slope (F[4,36] = 7.82, partial η^2 = .46, p < .0001) and into a triple interaction with slope and noise (F[12, 108] = 3.51, partial η^2 = .28, p < .001), confirming that participants adapted their responses to the prescribed slope and the probed position. These effects reveal that the median slopes associated to the extrapolation responses were increasingly steeper as the noise increased, with a bias, once again, that increased with noise level (ANOVA restricted to positive slopes: main effect of noise, F[1.64, 14.78] = 7.86, partial η^2 = .47, p <.01; ANOVA restricted to negative slopes: main effect of noise, F[1.67,15.07] = 6.22, partial η^2 = .41, p =.01).

DISCUSSION

The results of study 6 showed that participants were able to perform an intuitive extrapolation, meaning they could predict the location of a point outside the range of available data. Unlike the predictions of ordinary least squares, participants' estimates were biased and were affected by noise level. In agreement with the results of experiment 2, their estimates resembled again those predicted by Deming regression. We conclude that participants can approximate a linear regression from a noisy scatterplot, and do so with a comparable performance regardless of the details of the stimuli and the task: binary judgement on a flashed graph (study 2), slope adjustment on a flashed graph (study 5) or extrapolation on a long-exposure graph (study 6). The results of our three experiments converge to suggest that humans behave in a highly competent and consistent way when extracting statistical information from a scatterplot. They take into account not only the slope,

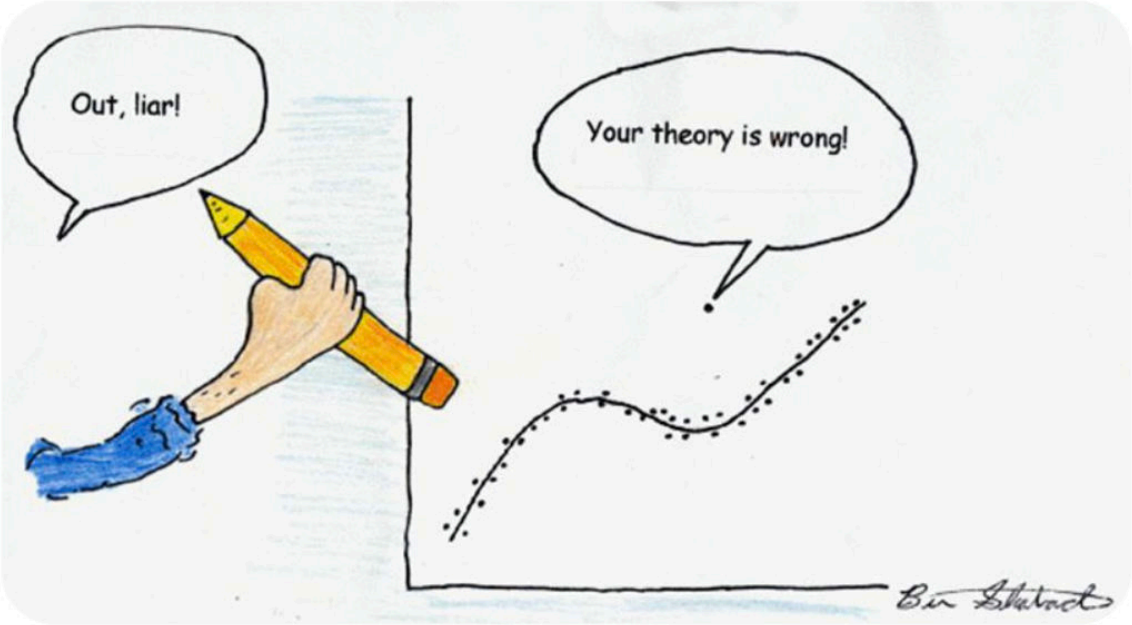
but also the noise level and the number of data points – but do not do so according to classical OLS regression, but to the lesser known Deming regression.

How could we explain that, independently from the response mode (line adjustment and extrapolation) humans use Deming regression rather than ordinary least squares, although the latter procedure would have been more adequate for our datasets, where the y values were noisy measurements of fixed x values? A Deming-like regression may be rational for several reasons. First, Deming is the appropriate procedure when the x measurements are noisy – and participants may not have been aware that x values were fixed. Second, Deming emerges naturally if participants treat the x and y axes symmetrically, which results in minimizing the distance of the points to the fit from both points' coordinates. Indeed, the asymmetry in OLS regression (resulting in a different regression line when y is regressed on x or vice-versa) is highly counter-intuitive. Our results suggest that humans spontaneously perceive the principal axis of the graph, which is defined as the straight line that minimizes the sum of the squared Euclidean distances to the set of points, and hence corresponds exactly to the Deming regression line. Interestingly, considerable research indicates that, during both the perception of objects and of the geometry of the environment, humans and other animals spontaneously extract the principal axis of simple shapes and use it in various computations including object perception (Cohen & Singh, 2006), object misperception in patients with hemineglect (Driver et al., 1994), object manipulation (Turvey et al., 1992), grasping (Cuijpers et al., 2004), visual search (Boutsen & Marendaz, 2001), or spatial reorientation (Bodily et al., 2011, 2018; Cheng, 2005). It therefore makes sense that, when confronted with the task of perceiving the direction of a scatterplot, human adults would spontaneously reuse this evolutionarily ancient ability to grasp an object's principal axis. According to this hypothesis, graph perception would constitute a novel instance of "neuronal

recycling" (Dehaene, 2005; Dehaene & Cohen, 2007), i.e. the repurposing of pre-existing and evolutionary older cognitive mechanisms, initially devoted to other purposes, for novel cultural uses. Just like the invention of writing repurposes part of the ventral visual system for object recognition towards the fast recognition of letter strings (Dehaene, 2009), the cultural invention of the scatterplot could be viewed as a clever way, starting from a large set of numerical data points, to display them in 2D or 3D space such that the resulting graph benefits from the human visual system's sophisticated parallel processing ability, resulting in an immediate extraction of its principal axis.

CHAPTER 3

ROBUST MENTAL REGRESSION: HUMAN RESISTANCE TO OUTLIERS



In the next study (study 7) we further probed human ability to perform mental regression by asking the following research question: are human intuitive statistical estimations robust to the presence of outliers? We asked participants to perform trend judgments and line adjustments in the absence and in the presence of outliers and we manipulated the level of attention towards such outliers: we either did not inform them about their presence (experiment 1); inform them that some outliers might be present and invite them to discard any outlier (experiment 2); inform them about their presence and, before adjusting the robust regression line over the scatterplot, invite them to actively detect their presence or absence (experiment 3).

STUDY 7: OUTLIER DETECTION AND REJECTION

Every scientist regularly deals with outliers, i.e. anomalous observations or measurements that appear very different from the others. The dilemma is always the same: should we consider them the result of normal variability ("noise") inherent in the data, or exclude them from the main analysis, because they “arise suspicions that they were generated by a different mechanism” (Hawkins, 1980)? The answer is never straightforward and often depends on the data format, the scientific field, the number of observations and many other factors; as a consequence, several methods for outlier detection exist, with their advantages and disadvantages (Smiti, 2020). They include distribution-based methods (defining outliers as a function of their variation from a standard distribution), distance-based methods (which compute the distances among all items in the dataset, and consider as outliers those items that do not have close neighbors), and density and cluster-based approaches (which define outliers on the basis of their local density and their belonging to a distinct data cluster). Crucially, all of these methods depend on a threshold, a point beyond which an outlier is considered as such – and once a threshold has been fixed, they are only meant to detect outliers and do not provide explicit guidance on their inclusion or rejection from further analysis. A notable exception is represented by Bayesian approaches, which will be discussed later in the paper. Interestingly, different graphical adaptations have also been proposed to facilitate the perceptual identification of outliers in graphs by human readers, through data visualization tools such as modifying the size, color and opacity of different data points (Micallef et al., 2017).

As discussed in the introduction, one of the most intuitive (but still efficient) techniques to detect outliers is to plot all observations in a bivariate visual format, the scatterplot (Friendly & Denis, 2005), and to let a human viewer decide on the presence of outliers. Indeed,

researchers in psychology consider scatterplots their elective tool in outlier detection (Orr et al., 1991). The other alternatives I have reviewed, such as box plots, hide the complexity of a dataset and may ultimately favor misleading conclusions about the data (Godau et al., 2016; Pastore et al., 2017). For example, scatterplots can be used to detect and reject response times that are either too fast or too slow relative to the average value. They can also be useful to detect the existence of a secondary pattern of data that should be analyzed separately or in interaction (see Sunday et al., 2019 for an example).

These studies, however, raise an important and understudied question: are humans really capable of spotting outliers when a large dataset is displayed as a scatterplot? Our findings on "intuitive statistics" (studies 2, 3, and 4) but also other studies indicate that human adults are remarkably accurate at performing several different statistical tasks on scatterplots. Crucially, the stimuli in the studies described so far were always graphs without outliers, whose data points were normally distributed around the regression line. Only a few studies specifically investigated the role of outliers in graph-based tasks. One found that human adults fail to fully reject outliers when asked to determine the Pearson r of the dataset (Bobko & Karren, 1979; Meyer et al., 1997). Similarly, correlation estimations are affected by outliers independently of the participants' statistical knowledge (Meyer & Shinar, 1992). Even when asked to adjust the trend on a scatterplot including outliers that were either extremely far or located at the boundaries of the main dataset, participants performed a linear regression that fell in-between a robust one (that excludes those outliers) and the line predicted by an ordinary least squares (OLS) algorithm (Correll & Heer, 2017; Liu et al., 2021). Taken together, these results suggest that, in the presence of clear and extreme outliers, participants are affected by them in their correlation judgments and regression estimations, although they

assign them a lower weight than that of other observations in the dataset. In other words, participants attempt to reject outliers, but are not completely successful in doing so.

Another line of research on so-called “ensemble perception", the human ability to automatically encode summary statistics of the visual environment (for a review: Whitney & Yamanashi Leib, 2018), found that, with stimuli other than graphs, humans do the exact opposite: they discard from their judgments all the items that considerably deviate from the other elements in the set (Epstein et al., 2020; Haberman & Whitney, 2010). This automatic filtering of outliers might indeed be highly beneficial in real-life contexts: avoiding deviant observations while focusing on the most representative information, allows to overcome our attentional limitations and to enhance our visual cognition (Alvarez, 2011).

As suggested by the findings from the line adjustment and the extrapolation task presented in studies 6 and 7, perceiving noisy graphical representations such as scatterplots might thus be a novel instance of ensemble perception (since humans manage to quickly and accurately extract a statistical trend from noise). At the same time, however, it does not seem robust to the presence of outliers, contrary to what the literature on ensemble coding would predict. Unfortunately, all past experimental investigations on outlier processing in graphs do not resolve this discrepancy. Indeed, previous studies share two fundamental limitations. First, they allowed participants to slowly inspect the scatterplot before providing any correlation judgment. Second, they always used outliers that diverged dramatically from the main distribution or that were located exclusively at its boundaries, without experimentally manipulating the strength of the outliers in terms of both their distance and their number. These experimental choices can surely be praised for their resemblance to ecological real-life situations: researchers usually take their time to inspect a graph and they often tend to reject only extreme outliers (Anscombe, 1960). However, they do not allow to characterize humans’

spontaneous processing of outliers and their role in affecting intuitive statistics. Furthermore, they do not clearly separate outlier detection from outlier rejection, two processes that we suggest should be carefully distinguished – indeed, the above results suggest, but do not prove, that humans may detect the presence of outliers, and yet continue to be dragged towards them in their mental regression evaluations.

In this study, we aimed to provide an in-depth psychophysical investigation of the perceptual processing of outliers in scatterplots. We tried to answer five open questions:

1) Do subjects spontaneously reject outliers when asked to perform a trend judgment or a regression estimation on a graph, without being told that there might be outliers? The aforementioned studies on ensemble perception (Whitney & Yamanashi Leib, 2018) found that outlier facial expressions (Haberman & Whitney, 2010) and oriented lines (Epstein et al., 2020) are spontaneously excluded when participants are asked to evaluate the average value of a set. We tested whether those findings extend to trend judgments and line fitting on scatterplots or whether, in this case, outlier items are automatically included.
2) Do the number of outliers and their distance from the main dataset modulate the bias that participants exhibit in estimating the slope or in judging the direction of the data's linear trend? Previous research (Bobko & Karren, 1979; Correll & Heer, 2017; Meyer et al., 1997; Meyer & Shinar, 1992) showed that correlation judgments and regression estimates are not robust to the presence of outliers, but this result could vary with the number and distance of the outliers. We thus measured if human performance in intuitive statistics is parametrically affected by those factors.
3) If outliers do bias participants' performance, is this bias modulated by the level of attention towards them? Across our three studies, we varied the level of attention to

outliers by either not providing any information about their presence (experiment 1: no attention); telling participants about their presence and inviting them to discard them in their judgments (experiment 2: medium attention); or explicitly asking them to detect the presence of any outlier, on every trial, before estimating the line through the remaining data points (experiment 3: high attention). There is very little prior research on this topic. Attention towards deviant stimuli has been shown to bias ensemble average estimations in the direction of the deviant item, but participants were never asked to discard outliers (de Fockert & Marchant, 2008). Our manipulation of participants' attention towards outliers thus provides a first test of the role of attention in outlier rejection.

4) How does outlier detection work? In experiment 3, we asked participants to detect as fast as possible the presence (or absence) of any outlier in the dataset. In this manner, we could directly investigate the variables that affect outlier detection and, ultimately, to propose a model of how humans decide whether a given data point is an outlier or not.

5) If outliers are correctly detected, does this mean that they can also be rejected? In experiment 3, we tried to disentangle outlier detection and rejection. On every trial, participants performed a task of outlier detection followed by slope estimation, thus allowing us to examine the contingencies between them. Participants might be well aware of the presence of outliers and the need to discard them, but still fail at doing so, thus suggesting that perceptually rejecting outliers is an ability impenetrable to cognition, as is the case for many visual phenomena (Stokes, 2013).

METHODS

Stimuli

All stimuli included two unlabeled lines denoting the x and y axes, which remained on screen for the duration of the experiment (figure 25). Each line was marked with three small ticks at locations corresponding to the values 0, 0.5, and 1 (those numbers were arbitrary and not shown to participants). Within the area comprised by those two axes, the stimuli were scatterplots comprising 18 white dots on a black background. The x coordinates of the 18 points were fixed and separated by an equal distance on the x axis. Each stimulus was the graphical representation of a dataset generated on the basis of three experimental factors, whose values were combined in a full factorial design. First, we varied the slope of the line (the "main slope") around which the main datapoints (except outliers) were located; the main slope could take value: -0.5, -0.25, +0.25, or +0.5. Second, we independently varied the slope of the line around which the outliers were located; this "outliers' slope" could take value: -0.5, -0.25, +0.25, or +0.5. Third, we varied the number of outliers (n = 0, 1, 2, 3 or 4). In detail, the stimulus generation algorithm worked as follows. First, the y coordinates of all points were determined according to the following equation: $y_i = main\_slope^* x_i + \varepsilon_i$, where the x_i are 18 numbers equally spaced between 0 and 1, and the ε_i are random numbers independently drawn from a normal distribution centered on zero and with standard deviation of 0.1. Afterwards, the desired number of outlier points (0, 1, 2, 3 or 4) were selected at random among all points in the dataset, excepting the six central ones, and their y coordinates were changed according to the following equation: $y_i = outliers\_slope^* x_i + \varepsilon_i$, again with $\varepsilon_i \in N(0,0.1)$. Because of the added noise, the OLS regression slope of the non-outliers dots could depart slightly from the prescribed one ("main slope"). To compensate for this, a small linear component was added to the main datapoints, calculated such that their

final slope corresponded precisely to the prescribed one, and always passed through the center of the screen (see figure 25B for an example with a main slope of 0.5, an outliers' slope of -0.5 and 4 outliers. Examples of stimuli from all experimental conditions are provided in appendix B). When the main slope was identical to the outlier slope, all data points were generated around a single slope, thus resulting in no outlier being presented (and such condition was considered equivalent to the one with 0 prescribed outliers). We generated outliers using a secondary process (namely another regression line with the outlier slope) because it offered a means to finely control their average distance from the main dataset, while still avoiding to impose an exact location to them. A different choice would have been to manipulate the distance factor by using different standard deviation distances from the main regression line but, in this case, all outliers would have had, for a given distance condition, *exactly* the same deviance, likely making the stimuli easily recognizable over trials.

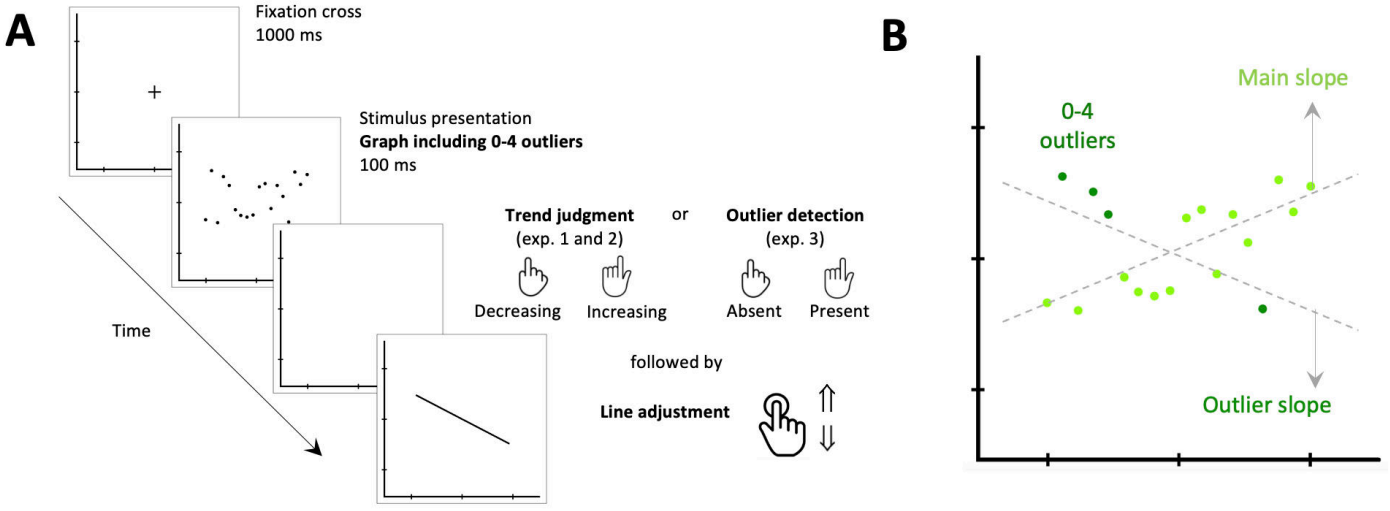


Figure 25. Experimental design. A, example trial. On each trial, participants were presented with a scatterplot and asked to judge, in experiment 1 and 2, if its trend was ascending or descending; or, in experiment 3, if there were any outliers. Immediately after their response, they had to adjust the slope of a line on screen by moving their finger on a trackpad, in order to provide an estimation of the regression line underlying the noisy scatterplot. In experiment 1, participants were not informed of the presence of outliers. Experiment 2 differed from experiment 1 only in that participants were informed that some outliers could be present, and were asked to try to ignore them in their judgments. Experiment 3 further emphasized outliers by first asking for explicit outlier detection before the slope adjustment task. **B**, Illustration of the stimulus generation process. In each scatterplot, the majority of dots were noisy samples around a line with a main slope of either 0.5, 0.25, -0.25, or -0.5. Between 0 and 4 dots were outliers (in this example, 4) generated as noisy samples around another line, whose slope could also take the values 0.5, 0.25, -0.25 or -0.5. Different colors are used for illustrations purposes only: outliers were not signaled in any way, since all stimuli were white dots on a black background.

Participants

30 participants (10 per experiment) were recruited (age: 26.2 $\pm$ 2.1, 16 females, 14 males). The sample size was the same as in our previous studies. All participants had normal or corrected to normal vision, no medical history of epilepsy, were right-handed, and did not take psychoactive drugs. The experiment was advertised through the mailing list of the first author's university. In order to ensure the homogeneity of the sample in terms of participants' cultural background, only participants with at least a master's degree were recruited. They all signed an informed consent and were paid 5 euros for their participation. The experimental sessions lasted approximately 30 minutes and were approved by the local ethical committee. One participant was excluded from study 7b analyses since he failed to perform the task appropriately (his performance was at chance level).

Experimental procedure

Participants were invited to sit on a fixed chair with their head at a distance of 50 cm from the screen. Each experimental session was divided into 5 blocks of 80 trials; the duration of each block was ~4 minutes. After each block, participants could take a short break. Before starting the actual experiment, 25 practice trials were run under the researcher's supervision, in order to control for the correct execution of the task (i.e., maintaining the correct distance from the screen, correctly placing their hand and fingers; familiarizing with the rapid presentation of the stimuli. No feedback on performance was provided). On each trial, as illustrated in figure 25A, a fixation cross first appeared for 1000 ms, immediately followed by a scatterplot flashed for 100 ms. The stimuli were flashed in order to promote spontaneous and fast responses and thus to avoid any possible explicit strategy, calculation, or complex eye movement patterns. The experimental procedure varied depending on the experiment.

Experiment 1. No information concerning the presence of outliers was given to participants. They were merely asked to respond (as fast and accurately as possible) by pressing with their left-hand ring finger on a key (signaled with a ⇓ sticker) if they thought that the trend in the scatterplot was decreasing or, conversely, to press with their left-hand index finger on another key (signaled with a ⇑ sticker) if they thought that the trend in the scatterplot was increasing. Immediately after this first response, an adjustable line appeared in the middle of the screen. The line was initially horizontal, but participants were asked to adjust it as accurately as possible by moving their right-hand index finger on the computer trackpad. The center of the line was kept fixed, so that moving the finger up or down the trackpad resulted in a rotation of the line around its center, whose angle was proportional to the finger displacement; moving the finger up tilted the line in the counterclockwise direction, whereas moving the finger down tilted it in the clockwise direction. For this second task, participants were invited to respond independently of their first trend judgment: they were explicitly told that they could orient the line in a direction opposite to their trend judgment, if they thought that they had made a mistake in the first task. When the adjustment was completed, they pressed the trackpad in order to confirm their answer and move to the next trial, which was preceded by a 1s fixation cross.

Experiment 2. Participants of this experiment were asked to perform the exact same task as participants in experiment 1. The only difference consisted in the information given to them before starting the experimental session: they were informed that 1 or more outliers, defined as points outside the main dataset, could be present in some trials. They were invited to try to exclude such outliers from their answers, and thus to perform both tasks of trend judgment and slope adjustment only on the main dataset.

Experiment 3. As in experiment 2, participants were informed that one or more outliers could be present in some trials. They were asked to detect them, as fast and accurately as possible, by pressing with their left-hand ring finger on a key (signaled with a "NO" sticker) if they thought that the scatterplot did not include any outliers or, conversely, to press with their left-hand index finger on another key (signaled with a "YES" sticker) if they thought that the scatterplot included one or more outliers. Immediately after this detection response, they moved to the slope adjustment task, identical to experiment 2, with the explicit instruction to try to estimate the slope of the main dataset only and, thus, to reject outliers.

RESULTS

Performance in trend judgment and line adjustment in graphs without outliers

First, in an attempt to replicate our previous studies, we analyzed participants' trend judgment performance in the absence of outliers. Figure 26A shows the percentage of trials classified as "increasing" as a function of the main slope and of the t-value associated with the scatterplot linear regression. As clear from the figure, in both experiments 1 and 2, participants' responses could be modeled as a sigmoid function of such t-value. Both the sigmoidal shape of their response rates (figure 26A) and the distance effect in their response times (i.e., slower responses for stimuli with a t-value closer to zero; figure 26B) could be jointly predicted by a classical decision-making model which assumes a noisy accumulation of evidence towards a decision bound (Gold & Shadlen, 2002). In figure 26B, the blue lines show the performance predicted by that model. These results replicate the findings described in the previous chapters.

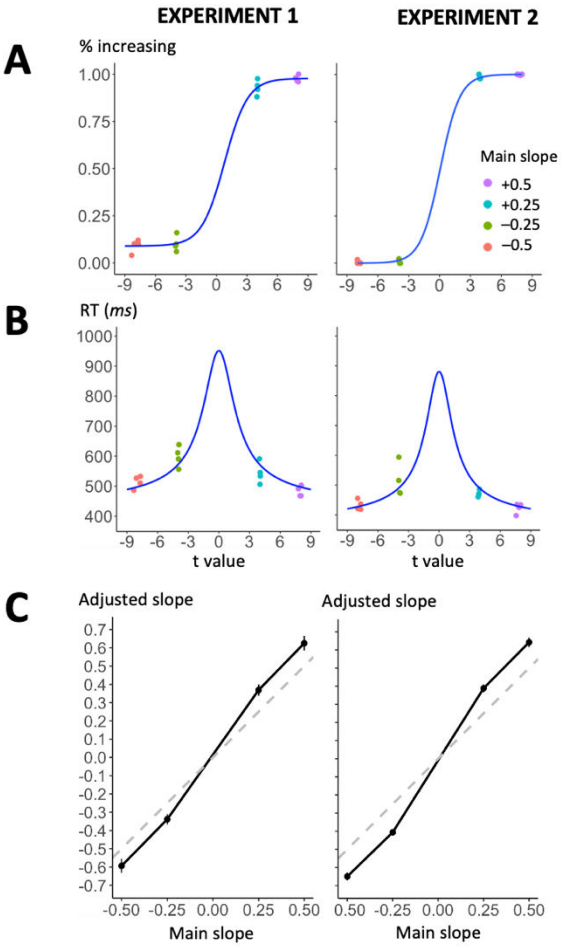


Figure 26. Performance in trend judgment (A, B) and line adjustment (C) in graphs without outliers. In both experiments, the percentage of "increasing" responses (**A**) and the response times (**B**) vary systematically with the t-value associated to the Pearson coefficient of correlation. The blue lines in the middle row indicate the response times predicted by a simple accumulation-of-evidence model (Gold and Shadlen, 2002). The plots in **C** show the slopes reported by participants (black lines) and predicted by ordinary least squares (OLS) regression (dashed grey lines), which corresponds to the process by which the scatterplots were generated. Participants responded with slopes exceeding those predicted by OLS, in agreement with the use of Deming regression. All of these results replicate our previous findings with similar mental regression tasks (Ciccione & Dehaene, 2021).

Influence of outliers on accuracy in the trend judgment task

We then looked at participants' performance in the same trend judgment task ("ascending or descending?") when outliers were present in the stimuli. Figure 27 shows the results from both experiment 1 and 2, which closely resembled each other. The average error rates for stimuli without outliers are simply indicated as a reference (the black dots). The top row indicates the error rate as a function of the number of outliers as well as two other driving variables: the absolute value of the main slope of the scatterplot (steep: 0.5; or shallow: 0.25), and the outliers' distance, quantified as the absolute difference between the outliers' slope and the main slope. For a main slope of 0.5, the available outliers' distances were 1, 0.75 and 0.25, which for simplicity are referred to, respectively, as "large", "medium" and "small".

Similarly, for a main slope of 0.25, the available outliers' distances were 0.75, 0.5 and 0.25, which are again referred to with the same labels.

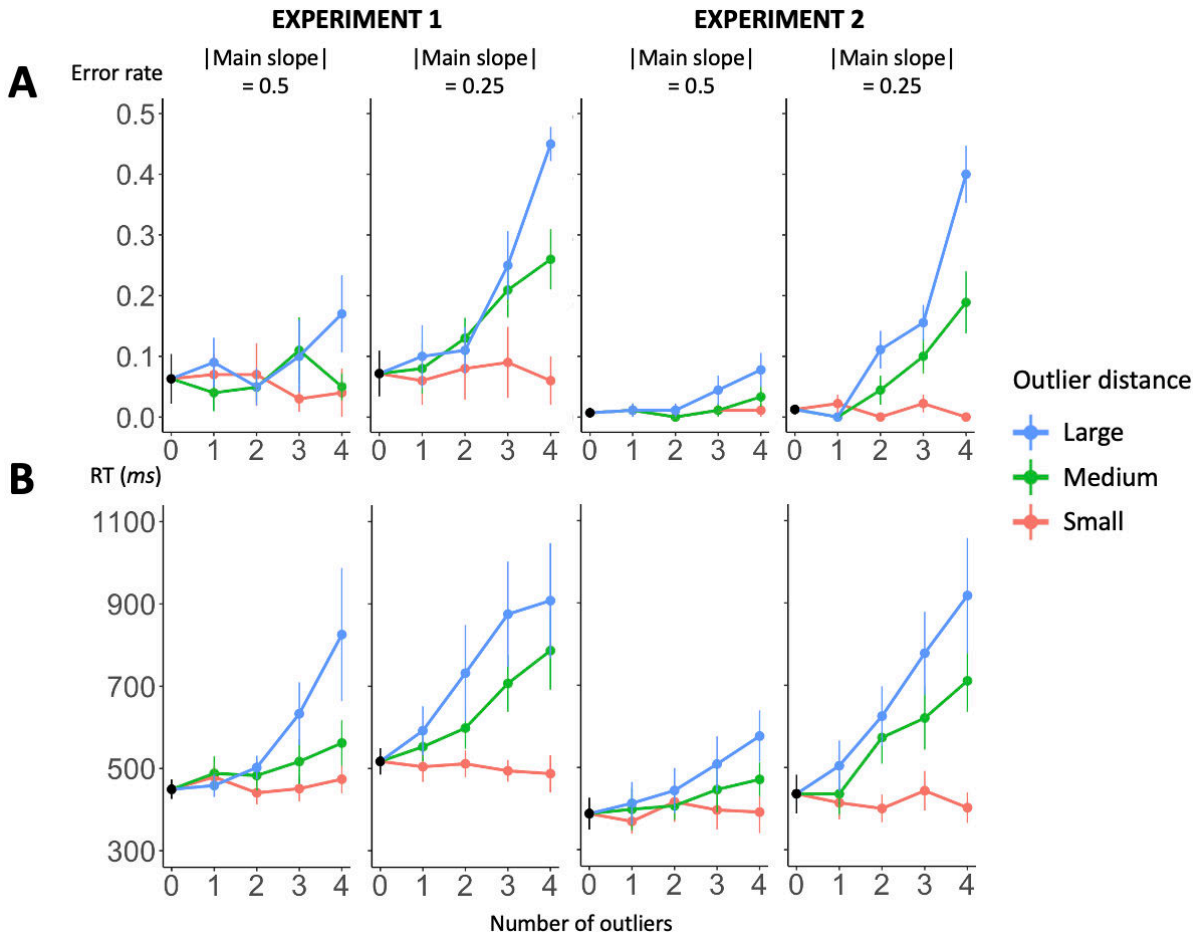


Figure 27. Influence of outliers on performance in the trend judgment task ("is the graph ascending or descending?"). Results are plotted as a function of the number of outliers, separately for graphs with steep (0.5) or shallow (0.25) main slopes. Both error rates (**A**) and response times (**B**) increase as a function of the number of outliers, as well as of the distance of the outlier slope from the main slope. When the graph has a shallower main slope (0.25), thus rendering the task more difficult, the influence of outliers becomes correspondingly larger. Error bars indicate one standard error of the mean across subjects.

As we can see, in both experiments, the error rate increased as a function of the number of outliers and their distance from the main dataset, indicating that the participants' responses were attracted towards the outliers. To test for the significance of these observations, we conducted an ANOVA on participants' error rates with the main slope, the outliers' distance

and the number of outliers as within-subjects' factors and the experiment number (1 or 2) as a between-subjects' factor. While the experiment number had no significant main effect nor any interaction (all related p values > .1), all other factors had main effects (*main slope*: F[1, 17] = 101.15, partial η^2 = .86, p <.001; *outliers' distance*: F[1.5, 25.8] = 90.97, partial η^2 = .84, p <.001; *number of outliers*: F[2.5, 42.9] = 39.55, partial η^2 = .7, p <.001) and interaction effects (*main slope and outliers' distance*: F[1.6, 27.4] = 18.78, partial η^2 = .52, p <.001; *main slope and number of outliers*: F[2.3, 39] = 25.6, partial η^2 = .6, p <.001; *outliers' distance and number of outliers*: F[3.2, 54.1] = 19.12, partial η^2 = .53, p <.001; triple interaction of *main slope, outliers' distance and number of outliers*: F[3.9, 66.9] = 6.54, partial η^2 = .28, p <.001).

Figure 27 clarifies the meaning of those interactions. First, error rates generally increase with the number of outliers, but more so when the main slope is shallow (0.25), thus rendering the main decision more difficult, than when the main slope is steep (0.5). Second, similarly, for the same number of outliers, their impact is larger when their distance to the main dataset is larger, i.e., when they deviate more from the main regression line. Those findings make sense: essentially, the more numerous the outliers, and the more they push towards a line with a different orientation from the main one, the more likely participants are to make an error. Indeed, it is worth noting that the small outliers' distance condition (red lines), with a main slope of 0.25, was the only experimental condition in which the outliers' slope was steeper than the main slope: in this situation, outliers were not expected to make the trend judgment harder to perform, since they made the overall trend of the graph steeper. Indeed, the error rates in these conditions did not increase as a function of the number of outliers (the red lines are essentially flat): this was confirmed by a non-significant main effect of the number of outliers in two ANOVAs restricted to those conditions (*experiment 1*: F[1.1, 9.7] = .28, partial η^2 = .03, p = .62; *experiment 2*: F[3, 24] = 1.73, partial η^2 = .18, p = .19).

Influence of outliers on response times in the trend judgment task

Response times in the trend judgment task for experiment 1 and 2 are plotted in figure 27 (bottom). Response times behaved in parallel to error rates, thus indicating the absence of a speed/accuracy tradeoff. They increased as a function of the number of outliers as well as of the distance of the outliers' slope from the main slope. The same ANOVA as above, now on median response times, again revealed no main effect or interactions involving the experiment factor (all related p values > .1). It also indicated that all within-subject factors had a significant main effect (*main slope*: F[1, 17] = 53.67, partial η^2 = .76, p <.001; *outliers' distance*: F[1.1, 18] = 26.88, partial η^2 = .61, p <.001; *number of outliers*: F[1.6, 26.8] = 20.02, partial η^2 = .54, p <.001) and entered into significant interactions (*main slope and outliers' distance*: F[1.4, 23.3] = 16.87, partial η^2 = .5, p <.001; *main slope and number of outliers*: F[2.4, 40.3] = 6.45, partial η^2 = .28, p <.01; *outliers' distance and number of outliers*: F[2.4, 39.9] = 15.04, partial η^2 = .47, p <.001; no triple interaction of the within-subjects factors was found). Again, those interaction effects are easily observable in figure 27: response times increased significantly faster with the number of outliers as the distance of the outliers increases, and also as the main slope gets shallower. Like for error rates, the experimental condition in which the outliers' slope was steeper than the main one resulted in no increase of response times (as evident from the essentially flat red lines in the plots for a main slope of 0.25), which was confirmed by two ANOVAs restricted to those conditions (*experiment 1*: F[2.6, 23.4] = .7, partial η^2 = .07, p = .54; *experiment 2*: F[2, 15.7] = .37, partial η^2 = .04, p = .7).

Lastly, as we can see from the response time plots for a main slope of 0.25, we found that the presence of a single outlier, at a large enough distance from the main dataset (blue lines), induced a substantial increase in response times. Indeed, a paired t-test on participants' response times from both experiments revealed a significantly slower median response time

in the presence of one outlier than in the absence of outliers ($t(18) = 2.78$, $p < .01$; respectively 550 versus 478 ms).

One could argue that a greater number of outliers simply made the overall slope of the dataset closer to zero, thus making trend judgement more difficult. Could participants' slower response times be explained by changes in slope rather than by the number of outliers? To test for this, we performed a multiple linear regression on response times with both the number of outliers and the absolute Deming slope as predictors. We found that both were significant ($\beta_{\text{number of outliers}}$= 20.7 ms/outlier, $p < .0001$; $\beta_{\text{absolute Deming slope}}$ = -822.5, $p < .0001$). We also computed a linear regression on the residuals of the response times as a function of the absolute Deming slope and found that the number of outliers was still a significant predictor (β= 19.2 ms/outlier, $p < .0001$). Thus, the results confirm that outliers influenced response times over and above their indirect effect on the overall trend, with a cost of ~20 ms per outlier.

Overall, performance in the trend judgment task indicated that, regardless of the instructions to exclude outliers, participants were always strongly influenced by them, especially when (1) they were more numerous; (2) their deviation was large; and (3) the decision was difficult because the main slope was shallow.

Influence of outliers on the line adjustment task

The second task, which was run in experiments 1, 2 and 3, consisted in a slope adjustment: participants were asked to adjust the line in order to best fit the scatterplot. As explained in the methods section, the three experiments differed only in terms of the induced level of attention about the presence of outliers in the stimuli: in experiment 1, no information about outliers was given; in experiment 2, participants were invited to exclude them in both their trend judgment and slope adjustment; in experiment 3, they were explicitly invited to

concentrate on them and detect their presence (or absence) before performing the slope adjustment task (after rejecting them).

We first examined the slope estimated by participants in the absence of outliers (figure 26C shows the results for experiment 1 and 2, which were similar to experiment 3). Confirming our previous studies, we found that the estimated slope closely tracked the actual slopes of the graphs, but were steeper than the ones predicted by a classic ordinary least squares (OLS) regression (the grey dashed lines in figure 26C). Their values were compatible with the minimization of the orthogonal distance of the points to the best-fitting line, a procedure known as Deming regression.

For each subject and each experimental condition, we then evaluated the impact of outliers relative to this no-outlier baseline. To this aim, we calculated "response bias" as the difference between the median slope that they reported in the presence of outliers and in their absence. For visualization and analysis' purposes, the sign of this difference was flipped such that a positive value always indicated attraction towards the outliers (in practice, this meant that we flipped the sign for all stimuli with an outliers' slope lower than the main slope). Figure 28 shows the mean response bias as a function of experiment, main slope, number of outliers, and outliers' distance from the main dataset. We can see that the outlier-induced bias increased with the number of outliers, but did so faster for a large outliers' distance, and more so in experiment 1 than in experiment 2 or, a fortiori, experiment 3. We confirmed these observations through a repeated measures ANOVA on participants' median bias with experiment number as between-subjects factor and main slope, number of outliers, and outliers' distance as within-subjects factors. All of the latter had a significant main effect (*main slope*: $F[1, 26] = 43.35$, partial $\eta^2 = .63$, $p < .001$; *number of outliers*: $F[1.57, 40.94] = 82.72$, partial $\eta^2 = .76$, $p < .001$; *outliers' distance*: $F[1.48, 38.53] = 22.08$, partial $\eta^2 = .46$, $p <$

.001). Although the main effect of *experiment* was close to significance (F[2, 26] = 3.20, partial η^2 = .2, p = .06), it entered in a significant interaction with both the *main slope* (F[2, 26] = 4.99, partial η^2 = .28, p = .01) and the *outliers' distance* (F[2.96, 38.53] = 5.76, partial η^2 = .31, p < .01).

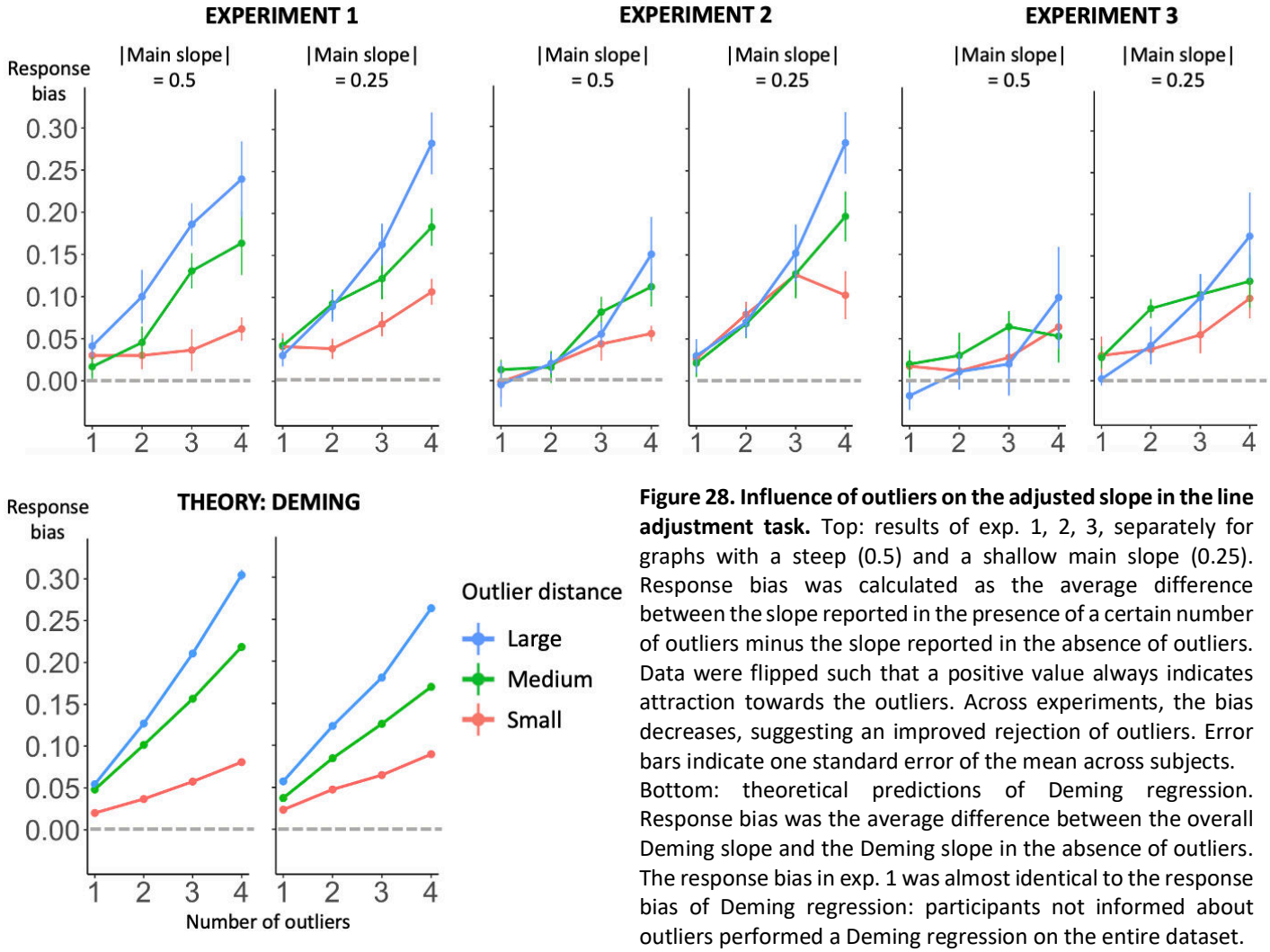


Figure 28. Influence of outliers on the adjusted slope in the line adjustment task. Top: results of exp. 1, 2, 3, separately for graphs with a steep (0.5) and a shallow main slope (0.25). Response bias was calculated as the average difference between the slope reported in the presence of a certain number of outliers minus the slope reported in the absence of outliers. Data were flipped such that a positive value always indicates attraction towards the outliers. Across experiments, the bias decreases, suggesting an improved rejection of outliers. Error bars indicate one standard error of the mean across subjects. Bottom: theoretical predictions of Deming regression. Response bias was the average difference between the overall Deming slope and the Deming slope in the absence of outliers. The response bias in exp. 1 was almost identical to the response bias of Deming regression: participants not informed about outliers performed a Deming regression on the entire dataset.

Indeed, as we can see from figure 28, the outlier-induced bias decreased across experiments, as the level of attention to outliers increased, and this effect was more pronounced for a larger number of outliers and for larger outliers' distances. It is worth noting that the *number of outliers* had also a significant interaction with both the *main slope* (F[2.17, 56.35] = 3.9, partial η^2 = .13, p = .02) and the *outliers' distance* (F[4.15, 107.81] = 12.36, partial η^2 = .32, p

< .001). Thus, the results of the line adjustment task (figure 28) closely paralleled those of the trend judgment task (figure 27).

If, as we suggest, uninformed participants did not spontaneously reject outliers, but included them in their regression estimates, then their response bias should be predictable by a global regression performed on the entire dataset. To test this idea, we examined whether the response bias from participants of experiment 1 (i.e., those who received no information about the presence of outliers) mirrored the theoretical predictions of Deming regression. As with the actual data, we first computed the response bias as the difference between the slope predicted when the regression was applied to the entire dataset, and when it was applied to a dataset without outliers. Figure 28 (bottom) shows the predicted biases for each experimental condition, plotted in the same way as the human data. Those predictions quantitatively match the observed data (linear regression between predicted and observed, $R^2 = 0.91$, slope = 1.02 ± 0.07, intercept = .01). In particular, Deming regression predicts that bias should increase with the number of outliers and with their distance from the main dataset, exactly as in human data.

Performance in outlier detection

On every trial of experiment 3, participants first performed an outlier detection task: immediately after the flashing of the scatterplot, they had to decide whether they had seen at least one outlier or not, by pressing one of two response keys as fast and accurately as possible. This experimental procedure allowed us to directly investigate whether and how humans detect the presence of outliers. Figure 29 shows the percentage of “yes” responses as a function of the main slope, the number of outliers and their distance. The results indicated that false alarms were quite high (40-50% of trials without outliers), but that correct detection increased as a function of the number of outliers, especially for large and medium

outliers' distances. Those observations were confirmed by an ANOVA on the percentage of "yes" responses with the above factors as within-subjects' factors (to obtain a full factorial design, we excluded the conditions with 0 outliers, which are presented in figure 29 only for reference). There was a main effect of both the *number of outliers* ($F[1.90, 17.13] = 11.52$, partial $\eta^2 = .56$, $p < .001$) and their *distance* ($F[1.59, 14.28] = 52.61$, partial $\eta^2 = .85$, $p < .001$). The *main slope* had no main effect ($p = .48$) but entered in a significant interaction with the *outliers' distance* ($F[1.79, 16.13] = 11.32$, partial $\eta^2 = .56$, $p = .001$): in fact, as clear from figure 29, for a steeper main slope of 0.5, the difference in correct detections between the three outliers' distances was more pronounced than for a main slope of 0.25.

We ran a similar ANOVA on participants' median response times for correct detections and found only a significant main effect of *outliers' distance* ($F[1.13, 10.17] = 5.99$, partial $\eta^2 = .4$, $p = .03$) and its interaction with the *main slope* ($F[1.81, 16.25] = 6.09$, partial $\eta^2 = .4$, $p = .01$).

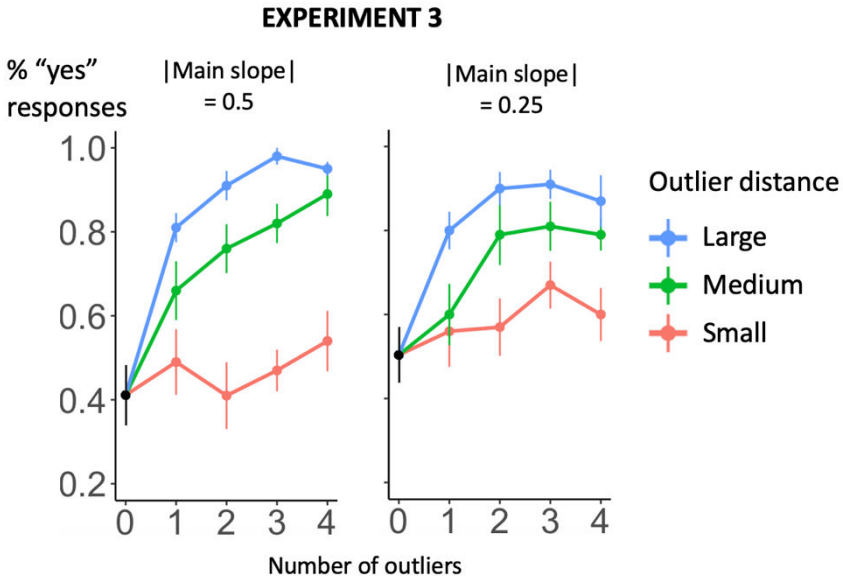


Figure 29. Performance in outlier detection in experiment 3. The percentage of trials in which participants reported seeing at least one outlier is plotted as a function of the true (i.e., "prescribed") number of outliers (0-4). This percentage increases as a function of the number of outliers, as well as their distance between their slope and the main slope. Error bars indicate one standard error of the mean across subjects.

Formulating and testing a theory of outlier detection and rejection

On what basis do participants decide on the presence of outliers? We formulated the hypothesis that, like a statistician, they might base their judgments on an estimate of how much a given data point departs from the rest of the cloud. A simple way of measuring such a departure is to compute a z-score for each point, i.e., a fraction with the numerator equal to the distance of that point to the regression line, and the denominator equal to the standard deviation of such distances. Such a z-score evaluates to what extent the observed data point is out-of-distribution compared to the other ones.

The specific model we propose is shown in figure 30A. First, we computed the Deming regression of each scatterplot and postulated that, for the numerator, participants use the perpendicular distance to that line. Second, for the denominator, since our graphs all had the same noise level (standard deviation = 0.1), we postulated that subjects could pool their noise estimates across trials and eventually converge to a fixed value. Note that this hypothesis may be revised in a different experimental setting – for instance if participants saw a single graph, or if the noise level varied across trials; then their estimate could be based on the observed graph. Here, however, we obtained a better account by postulating a fixed value of the denominator (as confirmed by a model comparison described later in this section).

In the end, we therefore calculated, for each point, a z-score equal to its perpendicular distance to the regression line divided by 0.1 (figure 30A). Our hypothesis predicts that this value is the decision variable on the basis of which participants decide whether that point is an outlier. Since they had to decide whether *any* outlier was present, the percentage of "yes" responses in outlier detection should be a logistic function of the maximum z-score over all 18 data points. Figure 30B shows the corresponding psychophysical curve (for visualization and analysis' purposes, the responses were binned according to the highest z-score). We ran

a multiple logistic regression on all participants' responses with two regressors: the highest z-score and the actual prescribed number of outliers; we found that the former was an excellent predictor of "yes" responses (β = 1.91, p < .0001), better than the actual prescribed number of outliers (β = .1, p < .0001). Indeed, as we can see from figure 30B, when the highest z-score was low (~0.8), the proportion of "yes" responses dropped to 15%, lower than the average rates of false alarms of 48% on trials where prescribed outliers were genuinely absent (figure 29). Conversely, at the opposite extreme, when the highest z-score exceeded about 3, the detection rate was close to 100%, higher than the average values of 65% when a single outlier was actually present (figure 29).

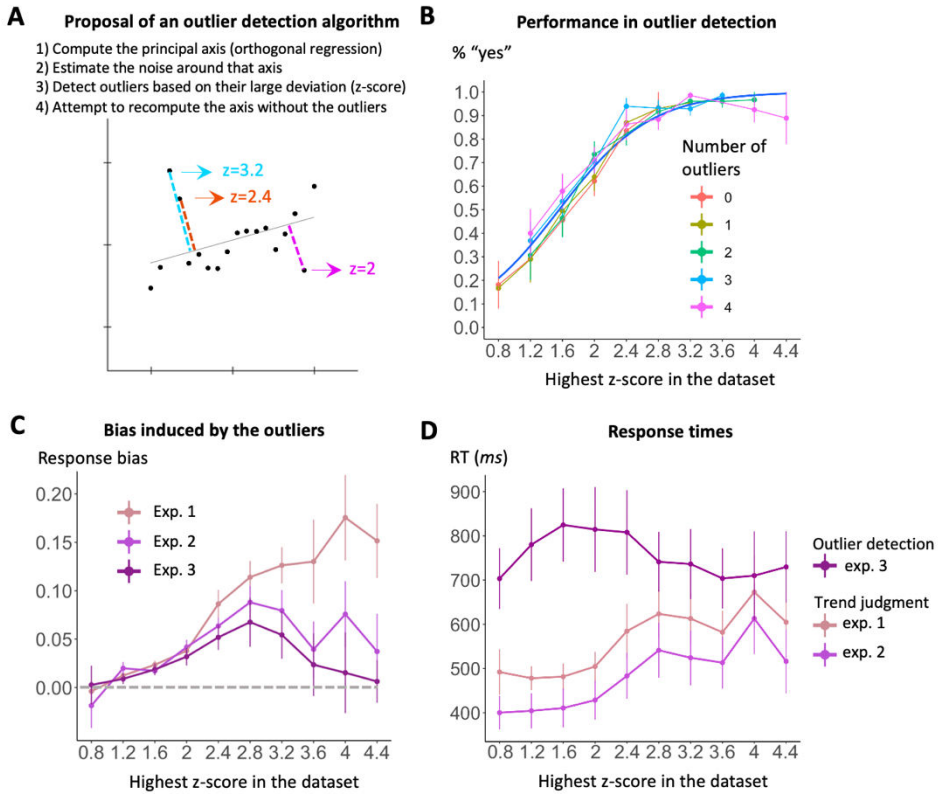


Figure 30. Participants may detect outliers by computing the significance of their deviation from the principal axis. A: example of a scatterplot with three outliers and proposal of an outlier detection algorithm. The outliers can be detected by calculating their individual z-scores, computed as their distance to the Deming regression line (i.e., the principal axis), divided by an estimate of the standard deviation of those distances. Outliers tend to have large z values. **B:** The percentage of trials in which the subjects reported seeing at least one outlier (detection task in experiment 3) is well predicted by the highest z-score in the stimulus graph, regardless of the prescribed number of outliers. Crucially, the highest z-score is a better predictor of outlier detection than the prescribed number of outliers, (shown in figure 5). **C**: the bias in the slope adjustment task also varies as a function of the highest z-score in the dataset. In experiment 1, where participants were not told about outliers, the increase is essentially monotonic. In experiments 2 and 3, the bias starts decreasing when the highest z-score exceeds ~2.8. In experiment 3, the bias returns to zero for larger z-scores, indicating that extreme outliers can be rejected when explicitly instructed. **D**: response times in trend judgment (exp. 1 and 2) increase as a function of the highest z-score in the dataset; response times in outlier detection (exp. 3) peak for highest z-scores around 2, where the model predicts the presence or absence of outliers to be most ambiguous. Error bars indicate one standard error of the mean across subjects.

One could rightfully argue that the highest z-score in the dataset does not take into account the number of other outliers. To focus on the simplest cases, we thus restricted our logistic regression to stimuli with either no prescribed outlier, or with a single prescribed outlier – and in both cases, we found that the highest z-score was still a significant predictor of the percentage of "yes" responses (respectively: $\beta_{0\_outliers} = 1.88$, $p < .0001$; $\beta_{1\_outlier} = 2.22$, $p < .0001$). Figure 30B makes it clear that a single function of the z-score provided an excellent account of the outlier detection responses, regardless of the actual prescribed number of outliers.

We tested several alternative ways of computing the z-scores. First, the distances (numerator) could be computed using the regression of all points (as we did) or the regression restricted to the main dataset. Second, they could be based on the perpendicular distance to the Deming regression, or the vertical distance to the OLS fit. Third, the standard deviation (denominator) could use the prescribed standard deviation of the distances (0.1) or the actual standard deviation, measured from the specific graph. We modeled the logistic regressions of the percentage of "yes" responses as a function of the highest z-score calculated through all eight combinations of those three parameters and found that the model with the significantly smallest Akaike Information Criterion (AIC), thus the one more plausible to be correct (Akaike, 1998), was the above-described model.

Given that the highest z-score accounted well for outlier detection in experiment 3, we next examined whether the same variable also predicted the capacity for outlier rejection, i.e., the influence of outliers on mental regression slopes. To this end, we went back to experiments 1, 2 and 3, and plotted the participants' response bias in the line adjustment task as a function of the highest z-score in the stimulus graph, separately for each experiment (figure 30C). Interestingly, for experiment 1, the response bias increased monotonously as a function of

the highest z-score (R^2 = .36, F[1, 98] = 57.21, p <.001): with no information concerning the presence of outliers, participants included them in their estimations, and the greater their deviance, the higher the bias they induced. However, for experiment 2 and 3, in which participants were explicitly asked to reject outliers, a similar increase in response bias was seen only up to a highest z-score of ~2.8, after which the bias started to decrease. Indeed, in experiment 3, which required an explicit outlier detection on each trial, the bias was statistically indistinguishable from zero for z-scores higher than 3.6 (mean bias = 0.001; t-test on all responses against zero: t(173) = .09, p = .93).

These observations were confirmed by a repeated-measures ANOVA on the outlier-induced bias with experiment (1, 2 or 3) as between-subjects factor and the highest z-score in the dataset as within-subjects factor: both had a significant main effect (*experiment*: F[2, 26] = 4.94, partial η^2 = .28, p = .02; *highest z-score*: F[3.17, 82.53] = 8.56, partial η^2 = .25, p < .0001) and entered into a significant interaction with each other (F[6.35, 82.53] = 2.57, partial η^2 = .17, p = .02). Crucially, the main effect of the experiment and its interaction with the highest z-score vanished when the ANOVA was computed only on stimuli with a highest z score limited to values at or below 2.4 (both p values > .47).

In summary, the data in figure 30 suggests the existence of two ranges. For highest z-scores below roughly 2.4, participants miss many of the outliers, while their influence on regression responses increase with z; and for highest z-scores above that value, outlier detection approaches 100%, and their influence on mental regression starts to decrease – but only if subjects are told to reject them.

This conclusion seems to suggest that, *on average*, outlier rejection closely parallels outlier detection. However, this was not true on a single-trial basis. We restricted the analysis to those trials of experiment 3 in which (a) a single outlier was prescribed; (b) that point had the

highest z-score; and (c) the participant responded that he had detected an outlier (most likely the prescribed one). On such trials, if outlier detection automatically led to outlier rejection, there should be no outlier-induced bias on the participants' slope estimates. This was true for scatterplots with one prescribed outlier with a z-score higher than 2 ($t(84) = -.31$, $p = .62$) but not for scatterplots with one prescribed outlier with a z-score at or below 2: for these stimuli, the bias was still significantly higher than zero ($t(49) = 2.75$, $p < .01$). This finding confirms that participants could remain influenced even by outliers that they have detected.

Lastly, we looked at whether the response times could also be predicted by the z-score of the datapoints (figure 30D). First, we considered the trend judgment task used in experiments 1 and 2, where we previously found that RT increased with the prescribed number of outliers, and examined whether it could be explained by the actual number of outliers. To estimate the latter, we calculated, for each graph, the number of outliers passing a threshold of z>2, and we included it as a predictor in a multiple regression on response times, together with the absolute Deming slope and the absolute main slope of the dataset. All predictors were significant ($\beta_{\text{number of outliers higher than z=2}}$= 25.3 ms/outlier, $p < .0001$; $\beta_{\text{absolute Deming slope}} = -1185.3$, $p < .0001$; $\beta_{\text{main slope}} = 579.2$, $p < .0001$). We then calculated the residuals of the regression with the two mentioned slopes as predictors and computed a linear regression on such residuals as a function of the number of outliers with a z-score higher than 2, finding it was still a significant predictor ($\beta = 13.9$, $p < .01$). Crucially, such a linear regression had an AIC of 112289, which was significantly smaller than the one calculated on the residuals as a function of the prescribed number of outliers (AIC = 112456, $\Delta_{AIC} = 167$, $p < .0001$), suggesting once more that the z-score of the datapoints was a better predictor of participants' performance than the prescribed number of outliers. This is evident when comparing figure 29 to figure 30B: if the prescribed number of outliers is taken into account (figure 29), outliers are wrongly

detected at a very high rate (~40-50% when no outliers were present); however, when the actual distance of those outliers is considered (figure 30B), the false detection rate turns out to be much lower (~20-30% for trials with a low z-score).

Next, we considered the response times in outlier detection (experiment 3). Our model predicts that participants take that decision by evaluating whether any point has a z score above a threshold value, close to z=2. Thus, the decision variable should be the difference between the highest-score and this threshold, and response times should be increasingly slower as this difference approaches zero. To test this prediction, for each graph, we calculated the absolute distance between its highest z-score and 2, and we used such value as a predictor in a linear regression of response times. The effect was significant (β = -70.8, $p < .0001$), and a plot of RTs indicated that indeed, RTs decreased with the distance from the putative decision boundary (figure 30D).

Comparing human performance with an optimal Bayesian model

As explained in the introduction, formal methods of outlier detection share two fundamental aspects: they possess a threshold beyond which a datapoint is dichotomously considered an outlier or not, and they do not provide any explicit indication on whether the outlier should be included or excluded from the analysis – and thus do not directly speak to our data, which are primarily about how participants' regression estimates vary in the presence of outliers, and of instructions to reject them.

An exception is given by Bayesian approaches, which compute the posterior probability that each observation is an outlier; such probability can be seen as the "weight" that each item has in the regression (a lower probability/weight has a smaller influence on the regression). How does this approach perform in comparison with our participants? In order to answer this question, we computed, for each trial used in our experiments, the posterior probability of

each item being an outlier, as formalized by Chaloner & Brant (1988). Then, for each such trial, we ran 1000 iterations, in which the points in the dataset were excluded depending on their probability to be an outlier (e.g., a point with a probability of 0.8 being an outlier, was excluded, on average, 80% of the times). We then calculated the Deming regression slope of each iteration (i.e., on the items that, on that occasion, were not considered outliers) and took the median of the 1000 iterations. This algorithm provided us with the regression slope predicted by the weighted Bayesian approach for each trial in each experimental condition of our experiments. Next, we calculated the response bias of such a model (figure 31) in the exact same way we did for our participants. For comparison, we also plotted in figure 31 the bias shown by a classic Deming regression algorithm. Indeed, Deming regression is also thought to be more robust to outliers than ordinary least squares, because outlier data points affect (i.e., "pull") the regression line to a smaller extent when they are orthogonally projected to it (as in Deming) than when they are vertically projected to it (as in OLS).

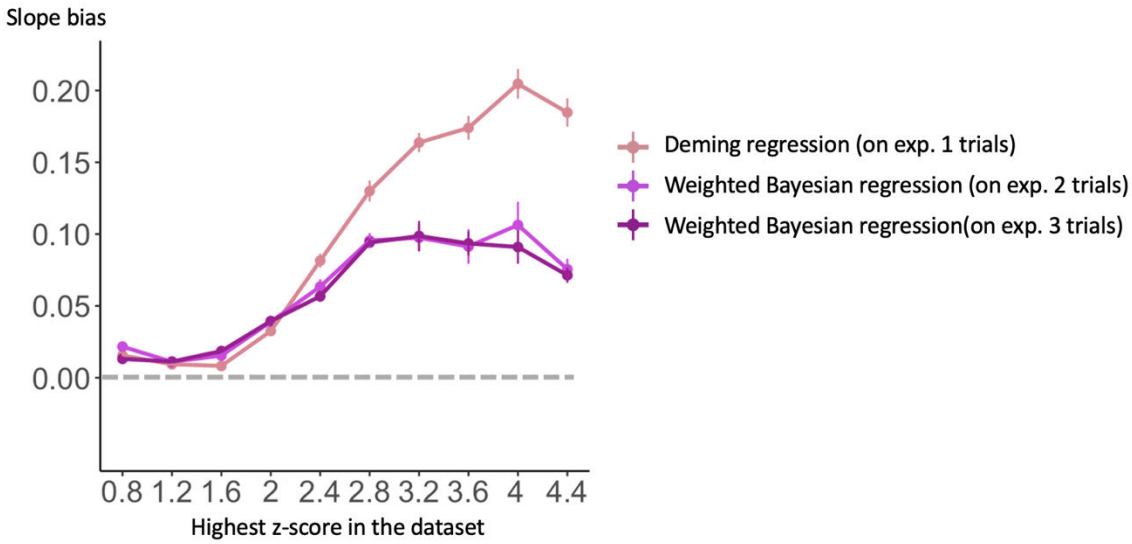


Figure 31. Slope bias of Deming regression (calculated on trials presented to subjects in experiment 1) and of the weighted Bayesian approach (on trials from exp. 2 and 3). The model is adapted by Chaloner & Brant (1988). As evident from the comparison of this figure with figure 6C, Deming regression well mimics participants' response bias in experiment 1: no outliers are excluded. In experiment 2 and 3, the Bayesian model mimics participants' outlier rejection behavior, although human bias (figure 6C) decreases to zero for very large highest z-score values.

The results show that Deming regression, once again, nicely mimics participants' performance in experiment 1, where participants were not explicitly told about outliers, but not experiments 2 and 3. In other words, even if Deming regression is partially robust to outliers, its robustness is modest and both participants (figure 30C) and Deming do not automatically exclude even distant outliers. However, we can see that the Bayesian model (set to probabilistically detect outliers beyond a threshold of z = 2) is much more robust to outliers and shows a behavior partially similar to humans in experiments 2 and 3: for highest z-scores larger than 2.8, its bias stops increasing. Crucially, however, a difference remains: whereas in humans such bias ultimately decreases as the z score becomes very large (figure 30C), the bias for the model remains essentially flat for increasing values of z-scores. The results indicate that the Bayesian model, while close to humans, still differs from them in that it misses a mechanism to sharply reject obvious outliers.

DISCUSSION

Across three experiments manipulating the number and distance of outliers in scatterplots and the level of attention towards them, we probed the human capacity for intuitive statistics in tasks of trend judgment, line fitting and outlier detection, investigating whether outlier items are spontaneously included (as suggested by the literature on graph perception) or rather excluded from any statistical judgment (as predicted by the literature on ensemble perception). We now examine how the results provided answers to the five research questions presented in the introduction of this chapter. We also try to integrate our findings, both with previous findings in the narrow domain of scatterplot perception and with the larger literature on ensemble and outlier perception (which did not use graphs as stimuli); indeed, as appropriately argued by Rensink (2021), studies on graphical representations can

provide fruitful insights not just for graph perception but also, more broadly, for vision sciences.

First, do subjects spontaneously reject outliers when asked to perform a trend judgment or a regression estimation on a graph, without being told that there might be outliers? Experiment 1 is quite clear: participants do not spontaneously reject outliers and they integrate these deviant points in both their trend judgments and their regression estimations. As summarized in the introduction to this chapter, recent studies on ensemble perception (e.g., Epstein et al., 2020; Haberman & Whitney, 2010) showed that, on the contrary, deviant items are easily discarded when participants are asked to provide an estimate of the average of a set. This contradiction might suggest that the intuitive extraction of visual statistics from a graph is not solely a form of ensemble perception. Indeed, when asked to fit a line or extract a trend from a graph, our participants performed a computation that goes beyond the simple "averaging" of a value on a common scale, as is the case for the ensemble perception of items of different hues or orientations. In these cases, the averaging is over a factor that is already present in each individual item: the average color of all items' color, the average orientation of all items' orientations (Whitney & Yamanashi Leib, 2018). In the case of scatterplots, the average item location is useless when assessing a trend, which arises from the relations between data points. Future research should try to disentangle the commonalities and differences between graph and ensemble perception (Cui & Liu, 2021). At the very least, our studies prove that the two processes are not fully overlapping. It is important to point out that, both in our stimuli and in the reviewed papers on ensemble perception, when multiple outliers were present, they were correlated with each other: more specifically, they either had the exact same level of deviation from the average value (Haberman & Whitney, 2010) or they were generated from a secondary value with the

addition of random noise (Epstein et al., 2020), as was also the case here. Future research should investigate whether the same results hold (both for graph and ensemble perception) if the outliers are fully uncorrelated.

Second, do the number of outliers and their distance from the main dataset modulate the bias they induce? Our results from experiment 1 and 2 show that yes, participants' errors and response times in the trend judgment task increase for a higher number of outliers and for a larger distance of these outliers from the main dataset. Likewise, the participants' slope estimates become increasingly biased (i.e., attracted towards outliers) for larger values of these factors. It is worth noting that those increases in error rate, response time and response bias were significantly less pronounced for a main slope of |0.5| than for a shallower main slope of |0.25|. In other words, when the main trend was steeper, outliers were less likely to affect participants' responses. This result makes sense: it is when the decision is most difficult, because the main slope is less pronounced, that outliers have the greatest influence. However, the effect of outliers on response times was still significant even when slope was regressed out, a finding that suggests a serial processing of outliers, with a cost of ~20 ms per item. Overall, our findings extend previous research on outlier processing in scatterplots (Bobko & Karren, 1979; Correll & Heer, 2017; Meyer et al., 1997; Meyer & Shinar, 1992) by showing that deviant points in a scatterplot affect the human capacity for mental regression more if they are numerous and further from the main dataset.

One might argue that the stimuli we used comprised a too small number of observations (18), which may not be sufficient to allow the viewer to form a reliable mental regression from which to detect deviant points. However, the results from figure 26 clearly show that our stimuli comprised enough evidence for subjects to accurately detect the regression slope. The reason we opted for 18 datapoints is double: first, we showed in study 1 that humans are able

to reliably compute mental regressions with as few as 6 datapoints, with a performance close to optimal for datasets like the ones we used in the current study (i.e., with 18 points generated from slopes steeper than 0.2); second, we wanted to avoid conditions in which outliers could too easily pop out, making the task trivial. Future studies could investigate the effects of the overall number of datapoints on outlier detection by parametrically varying this factor.

Third, can the outlier-induced bias be mitigated by drawing attention to them? In the fast trend judgment (first task of experiment 1 and 2), devoting attention to outliers did not significantly improve participants' performance (figure 27). This finding suggests that an extraction of the overall trend (including outliers) occurs fast and automatically – indeed, our hypothesis for outlier rejection suggests that it could be a necessary step prior to outlier detection and rejection. However, the comparison of the response bias in the line adjustment task from the three conditions of attention deployment (exp. 1: none; exp. 2: medium; exp. 3: high) revealed that, yes, outliers are more easily rejected when participants are aware of their presence and invited to discard them. It is worth clarifying that this finding does not imply that attention is needed for outlier processing itself: indeed, our findings from experiment 1 (no attention) clearly show that deviant items affect trend judgments and slope estimations even more if participants are not aware of their presence. In agreement with this, several studies showed that attention is not necessary for the perceptual processing of visual items (Kouider & Dehaene, 2007), which can still attract spatial attention even when subliminally perceived (Astle et al., 2010; Robitaille & Jolicoeur, 2006) and clearly deviating from the other items (Hsieh et al., 2011). However, our results are congruent with the finding that attention can modulate even subliminal processing (Kiefer & Brendel, 2006; Naccache et al., 2022).

When attention was deployed towards outliers (but, crucially, no rejection was asked), one study found that deviant items in size or brightness were integrated in judgments of average size or brightness and biased participants' judgments towards the outlier value (de Fockert & Marchant, 2008). Our findings show that this strong attraction, exerted by both unattended and attended outliers, can be reduced if participants are explicitly asked to exclude them, but experiment 3 suggests that it is hard to fully eliminate – even when a single outlier was present, and it was explicitly detected, it kept an influence on the participants' estimates of regression slopes. An interesting question for future studies is to what extent this strong attraction is resistant to training: in fact, a recent study showed that the estimation of correlation in a scatterplot improved significantly following long perceptual training sessions with feedback (Cui et al., 2018).

Fourth, how does outlier detection work? In the first task of our third experiment, we found that correct detection of outliers improved for larger distances from the main dataset, but also for more numerous outliers. The latter result might be due to at least two different reasons: a larger number of outliers may increase the probability for at least one of them to be seen; and/or it may make them globally more salient and recognizable (Kinchla, 1977). Future studies could try to disentangle these two hypotheses.

Interestingly, outlier detection exhibited considerably slower response times than trend judgments on the whole set (figure 30D for a direct comparison): this observation replicates previous evidence that visual judgments about the average value of the items in a set are faster than the detection of deviant observations present in those sets (Hochstein et al., 2018). This finding agrees with our model, according to which the extraction of the scatterplot trend is a necessary step prior to outlier detection, since the latter is based on their deviation from the main trend. Indeed, the paradox of outliers' detection (Epstein et al., 2020) is that

an outlier is defined as deviating from a summary statistic computed on the entire set, meaning that it cannot be computed without also extracting such a summary reference value. Therefore, the higher response times observed for outlier detection might be the result of a trend judgment phase followed by outlier detection per se. It should however be noted that, perhaps as a consequence of those successive stages, those response times were highly variable, and therefore any conclusion should be drawn with great caution.

We also formulated an explicit model of outlier detection, and tested it against many alternative models. The model hypothesizes that outliers are detected based on their elevated z-score, i.e., their large distance to the regression line, relative to the typical distance of other data points. Participants would compute a z-score for each data point, and evaluate whether the highest of these z scores exceeds a threshold of about 2. This model was supported by both response times and error analyses. In response times, we found a distance effect, whereby outlier detection became increasingly faster for stimuli whose highest z-score increasingly deviated from 2. This is exactly what the model predicts: for stimuli comprising points with smaller z-scores, the absence of outliers is quickly detected, whereas for stimuli with outliers with higher z-scores, their presence is recognized increasingly fast. Likewise, we found that the percentage of “yes” responses was best modeled as a function of the highest z-score, with a sigmoidal function showing an inflection point around about 2. Importantly, the best fit was obtained when the z score was calculated as the ratio between the orthogonal distance of the data point to the Deming fit, and the prescribed standard deviation of the datasets (i.e., the “noise” level). The explanatory advantage of the orthogonal distance over the vertical distance from OLS replicates our results from study 5 and 6 showing that participants minimize the perpendicular Euclidean distance of each point to the best-fitting line when computing a trend (Ciccione & Dehaene, 2021). On the other hand, the explanatory

advantage of the prescribed standard deviation over the actual standard deviation of each stimulus merits a brief discussion. It might have been rational for participants to compute the actual noise level in every individual scatterplot in order to determine if a point is or not an outlier. However, humans are remarkably accurate at encoding the variability in a set of items (Morgan et al., 2008; Solomon, 2010) and they do so automatically, even when not explicitly asked for it (Khayat & Hochstein, 2018). Furthermore, the standard deviation of orthogonal distances from the fit seems also to be used by humans when asked to perform correlation judgments (Yang et al., 2019). Therefore, it is reasonable to speculate that participants in our experiment computed the average noise level across trials, i.e., the prescribed standard deviation, and used it as their reference against which outliers were compared. This would be in agreement with previous evidence showing that human observers have access to a reliable measure of visual uncertainty in decision-making tasks (Barthelmé & Mamassian, 2009).

It is worth highlighting that we do not claim that humans are using explicit mental calculation to compute the z-score of each datapoint in the scatterplot. Indeed, the observed responses times would be incompatible with such a slow procedure. Our data simply suggest that, during fast graph perception tasks, humans deploy a fast process that tightly approximates a statistical model computing z-scores. As reviewed throughout the paper, the human visual system is known to be able to compute complex summary statistics over briefly presented sets of items: the automatic computation of z-scores merely adds to this set of computational abilities. However, whether or not the z-score hypothesis holds should be more precisely studied. Future research could manipulate, for instance, the noise level in successive graphs and asks (1) whether the actual noise level (i.e., the denominator in the z-score formula) can be computed on a trial-by-trial basis; and (2) whether an approximate division of dot distance by this noise estimate actually occurs and what is its accuracy. A more parsimonious

hypothesis is simply that the human visual system recycles its ability to detect objects' contours and principal axes and applies it to graphs, by extracting an estimation of the posterior distribution of all possible graph's contours (which would obviously depend on how noisy the graph is). Each datapoint would then be perceived either as part of such distribution (and therefore included in the trend estimation) or out of it (thus detected as an outlier).

Fifth, finally, if outliers are correctly detected, does this mean that they can also be rejected? Experiment 3 concludes to the negative: outlier detection does not necessarily lead to outlier rejection. When we modeled participants' bias as a function of the highest z-score in the dataset (figure 30B), we found that correct detection of the presence of outliers approached 90% for a highest z-score of 2.8. However, the response bias in the subsequent regression estimation (in which participants were asked to reject outliers; figure 30C) showed that, although the bias was reduced in experiment 3 (high attention) as compared with the two other experiments (none or medium attention), it was at its peak for a highest z-score of 2.8. It is only for stimuli with a highest z-score larger than 3.6 (i.e., with at least one extreme outlier) that the bias disappeared.

Interestingly, we also showed that an optimal Bayesian model that assigns a lower weight to outliers on the basis of their z-score (therefore, without fully rejecting them) behaves somewhat similarly to our participants, suggesting that human outlier detection and rejection may be a probabilistic computation. However, in this Bayesian model, the bias does not decrease sufficiently for large z-scores, whereas the human bias almost disappears then. This discrepancy may be due to the fact that the model uses the actual noise in the dataset, rather than an estimate of noise averaged over several trials (as used by humans). In fact, for larger highest z-scores, when more than one outlier is present, the z-score of those outliers necessarily decreases because high z-scores increase the overall noise level and, as a

consequence, decrease their weight in the regression. On the contrary, humans seem able to calibrate their rejections on the basis of the noise of the main generative process, as already discussed in a previous section.

Taken together, these findings suggest that outlier rejection depends on two factors: the degree of attention towards them, and their deviation from the main dataset. Both factors seem to influence participants in placing a threshold past which they would be more likely to consider a data point as an outlier, beyond the normal noise in the dataset. In other words, the same data point could be seen either as the result of normal variability in the graph or as a significantly deviant observation, depending on task instructions. However, even when participants were maximally invited to pay attention to outliers and to detect and reject them before performing any regression estimation (experiment 3), non-extreme outliers still biased their performance, even when they were correctly detected. This finding suggests that, to some extent, mental regression may be cognitively impenetrable (Pylyshyn, 1999; Stokes, 2013): correctly detecting outliers does not prevent them from influencing the participants' mental regression estimates. We can reasonably conclude that outliers in a graph are not treated as sets of items, thus confirming that graph perception does not operate identically to ensemble perception. We speculate that trend judgment and regression estimation are fast and largely automatic and that outliers, if present and detected, are rejected at a later time, with cognitive effort and following a probabilistic computation. In support of this hypothesis, a recent fMRI study on the neural bases of outlier processing for sets of colored objects (Cant & Xu, 2020) found that voluntarily discarding outliers led to activations that were not confined to early visual areas but involved fronto-parietal areas. Thus, two different types of processes (Kahneman, 2003) seem to be deployed during graph perception. Visual

perception, including the automatic computation of the principal axes of an object or a graph, seems to interact with higher-level cognition, including the deliberate rejection of outliers, with the second process not always able to counteract the information coming from the first (Pylyshyn, 1999).

Lastly, it is important to point out that our experimental tasks focused solely on the psychophysical aspects of graph perception, and did not include any specification of the names, characteristics and meaning of the x and y variables, as one would expect from "real" bivariate graphical representations. It seems likely that participants would have behaved differently if the stimuli were referring to actual data: indeed, outliers are usually either included or rejected from main analyses depending on several factors, including the statistical framework adopted by the scientist (frequentist or Bayesian), the experimental procedure of data acquisition, the type of variables, and their meaning. While our studies investigated the perceptual stages of outlier detection and rejection, future work should also consider using more ecologically valid stimuli in order to evaluate to what extent explicit knowledge of the data affects participants' biases and their probability to include or reject outliers.

EVIDENCE-BASED SUGGESTIONS TO IMPROVE DATA VISUALIZATION OF OUTLIERS IN SCATTERPLOTS

Based on the findings presented in this manuscript, we conclude by proposing a few suggestions to improve outlier detection and rejection in data visualizations. Since these guidelines are speculative, although evidence-based, future research should empirically test their utility through appropriate behavioral studies.

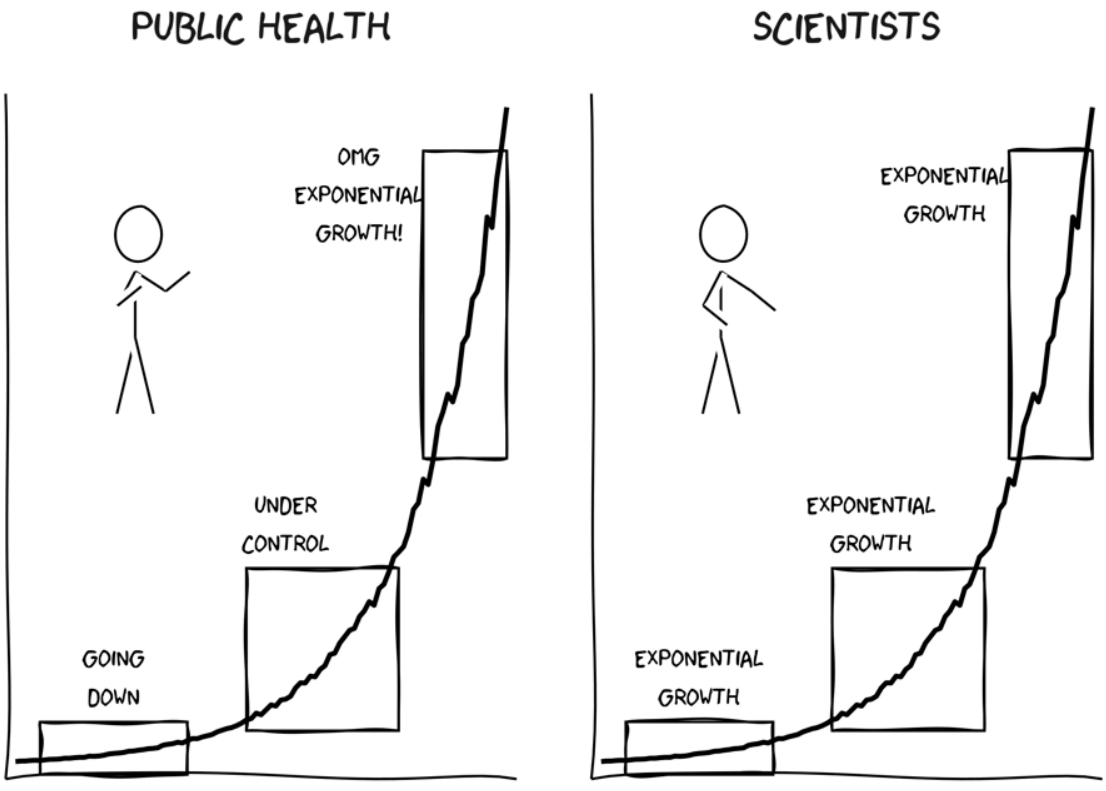
1) Given that outliers are not spontaneously rejected, it could be helpful to explicitly identify all datapoints that exceed a predetermined z-score deviation from the

overall linear regression. For instance, they could be put in a different color or, preferably, a smaller size or luminance. Such a manipulation of size and luminance was shown to be successful at modifying people estimations in a barycenter task (Hong et al., 2021) and perceptual judgments have been proved to operate correctly even in parallel, over multiple sets within a scatterplot (Gleicher et al., 2013).

2) Since human mental regressions tend to be performed on the whole dataset, even when outliers are correctly detected, scatterplots could include both the regression applied to all points and the regression after exclusion of the points that exceed a predetermined z-score. The direct comparison of a robust regression with a non-robust one could help make the discrepancy between the two models more salient to the reader.
3) Since outlier detection is better than outlier rejection, interactive visualizations may help. A regression line would first be calculated over the entire dataset, and then the user would select potential outliers. The regression slope would then instantly adapt to exclude those points, which would allow for an interactive, on-line visualization of how outlier rejection changes the regression. In this manner, the defects of human intuition would be supplemented by human-machine interaction.

CHAPTER 4

PREDICTING THE UNCERTAIN FUTURE: EXTRAPOLATION FROM NON-LINEAR NOISY TRENDS



XKCD comics

In the studies presented so far, we showed that humans can accurately judge if a noisy linear trend is increasing or decreasing, fit a regression line over it and even extrapolate from it. Performance was highly consistent across these three tasks of trend judgment, line fitting and extrapolation and was characterized by what we called a Deming bias: humans minimize the orthogonal fit to the points rather than the vertical one. We also showed that human mental regression is not entirely robust to outliers, even when attention is deployed towards them. We next wanted to challenge humans "beyond linearity": we wondered whether we could characterize their ability to forecast the evolution of noisy scatterplots generated from non-linear trends. In this chapter I present findings obtained in the laboratory, in which we asked 10 participants to predict future datapoints from piece-wise, quadratic and sinusoid graphs (study 8). I also present a large-scale online study (N = 521), in which participants were asked to extrapolate from exponential functions (study 9).

STUDY 8: EXTRAPOLATION FROM NON-LINEAR NOISY SCATTERPLOTS

Linear regression is a standard procedure that, in most statistical packages, is typically applied without any verification of whether a linear fit does or does not make sense (e.g. if the data actually follows a non-linear trend). A famous example of the risks of performing linear regressions without visually exploring a graphical representation of the data is given by the famous Anscombe quartet (Anscombe, 1960). In study 8 I ask if humans are superior to standard linear regression, in the sense that they can recognize whether and when such a regression is appropriate. We used the same extrapolation task as in study 6, but we extended the scatterplots to non-linear functions, and asked whether humans could adequately adapt their extrapolations to the specific function exemplified in the graph. Our hypothesis was that human adults may be able to go beyond linearity and spontaneously identify non-linear statistical trends, although their performance might still be affected by a preference for linear trends (Kalish et al., 2007; Little & Shiffrin, 2009; Mcdaniel & Busemeyer, 2005). Specifically, our experiment was designed to disentangle two possible views of human extrapolation. Under hypothesis 1, participants would be restricted to linear extrapolation, based either on the last few points of a curve, or the tangent to the curve. Under hypothesis 2, participants would infer the nature of the curve and adapt their extrapolation correctly, either by taking into account the curvature in the last few points of the curve, or even the entire underlying function.

METHODS

Participants

10 participants were recruited for the experiment (age: 23.1 $\pm$ 3, 5 females, 5 males), with the same inclusion criteria of studies 1, 5, and 6. They were paid 5 euros for their participation.

The experiment lasted approximately 25 minutes and was approved by the local ethical committee.

Experimental design and stimuli

The stimuli were generated according to the same algorithm as studies 1, 5, and 6, but five different functions were used to generate the scatterplot. All had the same absolute value of the derivative at their rightmost point within the interval plotted (x = 1), but half of them had a positive derivative (figure 32, left column) and half of them had a negative derivative (figure 32, right column). The functions were the following:

1) Two linear functions having equations $l_1(x) = 0.65x + 0.18$ and $l_2(x) = -0.65x + 1.47$.

2) Four piecewise linear functions composed of two straight-line segments. Two functions had their inflection point early on, at 1/3 of their length; equations: $pl_1(x) = 0.65\left|\frac{1}{3} - x\right| + 0.39$ and $pl_2(x) = -0.65|\frac{1}{3} - x| + 1.26$. The other two piecewise linear functions had a late inflection point, i.e. at 2/3 of their length; equations: $pl_3(x) = 0.65\left|\frac{2}{3} - x\right| + 0.61$ and $pl_4(x) = -0.65|\frac{2}{3} - x| + 1.04$.

3) Two quadratic functions, either convex or concave, with equations $q_1(x) = 0.92x^2 - 1.19x + 1.09$ and $q_2(x) = -0.92x^2 + 1.19x + 0.56$. Crucially, the x coordinate of, respectively, the minimum and maximum of these two functions (x = 0.65), was almost identical to the x coordinate of the inflection point of the two piecewise linear functions (x = 0.66). These quadratic functions allowed us to examine if participants would correctly estimate the curvature of the graph, or even its entire quadratic trend.

4) Two sinusoidal functions completing 1.5 periods on screen and with two opposite phases, which corresponded to the following equations: $s_1(x) = 0.2\sin(11.28x) + 1.02$ and $s_2(x) = 0.2\sin(11.28x + \pi) + 0.63$. These functions were the only ones for which the

extrapolated points at x = 1.3 and 1.6 were not monotonically ordered (see figure 33). They were selected to examine if participants would correctly take into account the global oscillatory nature of the sine function.

For this experiment, the noise added for the generation of the actual scatterplots was kept fixed at 0.05 and the number of points was always 66. These values were chosen in order to allow participants to determine the nature of the curve without making it a trivial task. As in experiment 3, the location of the scatterplot was vertically jittered by a random amount (which was later corrected for in our analyses) and we included a considerable margin (12.5% of the screen) above and below the locations of the correct answers. Figure 32 shows an example of stimulus for each generative function.

Procedure

Display parameters, procedure and task were the same as in study 6. Participants were explicitly asked to give an intuitive answer (at one of the two "probed positions": either x = 1.3 or x = 1.6, exactly as in study 6) and to locate the point on their best estimate of the function from which the scatterplot was generated. The experimental design was a factorial design comprising all possible combinations of 5 generative functions, each with two possible signs of the derivative at x = 1, and two extrapolation positions, for a total of 20 conditions. The task was divided into 10 blocks, each comprising all 20 conditions in random order, for a total of 200 trials; the duration of each block was ~3 minutes. After each block, the participants could take a short break. No feedback was given. Before the beginning of the actual experiment, 25 practice trials were conducted under the supervision of the researcher in order to evaluate the correct execution of the task (but no feedback on correct responses was provided).

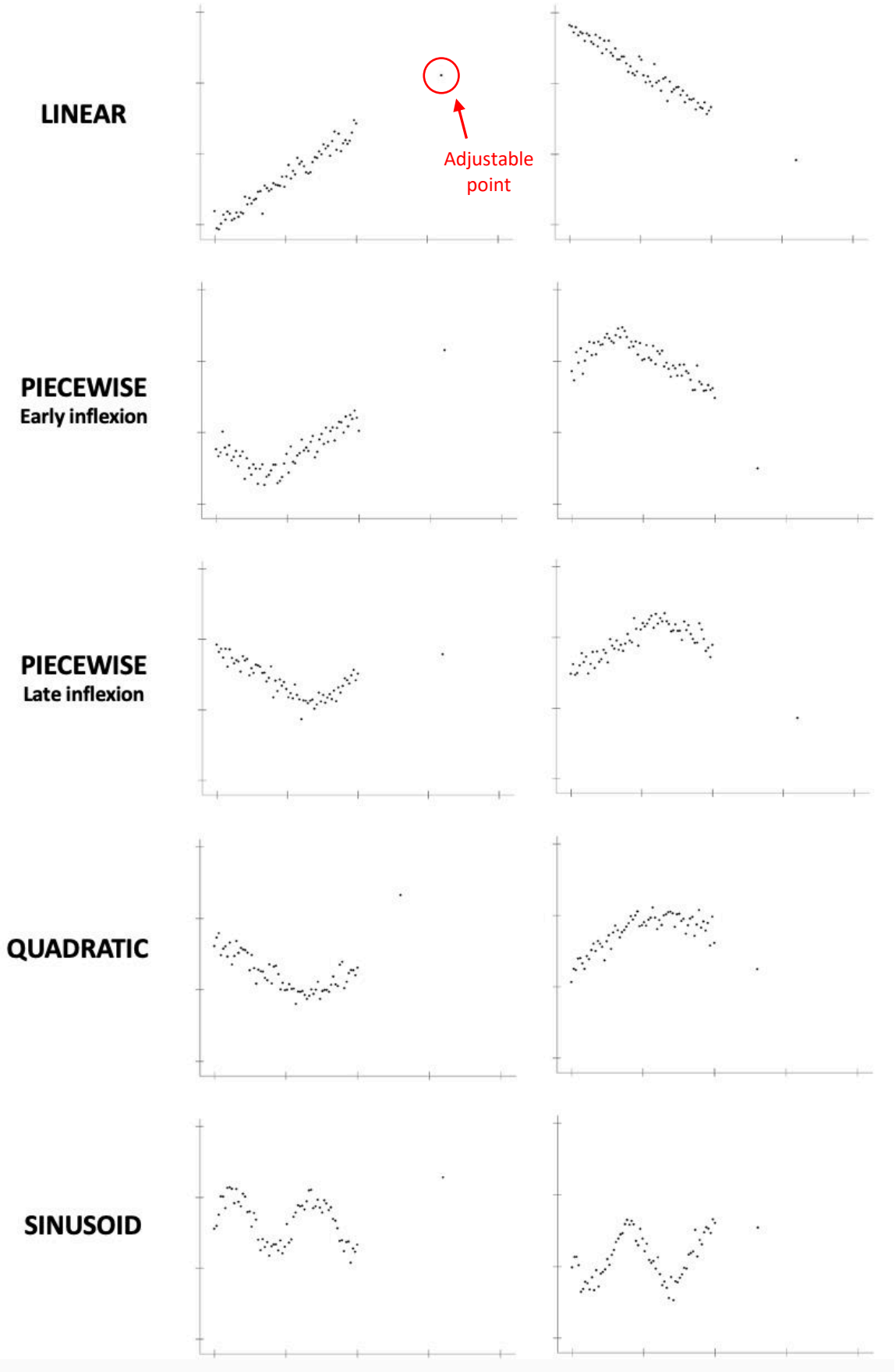


Figure 32. Design of study 8 (linear and non-linear extrapolation). The figure shows examples of stimuli for each generative function used. The movable dot is indicated by a red circle.

RESULTS

Location of the extrapolated point

First, we looked at the distribution of the participants' extrapolation responses on the y axis (figure 33). In all 20 conditions (except for quadratics, to which we later return), the center of the distribution fell close to the ideal answer (black dots in figure 33) and was clearly adapted to the function on the left. To verify this, we conducted a repeated measures ANOVA on participants' median extrapolation value with the type of function, the sign of the derivative and the probed position as within-participants factors, and we found a significant effect of function type ($F(2.65, 23.85) = 3.19$, partial $\eta^2 = .26$, $p < .05$), the sign of the derivative ($F(1,9) = 177.64$, Partial $\eta^2 = .95$, $p < .0001$), an interaction of the sign of the derivative with the probed position ($F(1,9) = 15.03$, partial $\eta^2 = .63$, $p < .01$), and a triple interaction of sign, probed position, and function type ($F(1.61, 14.52 = 33.39$, partial $\eta^2 = .79$, $p < .0001$). The effects and interactions involving function type indicated that participants varied their answers, not only according to the probed position (x=1.3 or 1.6) or the derivative at the end point, but also, crucially, according to the type of function underlying the scatterplot. Since all functions ended with the same derivative at x=1 (the rightmost point of the graph) this finding allows us to reject the hypothesis that participants were confined to a linear tangential extrapolation of the data.

Indeed, examination of the distributions made to the sinusoidal function made it clear that participants readily identified this function and gave adequate non-monotonic responses. In this condition only, the extrapolation at x=1.6 significantly reverted and became closer to the graph mean (y = 0.825) relative to the extrapolation at x=1.3 (for sinusoid with positive derivative: $mean_{x=1.3} = 1.20$, $mean_{x=1.6} = 0.87$, $t(9) = 6.89$, $p < 10^{-5}$; for sinusoid with negative derivative: $mean_{x=1.3} = 0.44$, $mean_{x=1.6} = 0.73$, $t(9) = -4.69$, $p = .001$). This observation is

compatible with the oscillations of the sinusoidal function, but incompatible with any linear extrapolation, either based on a subset of the data points or on the tangent at x=1. In all other conditions, the participants' extrapolations at x=1.6 deviated more than those at x=1.3, in agreement with the monotonicity of the underlying generative functions

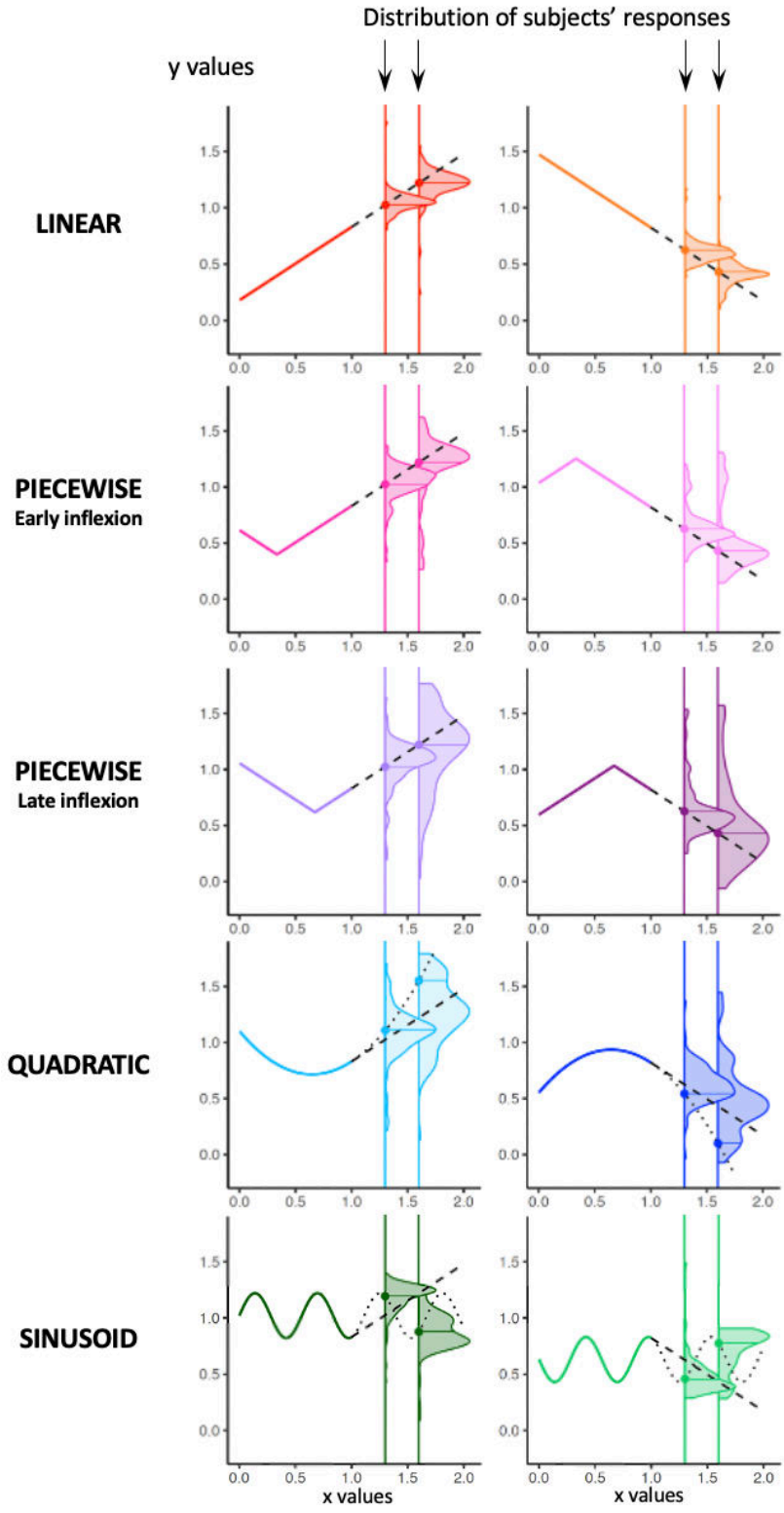


Figure 33. Results of study 8 (linear and non-linear extrapolation). The functions from which the scatterplots were generated are shown on the left, and are prolonged by dotted lines. Large dots indicate the ideal answers at the two probed x positions (x = 1.3 and x = 1.6). Dashed lines show the linear extrapolation, based on the function's derivative at the last point. Density plots show the distribution of all given answers from all subjects for a given x position. For greater readability, all densities were normalized to the same peak height.

Response bias

We next examined how accurately the participants' extrapolations coincided with the ideal location, as derived from the functions used to generate the graphs. For linear and piecewise-linear functions, responses were biased towards extrapolations further away from the expected point, as expected from Deming regression. 718 out of 1200 extrapolations resulted in regression lines steeper than the ideal response (59.83%, one sample proportion test: $\chi^2 = 46.413$, df = 1, p<.0001). Furthermore, examination of the mode of participants' responses revealed a systematic over-estimation of the absolute slope in 12 out of 12 combinations of sign, probed position, and function (see figure 33). Those results confirm the findings from studies 5 and 6: participants' linear regressions are not unbiased, as would be expected under OLS assumptions, but are biased towards steeper slopes, as predicted by Deming-like regression.

For quadratic functions, participants' extrapolations were inaccurate, since their answers were considerably further from the expected location (see figure 33). One sample two-tailed t-tests on participants' median answers revealed a significant difference with the ideal answer for both convex and concave quadratic functions at the probed position of x = 1.6 ($mean_{convex}$ = 1.27, $ideal_{convex}$ = 1.54, t(9) = -4.39, p = .002 and $mean_{concave}$ = .47, $ideal_{concave}$ = .11, t(9) = 4.60, p = .001) and for the concave quadratic at the probed position of x = 1.3 (mean = .63, ideal = .55, t(9) = 2.656, p = .026); for the convex quadratic, the difference was in the proper direction but did not reach significance (mean = 1.07, ideal = 1.10, t(9) = -0.569, p = .6). Crucially, no significant difference was found between participants' answers and the answer expected if participants computed a linear extrapolation based on the derivative at x = 1 (dashed line in figure 33; all p values > .38).

Response variability

As we see from figure 33, the variability in responses greatly varied depending on the type of function and on the probed position; to investigate the statistical significance of this variation, we conducted a Fligner Killeen test, which is a test for homogeneity of variances, robust to departures from Gaussian distributions (see Conover et al., 1981). It revealed no significant differences in the variance of the responses to functions having a positive or a negative derivative (FK med $X^2 = 0.137$, df = 1, p = 0.71). However, both the type of function and the probed x position had a significant impact (respectively: FK med $X^2 = 63.90$, df = 4, $p < 10^{-16}$ and FK med $X^2 = 94.37$, df = 9, $p < 10^{-16}$). Also, the variance of responses for piecewise linear functions with a late inflection was significantly higher than either the one for purely linear functions (FK med $X^2 = 14.33$, df = 1, $p < .001$) or for piecewise linear functions with an early inflection (FK med $X^2 = 6.63$, df = 1, p = .01). This increase in response variability is consistent with participants estimating a linear regression based on increasingly fewer data points (those on the right-hand side of the inflection for piecewise linear curves). It allows us to reject the hypothesis that participants used the fact that the piecewise linear functions were symmetrical, with the same absolute slope on both sides: if that was the case, there should have been no increase in variability, as the same number of points would have been available to estimate the slope for both types of piecewise linear functions. Crucially, no difference in responses' variance was found between quadratic and piecewise linear functions with a late inflection (FK med $X^2 = .004$, df = 1, p = .95). This finding, together with participants' inaccuracy for quadratics, suggests that quadratics might have been misperceived as ending with a linear trend.

To further confirm these findings, we conducted an ANOVA on the standard deviation of the participants' responses, with the type of function, the sign of the derivative and the probed

position as within-participants factors. There was no effect of the sign of the derivative ($F(1, 9) = .02$, partial $\eta^2 = .003$, $p = .88$), nor of its interactions, indicating that participants treated symmetrically the upward and downward-going functions. As expected, we observed a main effect of the probed position ($F(1, 9) = 35.78$, partial $\eta^2 = .80$, $p < .001$), indicating that the standard deviation increased when the probed position went from x = 1.3 to x =1.6, i.e. with greater extrapolation distance ($mean_{x=1.3} = .159$, $mean_{x=1.6} = .276$). Furthermore, there was a significant effect of the type of function ($F(2.63, 23.70) = 10.10$, partial $\eta^2 = .53$, $p <.001$) and its interaction with probed position ($F(2.48, 22.31) = 3.57$, partial $\eta^2 = .28$, $p = .04$). Post-hoc Tukey tests on the standard deviation of participants' given responses confirmed that response variability increased from linear (mean = .08) to both early (mean = .18, $p < .001$) and late (mean = .20, $p < .0001$) piecewise linear functions. It also significantly increased from linear to quadratic (mean = .18, $p < .01$) but not to sinusoid functions (mean = .12, $p = .41$). Also, we found no difference between the standard deviation of participants' answers for quadratic functions and for early ($p = .94$) and late ($p = .78$) piecewise linear functions.

Modelling of an optimal observer

One possible explanation of the poor performance with quadratics is that, given the noise level, there might not have been sufficient evidence to distinguish them from piecewise-linear functions. To clarify this point, we modeled a Bayesian optimal observer (as often done in the domain of visual perception; for a general framework and limitations, see Maloney & Mamassian, 2009; Mamassian et al., 2001) capable of selecting the best-fitting function within four families of functions: linear, piecewise-linear, quadratic or sinusoid. To find the best-fitting curve, the algorithm first used a minimization algorithm to identify, separately for each family of functions, the parameters that minimize the sum of the squares of the vertical distances of each data point to the curve (classical least squares). As a forward model, the

algorithm assumed (correctly) that distance to the curve was drawn at random from a Gaussian distribution with standard deviation σ. Thus, the total Log likelihood of a graph was

$$\ln(L) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n} {d_i}^2$$

where d_i are the vertical distances of each data point to the curve (i.e. $d_i = y_i - y_{model}$).

For the best-fitting model within each family of functions, we computed ln*(L)* using the above formula, with σ equal to the standard deviation of the residual distances. Finally, we selected the function for which the Bayesian Information Criterion (BIC) was minimal, i.e. the one which achieved an optimal trade-off between the number of free parameters (k) and the fit of the data points. We used the following formula: *BIC = k ln(n) – 2ln(L)*, where k is the number of estimated parameters in a given model (including the fixed variance term, i.e. k = 3 for linear regression), and ln(L) is the likelihood of the data for that model. We preferred BIC over AIC since it has been shown that BIC is asymptotically consistent, which means that it will select the true model if it figures among the models considered (Vrieze, 2012); also, BIC tends to penalize complex models more than AIC (Kuha, 2004).

We applied the model to the stimuli presented to the participants and we found that it was almost always able to select the correct underlying generative function (overall accuracy = 93.5%), even for datasets generated from quadratic functions (accuracy = 94.25%). This is crucial, since it means that the noise in our graphs was sufficiently low to make such a task possible. The optimal observer reproduced some, but not all of the features of the data. Like human participants, its response variability increased with the extrapolation distance, from x = 1.3 to x = 1.6. Like humans, the ideal observer was more precise for fully linear than for early-piecewise-linear and late-piecewise-linear functions. However, unlike humans, its estimates were unbiased and centered on the ideal location (see figure 34A). To reproduce

human biases, we returned to Deming-like regression by replacing d_i in the above equations with the Euclidean distance of each data point to the nearest point of the curve. This procedure reduces to Deming regression when the function is linear, but extends the concept to any arbitrary function. When the sum of the squares of this Euclidean distance was minimized, the model, like human participants, revealed a bias towards extrapolations further away from the ideal point, in agreement with Deming regression (see figure 34B).

The last important misfit of the model was that it failed to reproduce the observed human inaccuracy with quadratic functions. Indeed, as expected from Deming regression, the mean extrapolations of the model for the quadratic function fell slightly *beyond* the ideal ones (e.g. for the convex quadratic and x=1.6, correct extrapolation = 1.54, mean model response = 1.57), whereas the converse was true for our participants (mean response = 1.26, which is quite close to the linear tangent-based extrapolation = 1.22). We therefore considered a model that did not include the quadratic functions as one of the possible fits. In this case, quadratic functions were classified as piecewise linear functions 100% of the time. However, as seen in figure 34C, the fit to human data remained inadequate, since the model now predicted values that were almost constant as a function of x position, and way too close to the mean of the data (e.g. for the convex quadratic and x=1.6, mean response = 0.96).

Note, however, that for quadratics, the participants' responses exhibited a large variance and a distribution with multiple peaks roughly coinciding with the three above possibilities (constant response, linear tangent, or quadratic extrapolation); it is therefore possible that they adopted a mixture of the above strategies, and/or that they correctly identified the quadratic but failed to appropriately follow its curvature and, instead, based their extrapolations on the tangent line. We can at least conclude from our theoretical analysis that

(1) all functions were clearly discriminable, including the quadratic and piecewise-linear functions; and yet (2) participants performed poorly only with quadratic functions.

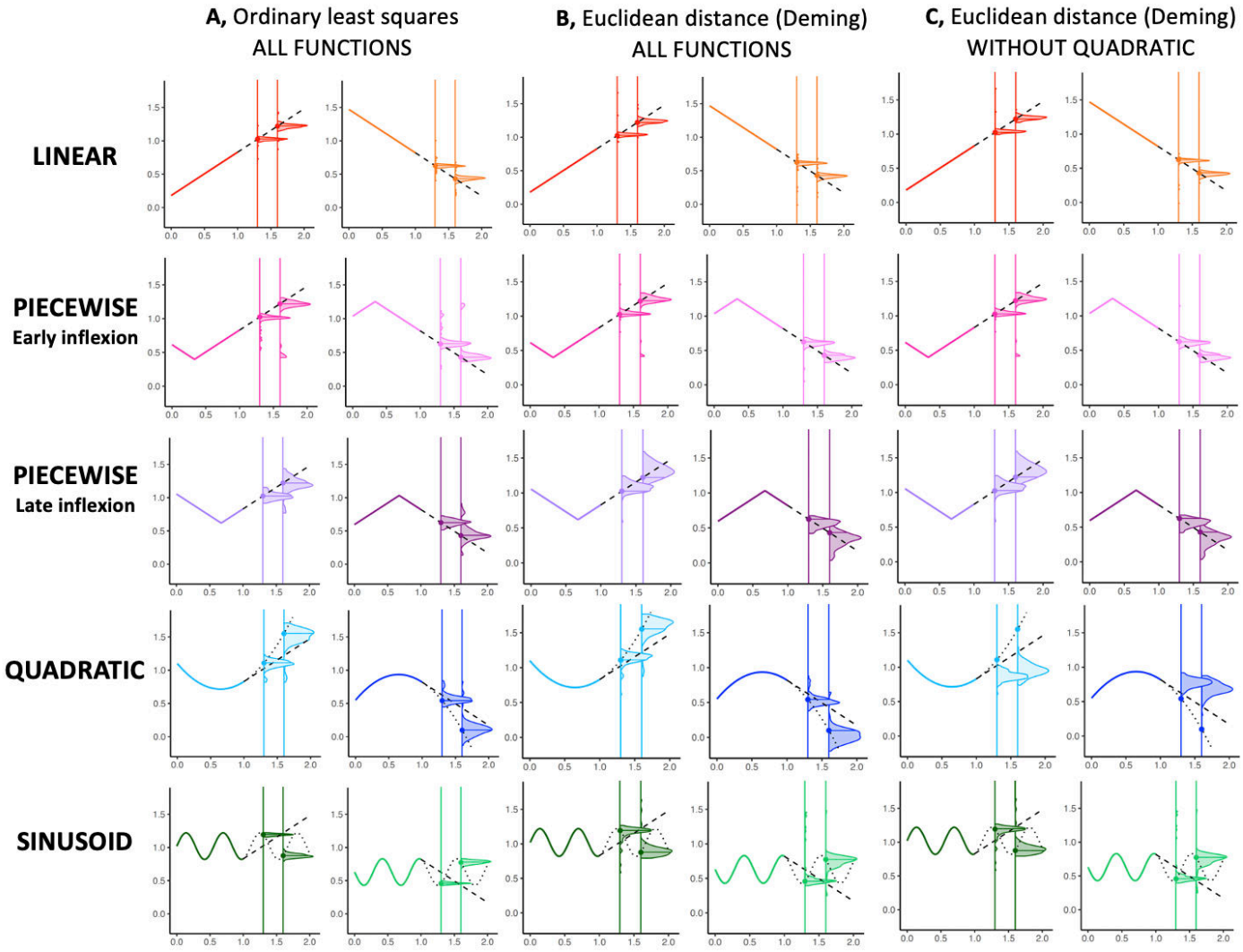


Figure 34. Predictions of three different models. A: Bayesian ideal-observer with full knowledge of the set of functions used, and using ordinary least squares (as appropriate given how the graphs are generated). **B**: Same as the ideal observer, but using Deming-like minimization of the Euclidean distance of each data point to the regression curve. **C**: Same as B, but after removal of the quadratic functions from the hypothesis set.

DISCUSSION

Study 8 showed that, unlike a fixed linear regression package, participants could flexibly adapt their regressions to the nature of the graphs they were exposed to. They understood linearity, bi-linearity, and periodicity (sine function) and used this knowledge to extrapolate points

outside the observational range, even for non-linear functions. The sole exception was quadratic functions, for which participants systematically underestimated the size of the variations of the function on most trials, and often chose to follow the tangential line rather than the actual curvature of the quadratic. This finding is consistent with prior evidence for a human preference for linear trends (Kalish et al., 2007; Little & Shiffrin, 2009; Mcdaniel & Busemeyer, 2005). Several previous studies reported a so-called "exponential growth bias", according to which humans, when facing a series of data points that underlie an exponential increase, tend to underestimate the position of a point outside the observational range (Andreassen & Kraus, 1990; Eggleton, 1982; Wagenaar & Timmers, 1978; Wagenaar & Sagaria, 1975; Wagenaar & Timmers, 1979); I will specifically investigate the nature of this bias in the next study. Our results uncovered such an underestimation for quadratic functions as well, thus suggesting that it might occur more generally for curved functions, not just exponentials. Whether this bias is proportional to the function's acceleration should be investigated by future studies. Also, we show that this bias may be restricted to non-linear functions that are, in their extrapolation area, monotonic, since participants correctly extrapolated sinusoids with a high degree of precision.

The consistent misperception of quadratics is compatible with the hypothesis that human compositional grammar (Piantadosi et al., 2016) of functions may be limited to a certain set of primitives such as linear and oscillating functions (Schulz et al., 2017) but not quadratics. It fits with the just mentioned previous observations that exponentially accelerating progressions are consistently underestimated. Tentatively, the misperception of accelerated functions may also be related to the history of science and, specifically, the remarkably slow discovery of the quadratic law of falling bodies. For centuries, the accepted theory of the "impetus" misconstrued the trajectory of a projectile as a line, followed by an arc of a circle,

followed by a vertical fall. Galileo finally managed to correctly describe the projectile motion as a quadratic function and formalized the concept of uniform acceleration of a falling body, which is crucial in Newton's second law of dynamics. Interestingly, it has been shown that most students hold pre-Newtonian intuitive views of motion (Espinoza, 2005; McCloskey et al., 1983). The misperception of quadratics might therefore be considered as part of a more general misunderstanding of the concepts of acceleration and deceleration. It would be interesting to further investigate the acquisition of quadratic functions in mathematics students and the specific difficulties that it raises.

With the exception of quadratics, study 8 reveals that humans are not limited to the perception of a single straight axis (the principal axis, as discussed for linear extrapolation in study 6), but can accommodate the perception of complex spaces comprising multiple, possibly flexible subparts. The visual system appears to extract multiple axes of quasi-symmetry that may correspond to the mathematical concept of "medial axis" or "shape skeleton" (Ayzenberg & Lourenco, 2019; Cohen & Singh, 2006; Kelly & Durocher, 2011; Kovacs & Julesz, 1994; Lowet et al., 2018). Mathematically, a medial axis is the locus of points equidistant from the two closest shape boundaries (Blum, 1967, 1973). The concept of medial axis generalizes the notion of principal axis to curved shapes (e.g. the skeleton of a snake). As such, it may explain why we found that human participants could extrapolate the scatterplots of non-linear functions, an ability that goes beyond the mere extraction of the principal axis. The capacity to approximate the medial axis of a graph, i.e. the locus of the curve that minimizes the sum of the square Euclidean distances to the points, would readily explain how humans extract non-linear curves from scatterplots, and why they show a Deming-like bias when doing so. More studies will be needed to test the reliance of participants on medial axis and to verify to what extent the neural recycling proposal applies to graphicacy.

STUDY 9: EXTRAPOLATION FROM EXPONENTIAL FUNCTIONS

As pointed out in the introduction, scatterplots are widely used to visually convey the evolution of measures such as global temperature, salary, or mortality over time (Friendly & Denis, 2005). They allow humans to efficiently grasp a functional relationship at a glance, without serially processing the data in numerical form, and to predict values beyond the available range. The studies presented so far indicate that human adults can extract the trend underlying a graph and that their extrapolation abilities partially extend to noisy representations of linear and non-linear functions. Would they extend to exponentials as well?

The case of exponential functions is important because it captures many real-world processes such as the increase in insect populations (Maino & Kearney, 2015), the compound interest on a financial capital (Blackman, 1919), the area damaged in a fire (Ramachandran, 1986) or the spread of epidemics and rumors (Bernoulli, 1760; Dietz, 1967; Ma, 2020). Exponentials are mathematical functions of the form $\mathrm{y} = \mathrm{a}\ b^x$ where a is the intercept with the y-axis and b is a positive real number. Exponential growth, which occurs when b>1, may initially look slow, but it will eventually overtake any linear or polynomial function.

Humans are known to consistently underestimate exponential evolutions. This “exponential growth bias” was originally described in behavioral studies in which participants extrapolated a new value from numerical real-world data: their numerical extrapolation was consistently lower than the correct one (Andreassen & Kraus, 1990; Eggleton, 1982; Wagenaar & Timmers, 1978), even when the data was presented in a graphical format (Wagenaar & Sagaria, 1975) or as a computerized representation of growing plants (Wagenaar & Timmers, 1979). The magnitude of this bias was slightly reduced after a short lecture on it (Wagenaar & Sagaria, 1975), but did not correlate with participants’ education level (Levy & Tasoff, 2016).

Misunderstanding exponential functions may have dramatic consequences. During the 2019 coronavirus pandemic, people systematically misunderstood its exponential evolution (Lammers et al., 2020): when asked to estimate the number of Covid-19 cases over the past days, they strongly underestimated its acceleration, and this underestimation correlated with low personal adherence to sanitary measures. Correcting their misperception significantly increased the support for social distancing measures (Lammers et al., 2020; Schonger & Sele, 2020), thus confirming the importance of a correct understanding of exponential functions to improve decision making and social behavior.

In study 9, we investigated whether experimental conditions can be found in which the exponential growth bias is reduced or even eliminated, in order to provide evidence-based suggestions for a better graphic visualization of exponential trends. To this aim, we analyzed in fine detail the factors that determine the misperception of exponential growths in graphs, by rigorously applying the same psychophysical methods and procedures I have been advocating for in this thesis. First, we investigated the role of the response modality: most previous studies asked participants to venture a numerical guess, and thus their biases might have arisen from the numerical estimation stage rather than from the analysis of the function itself. To separate those possibilities, we compared two extrapolation conditions from visually presented scatterplots: participants were either asked to point to the expected location of the extrapolated dot, or to provide a numerical answer.

Second, we analyzed the effect of noise in the dataset. Our studies of linear scatterplots showed that, with increasing levels of noise, participants' extrapolations and regressions' estimations depart more from the statistical ideal. In the case of exponentials, it is not known whether the underestimation bias is already present for noiseless functions or if it only arises

when extracting an exponential trend from noise. To answer this question, we compared extrapolation performance for noiseless versus noisy plots.

Third, we measured the costs and benefits of plotting the data on a log scale. When the y axis is logarithmic, exponential growth looks linear, and such linear trends can be perceived without much bias, as we showed in our first studies. Thus, we hypothesized that the advantage of extrapolating from a visually linear trend could eliminate the exponential growth bias altogether. In support of this hypothesis, framing exponential evolution in terms of doubling times (i.e., taking a logarithmic perspective on data) was shown to reduce the bias (Schonger & Sele, 2020). However, we also investigated whether this advantage was mitigated by the difficulty of interpolating numbers on a log scale.

Finally, we checked whether education in science and mathematics can help overcome the exponential growth bias. To this end, we asked whether the exponential bias varied with participants' knowledge of mathematics as well as their type and level of education.

METHODS

Experimental design

We used a factorial design with factors of response modality, noise, axis scale, and function type (figure 35A). On each trial, participants saw a scatterplot and extrapolated a new point outside the original data range. In two distinct blocks, they either placed a dot (pointing task) or ventured a numerical value (number task). Within each block, the stimuli consisted in a mixture of noisy or noiseless scatterplots, whose y-axis scale could either be linear or logarithmic. Importantly, the same five equidistant ticks were shown, and only their labels were changed (0-250-500-750-1000 in the linear scale condition; 1-10-100-1000-10000 in the log scale condition). Finally, the type of function could either be linear or exponential. In the

"same-graph" condition, the function was *visually* linear in both the linear and the log scale conditions (across subjects, the same exact displays were used). In the "same-function" condition, the function was *numerically* exponential in both the linear and the log scale conditions (across subjects, the same exact numerical values were used), thus resulting in an exponential-looking graph when plotted on a linear scale and in a linear-looking graph when plotted on a log scale (figure 35A). Each trial consisted in a scatterplot comprising 33 horizontally equally spaced white points on a black background inside a square graph.

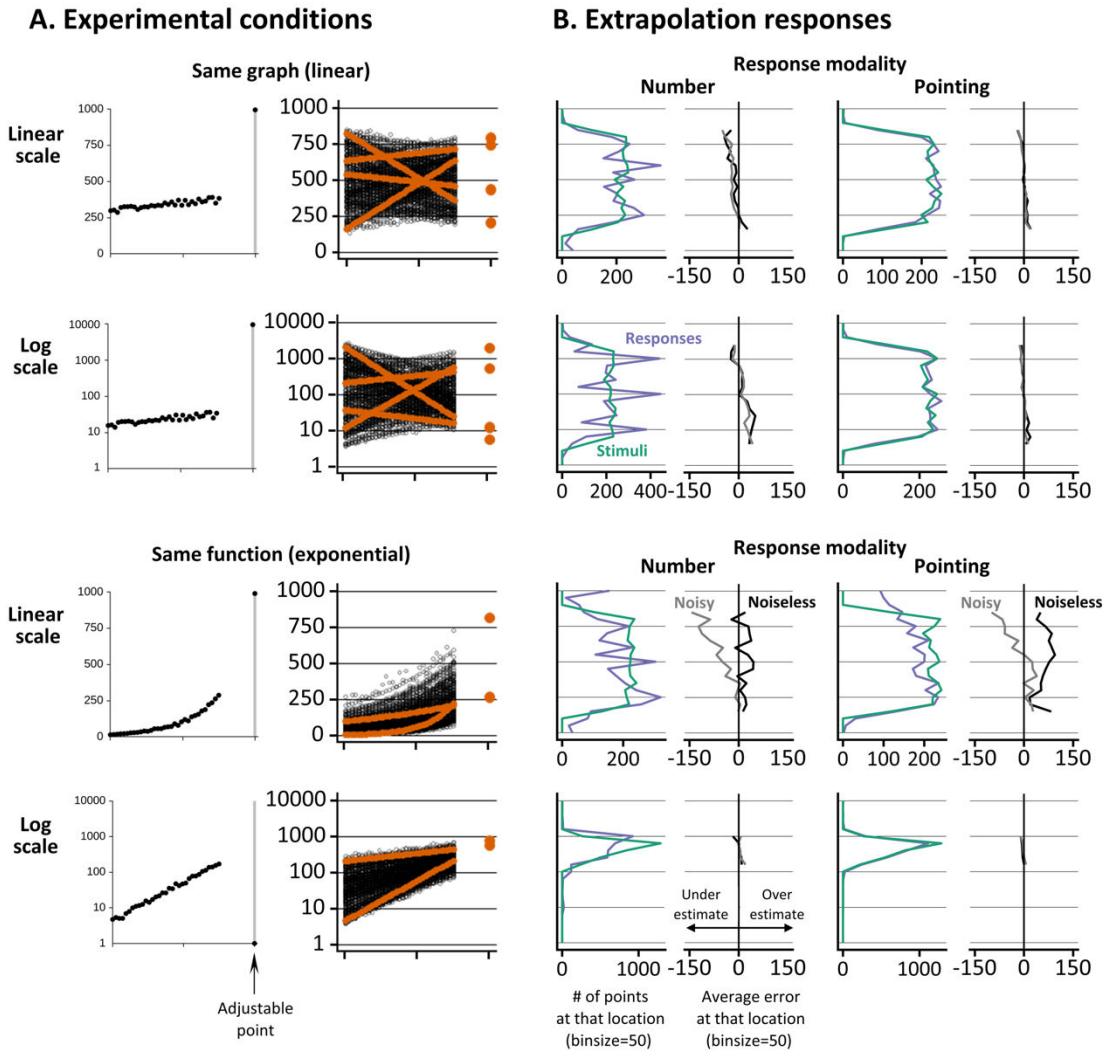


Figure 35. Experimental paradigm to study the human extrapolation of linear and exponential graphs. A: the four experimental conditions, with a sample noisy stimulus for each. Right, distribution of the stimulus datapoints, with some sample noiseless curves in red. **B**: distribution of the participants' extrapolation responses (in purple) compared with the distribution of correct responses (in green). Numerical (left) and pointing (right) responses are separated for sake of clarity. At the right of each subplot, the average error (i.e., the average bias, which simply refers to the difference between participants' response and the correct one) is indicated, for each location of the extrapolation area, separately for noisy and noiseless conditions.

Stimuli

To avoid possible biases due to a reluctance to use the extremes of the available scale, we matched the linear and exponential stimuli in two different ways, as explained below.

In the "same-graph" condition, the slopes and intercepts were randomly varied in order to generate an approximately uniform distribution of desired *pointing* responses in the range 150-850. Concretely, stimuli were generated by uniformly sampling a value Y in [150, 850] and then repeatedly uniformly sampling a slope A greater than 50 or smaller than -50 until the line of equation $\mathrm{y} = \mathrm{Ax} + (\mathrm{Y} - \mathrm{A})$ would take a value between 150 and 850 for x=-1 (the leftmost value). Then, 33 equally spaced points were sampled along the line of equation $\mathrm{y} = \mathrm{Ax} + (\mathrm{Y} - \mathrm{A}) + \varepsilon_\sigma$ for x in [-1, 0.5], where ε_σ was centered noise of variance σ, independent for each point. Finally, either a linear or a log scale was displayed on the left, without changing the screen locations of the dots.

In the "same-function" condition, the two parameters of the function were randomly chosen in order to generate an approximately flat distribution of desired *numerical* responses in the range 150-850. Concretely, we first sampled a value Y' in [150, 850] and then computed $Y = \frac{1000}{4} log_{10}(Y')$. Then, we sampled parameters with the same constraints as the "same-graph" condition (plus the requirement that the slope would be positive), to define a set of 33 points from the equation $\mathrm{y} = \mathrm{Ax} + (\mathrm{Y} - \mathrm{A}) + \varepsilon_\sigma$. Then, we either presented the points with a log scale on the left (log scale condition), or computed, for each point, the numerical value when shown on the log scale, and presented these values on a linear scale, thereby displaying a function of equation $y = 10^{\frac{4}{10} * (Ax + (Y-A) + \varepsilon_\sigma)}$, or equivalently $y = 10^{\frac{4(Y-A+\varepsilon_\sigma)}{10}} \times \left(10^{\frac{4A}{10}}\right)^x$, where the second term is the base, or growth rate, of the exponential. By doing so, we

ensured that the correct response ranged from 150 to 850 in the linear scale condition, and this procedure yielded bases ranging from 1.59 to 14.21.

The noise level σ was set to either zero (noiseless condition) or 10. This value was chosen to add a moderate amount of noise without hiding the underlying linear or exponential trend.

Participants

625 participants took part in this on-line experiment. The experiment was advertised on Twitter and Facebook, starting from the 18th of February, 2021. People interested to participate could simply click on the provided link, choose their preferred language among six (English, French, Italian, Spanish, Portuguese and Chinese) and read and accept a written consent, in which they declared not to be legally minor. Participants were informed they could withdraw from the experiment at any moment by simply quitting the webpage. The procedure and the consent were approved by the local ethical committee. Data collection for the purpose of the study was stopped two months later, the 18th of April, 2021. 521 participants met our criteria for data analysis, i.e., answering to all questions proposed and not giving the default answer to the four questions on self-evaluation. Note that 33 participants declared passing the experiment more than once, but the results were unchanged when their data was excluded. Their data was merely discarded from correlational analyses.

Experimental procedure

The experimental procedure started with a series of questions on demographic aspects (country of origin, gender, age, highest degree obtained; average grade in mathematics and academic field) and on subjective self-evaluations assessments, with answers on a Likert scale from 1 to 10: current skills in mathematics; current skills in first language; familiarity with graphs; knowledge of statistics. Then, participants performed one experimental task (the

pointing task or the number task), preceded by short instructions. At the end of their first task, participants were invited to continue with the second task, which was again preceded by the instructions. The order of the tasks was randomly assigned for each participant. Both instruction texts were each accompanied with a simple explanatory figure. The instructions to the pointing task were the following: “In this part of the experiment, you will see scatterplots such as in the figure below. Place a point in the grey zone where you think that the curve will go. When you are happy with the location you chose, just click. Try to be as precise as possible but do not use rulers and do not make calculations: we want your intuitive response!”. The instructions to the number task were the following: “In this part of the experiment, you will see scatterplots such as in the figure below. Consider the grey zone and estimate the value that the curve would reach if it continued till that zone. Try to be as precise as possible but do not use rulers and do not make calculations: we want your intuitive response!”.

Each participant performed 48 trials: 3 stimuli were presented in each of the 16 experimental conditions resulting from the combination of response modality, noise and y axis scale, in each of the same-graph and same-function conditions. Feedback on overall performance was provided only at the end of the experimental session, which lasted approximately 10 minutes including the questionnaire.

RESULTS

Evaluating bias versus variance in participants’ responses

Using a terminology borrowed from machine learning literature (Geman et al., 1992), participants could be inaccurate in two ways: bias versus variance (as already shown in study 8). Bias refers to the average error: a positive bias means that participants overestimate their

extrapolation, whereas a negative bias means that they underestimate it. Variance, or rather variability, here measured by standard deviation, describes how much participants' errors are spread around their average value, thus indicating erratic responding.

On each trial, we measured the signed difference between a subject's answer and the correct answer. Pointing responses were evaluated to the nearest pixels and linearly rescaled to range from 0 to 1000, thus coinciding with the linear axis values. Numerical responses were similarly transformed to screen units (see appendix C for discussion and alternative measures). We could then compute each subjects' mean error (bias) and standard deviation (variability) in each condition. Figure 35B shows the distributions of participants' responses and their average bias for each desired response, and figure 36 shows the average bias and standard deviation in the sixteen conditions of the design. Bias and variability, our dependent measures, were primarily analyzed with repeated-measures ANOVAs, separately for the same-graph (tables 1A and 1B) and the same-function condition (tables 2A and 2B; tables are all in appendix D). For each condition, we included the response modality (pointing vs numerical guess, categorical), the presence of noise (noisy vs noiseless functions, categorical) and the scale of the y axis (linear vs logarithmic, categorical) as within-subjects factors. We first present the results for linear-looking functions (same graph condition), then for numerically exponential functions (same function condition).

Extrapolations from linear functions ("same-graph" condition)

ANOVAs for the "same-graph" condition are presented in tables 1A (for bias) and 1B (for variability). On bias, there were main effects of noise and axis scale (see table 1A). Although there was no main effect of response modality, this variable interacted with axis scale. Similarly, on variability, there were main effects of all variables, and also several 2-factor interactions (table 1B). To understand those effects, we separate the data according to the

axis scale: we first examine linear functions on a linear axis (in the following three paragraphs) before evaluating the impact of a logarithmic axis.

Pointing leads to more precise linear extrapolations than venturing a number. On a linear scale (top line in figure 35, and leftmost points in figure 36), on average, participants were very precise at pointing towards the correct extrapolation value: no bias was found (mean bias=.79; t(520)=.91, p=.4). When asked to type in a numerical answer, the bias remained small, although significantly different from zero (mean bias=-14.6, t(520)=-5.55, p<.001, *d*=.2). As concerns variability, pointing answers were also less variable than numerical ones ($sd_{pointing}$ = 23.68, sd_{number} = 44.91, t(520) = -10.95, p < .0001, *d*=.48). Combined, these results reveal that participants were accurate at extrapolating linear trends, and that their predictions were significantly less erratic when pointing than when venturing a numerical guess.

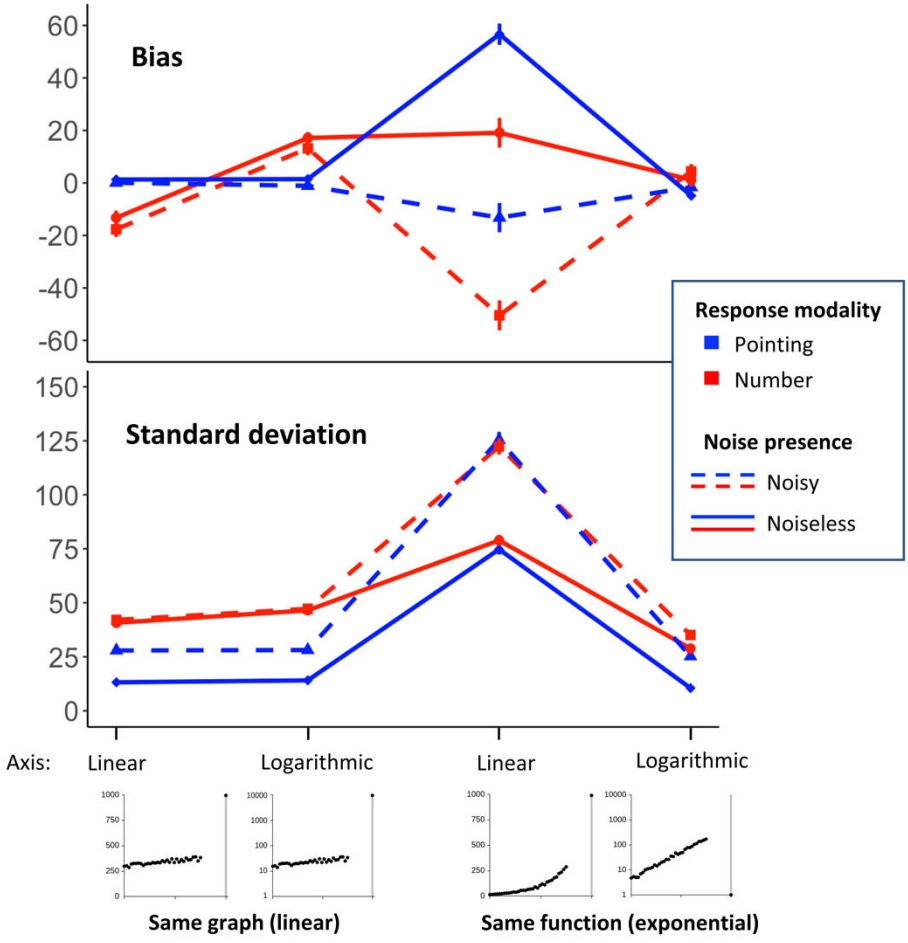


Figure 36. Mean bias and mean standard deviation of extrapolation responses, separately for each response modality, y axis scale and noise level. On the left: results from the same-graph (= linear) conditions. On the right, results from the same-function (= exponential) conditions. The figure makes it clear that responses to exponential curves on a linear axis differ considerably from the other conditions, both in bias and in variability across trials.

Numerical answers are anchored to round numbers. Examination of the distribution of responses indicated that numerical answers displayed a strong anchoring effect: responses spiked for round numbers (purple peaks in figure 35B) while pointing responses followed the flat distribution of desired responses. Reference tick values (0-250-500-750-1000) were used significantly more commonly as numerical guesses (433/3126=13.9%) than as pointing responses (31/3126=.01%; χ^2=374.34, df=1, p<.0001). Furthermore, responses corresponding to tens and hundreds were significantly more frequent than the expected frequency under the null hypothesis that each number had the same probability to be chosen (binomial test, empirical proportion 2597/3126=83%, expected proportion 313/3126=10%, p<.0001). Those findings fit with the known human preference for round numbers, particularly powers of ten (Dehaene & Mehler, 1992; Sigurd, 1988).

Extrapolations from noisy linear trends are more variable but still accurate. We next tested the effect of noise on linear extrapolations: by looking at figure 36, we can see that participants' bias (on a linear scale) was small when extrapolating both noiseless and noisy scatterplots. Their average errors were virtually identical in both cases, with no significant bias in the pointing condition (t-tests against zero, both p>.05), and a modestly larger bias for noisy stimuli in the numerical condition ($bias_{noisy}$=-17.71, $bias_{noiseless}$=-13.27, t(520)=-1.97, p<.05, *d*=.09). Their pointing responses were more variable for noisy scatterplots, as confirmed by a t-test restricted to the two conditions (sd_{noisy}=27.92, $sd_{noiseless}$=13.20; t(520)=9.93, p<.0001, *d*=.44), and in agreement with the findings from study 8.

For numerical answers, surprisingly, additional noise did not increase participants' response variability (two-sample t-test, p=.54), probably because the data was dominated by variability due to the anchoring effect.

Participants are poor at reading numbers from a log scale, except at tick values. We next examined the pure effect of asking participants to read out the numbers from a log scale rather than from a linear one, using identical linear graphs (second line in figure 35, and second data point in figure 36). As expected, the scale had no effect when participants responded by pointing (bias: $bias_{log}$=.51, $bias_{linear}$=.79; t(520)=-.49, p=.63; variability: sd_{log}=24.7, sd_{linear}=23.68; t(520)=.81, p=.42). However, there was a large influence on numerical responses. While response variability was only slightly larger on a log scale (sd_{log}=51, sd_{linear}=44.91; t(520)=3.32, p<.001, *d*=.15), response bias became systematically positive ($bias_{log}$=15.58) and was significantly different from both zero (t(520)=7.5, p<.0001) and from the linear axis condition where bias was negative ($bias_{log}$=15.58, $bias_{linear}$=-14.64; t(520)=13.62, p<.0001, *d*=.6).

What could be the reason for this effect? It appears that, on a log scale, participants have trouble finding which number corresponds to a given location. Indeed, the organization of a logarithmic scale can be counterintuitive. For instance, the value that falls in the middle of 10 and 100 is not 55, as on a linear scale, but 31.6 (the geometric mean of 10 and 100). Participants' failure to understand this property (or, at the very least, to appropriately convert a location on the log scale into a numerical value) could explain their numerical overestimations on a log scale: they would know the correct location (as indicated by their accurate pointing), but fail to turn it into a correct number, instead choosing either the closest tick mark, or using a linear interpolation of the two nearest tick marks, two strategies that would lead to overestimation.

Indeed, both strategies were attested in our data. First, for log scales, response distributions showed prominent spikes at tick values 10, 100 and 1000, which were absent at corresponding points 250, 500 and 750 on the linear scale (figure 35B). Indeed, the anchoring

effect was much stronger on a log scale: participants were more likely to make numerical guesses that fell at or close to the reference ticks' values (±25 screen units) on a log scale (1268/3126=41%) than on a linear scale (814/3126=26%; χ^2=147.77, df=1, p<.0001).

However, there was no significant difference in the frequency of answers corresponding *exactly* to the ticks (log, 452/3126=14.5%; linear, 433/3126=13.9%; χ^2=.43, df=1, p=.51). Thus, participants were still trying to venture an accurate guess, not just responding with the tick value. We next tested the idea that they performed a linear rather than logarithmic interpolation between ticks. We extracted all log-scale trials whose desired response location fell between the second and third tick marks (i.e., between 10-100 on a log scale) and submitted subjects' numerical answers to a multiple regression analysis with two regressors: the correct numerical answer and the answer that would have been correct under linear interpolation. We found that the latter regressor entered as a significant predictor (β=.97, p=.001), whereas the correct numerical response did not (β=.14, p=.65). Similar results were found in the range 100-1000 (but both predictors were significant: $\beta_{linear\_interpolation}$=4.94, p<.001; $\beta_{correct\_response}$=.5, p<.001).

Extrapolations from exponential functions ("same-function" condition)

We next turn to the perception of exponential functions. Overall ANOVAs for the "same-function" condition are presented in tables 2A (for bias) and 2B (for variability). On bias, there were main effects of noise and response modality (see table 2A). Although there was no main effect of axis scale, this variable interacted with both noise and response modality. Similarly, on variability, there were main effects of all variables, and also all 2-factor interactions (table 1B). To understand those interactions, we first examine what happens when exponentials are plotted on a linear scale, then investigate whether their perception can be improved by using a log scale.

Only noisy scatterplots lead to an underestimation bias for exponential functions. On a linear scale (third data point in figure 36), we observed the classical exponential underestimation bias. Importantly, this negative bias was only found for noisy functions, for both pointing responses (bias=-13.23, t(520)=-2.37, p=.02, *d*=.1) and numerical guesses (bias=-50.44, t(520)=-8.85, p<.0001, *d*=.39). On the contrary, participants' answers often overshot the correct prediction in the noiseless condition, for both numerical (bias=56.67, t(520)=13.93, p<.0001, *d*=.61) and pointing answers (bias=19.12, t(520)=3.37, p<.001, *d*=.15). These two opposite biases are clearly visible in the heatmaps in figure 37, which show the entire set of participants' responses for exponential functions on a linear scale. In the noisy condition, the green line, which represents the median response given by participants as a function of the correct location, is consistently lower than the red line, which represents the desired response. For noiseless functions, however, the green line is consistently above the red one, meaning that participants overestimate their extrapolations.

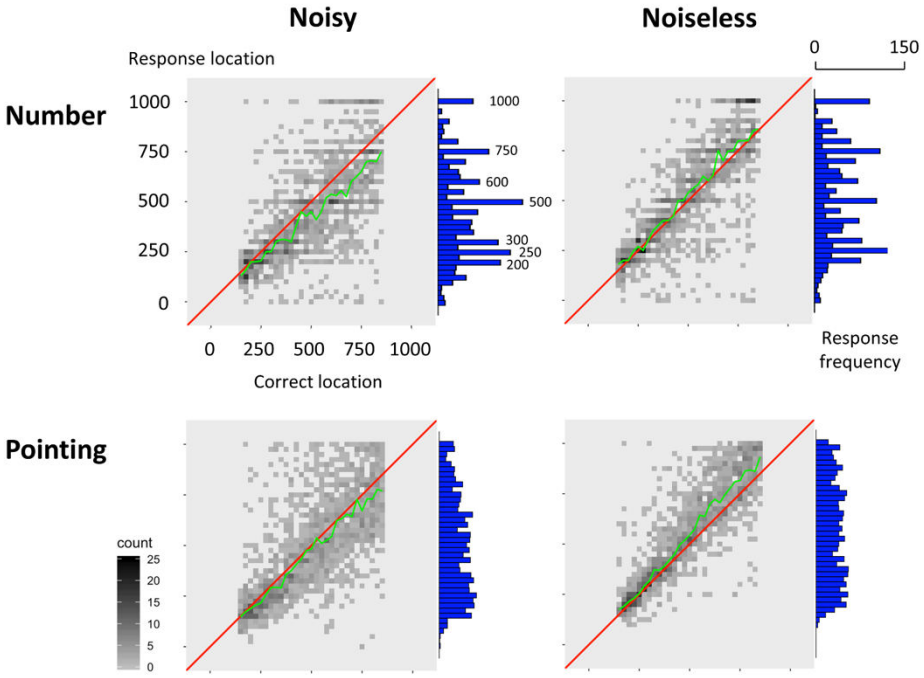


Figure 37. Distribution of the participants' responses to exponential functions on a linear scale, as a function of the expected correct locations (binsize = 25). Plots are shown separately for each response modality and noise level. On the right of each plot, a histogram of the frequency of responses is shown in blue. The diagonal red line indicates optimal performance. The green line shows the participants' median response for each bin of correct location.

Note also that the underestimation bias for noisy exponentials becomes noticeably larger as the exponential growth rate increases and, thus, as the location of the correct answer increases. This trend was confirmed by a linear regression between the bias and the exponential growth rate (R^2=.01, F[1,3124]=44.85, p<.0001) and by a linear regression between the bias and the correct location (R^2=.04, F[1,3124]=145.8, p<.0001). Because of the way our stimuli were generated, correct location and growth rate were correlated (r(3124) = .24, p < .0001). In an attempt to separate those variables, we performed a multiple regression on their normalized values, which revealed that both had a significant effect on bias ($\beta_{\text{correct location}}$=-34.1, p<.0001, $\beta_{\text{growth rate}}$=-12.6, p<.0001; R^2=.05, F[2,3123]=81.16, p<.0001). Thus, although the higher the growth rate, the higher the bias, the bulk of the effect came from the expected correct location: participants misperceived to a larger extent those exponentials that were expected to land at a higher location in the extrapolation area.

In passing, the marginal histograms in figure 37 (top) exhibit spikes at ticks' and round numbers' locations. We confirmed the presence of an anchoring effect for numerical responses, similar to that found for linear functions (see supplementary material).

As concerns response variability, participants' responses on a linear scale were dramatically more variable for exponential functions than for linear ones (figure 2; t(520)=38.88, p<.0001, *d*>1). This was true for both noise levels and response modalities (all p<.0001), confirming that participants always experienced considerable difficulties in correctly estimating the behavior of exponential functions. Extrapolations of exponentials on a linear scale were also more variable for noisy data than for noiseless functions: this was true for both pointing (sd_{noise}=125.38, $sd_{noiseless}$=74.77, t(520)=11.84, p<.0001, *d*=.52) and numerical responses (sd_{noise}=122.33, $sd_{noiseless}$=78.91, t(520)=10.55, p<.0001, *d*=.46).

The underestimation bias disappears when exponential functions are plotted on a log scale. Finally, we examined whether the perception of exponential trends could be rescued by plotting the same data on a log scale. The ANOVA (table 2A) revealed a main effect of the axis scale and its interaction with response modality, confirming what we can observe in figure 2B: when participants were asked to extrapolate exponentials plotted on a log scale, their answers became much less biased and more precise for both response modalities and noise levels. Pointing answers were significantly less biased on a log axis than on a linear axis, for both noisy ($t(520)=2.18$, $p<.05$, $d=.1$) and noiseless scatterplots ($t(520)=-15.39$, $p<.0001$, $d=.67$). The same was true for numerical guesses, which were significantly less biased on a log axis for both noisy ($t(520)=8.65$, $p<.0001$) and noiseless scatterplots ($t(520)=-3.78$, $p<.001$, $d=.17$). Analogously, participants' extrapolations of exponentials were considerably less variable on a log scale than on a linear scale ($t(520)=-42.19$, $p<.0001$, $d>1$).

Those results (figure 37) hold because we compare performance on both axes using the common currency of screen location units. When considering the actual numbers given, the variability for answers on a log scale increases relative to the linear scale but, crucially, the underestimation bias still vanishes (see appendix C). Taken together, these results suggest that, for the very same exponential trend, plotting the data on a log rather than a linear scale eliminates the underestimation bias.

Mathematical education mitigates the exponential growth bias

We next examined inter-individual variability in the underestimation bias. For each participant, we computed the magnitude of their exponential growth bias for noisy scatterplots. The distribution of individual biases, shown in figure 38A, covered a broad range from ~0 (no bias) to -100 or lower (strong bias). Was this variability just noise, or did it vary systematically with knowledge and education? Before the test, participants self-evaluated

their skills in mathematics on a scale from 1 to 10. For the following analysis, we excluded participants that answered 1 or 10, since they were considerably fewer than the other ones (only two and nine respectively).

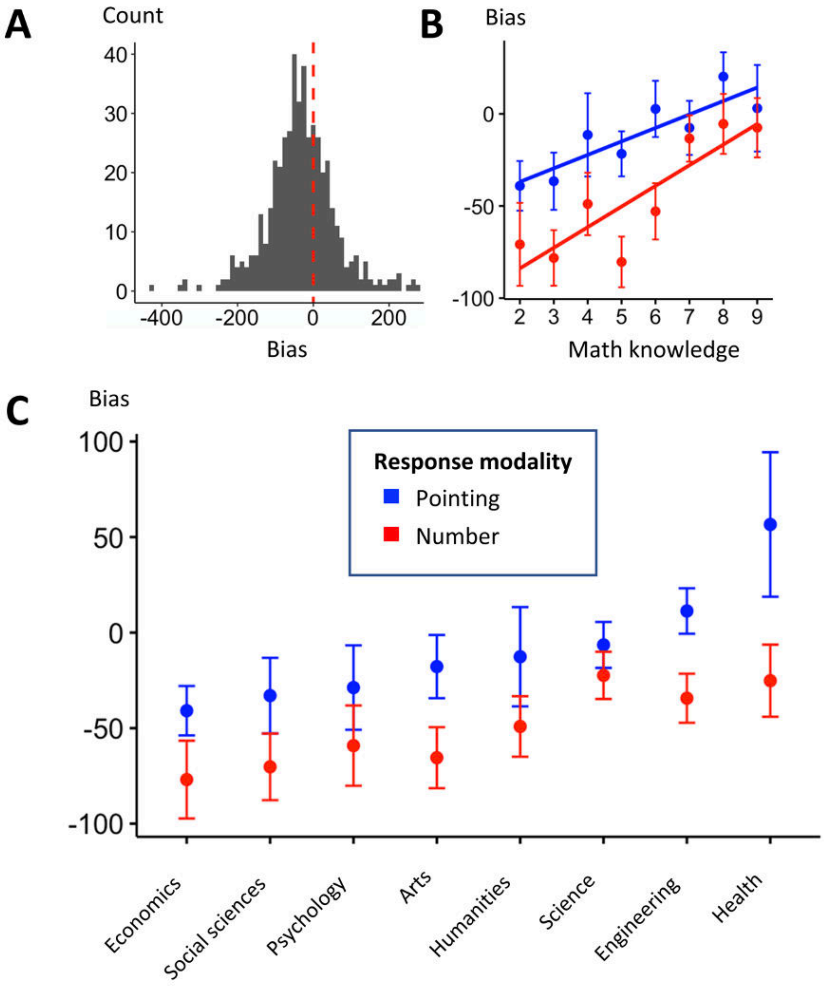


Figure 38. Inter-individual variability in the underestimation bias for noisy exponentials. A: distribution of the bias across participants. **B**: participants' bias as a function of their self-evaluated mathematical knowledge. **C**: participants' bias as a function of their academic field of study.

Participants' self-evaluated mathematical knowledge predicted the size of their underestimation bias for noisy exponentials (figure 38B), for both pointing (R=.12, df=476, p<.01) and numerical answers (R=.19, df=476, p<.0001). This correlation was specific to mathematics and was not found between participants' bias and their self-evaluation of first-language skills (regardless of noise and response modality, all p>.05). For noiseless exponentials, no such correlation was found for pointing responses (R=.05, df=476, p=.26)

and only a moderate one for numerical answers (R=.09, df=476, p<.05). To further analyze how the exponential growth bias for noisy data varied with education, we analyzed it as a function of the academic field in which participants obtained their university degree (figure 38C). We observed that graduates in science, engineering and health (n=221; average bias=-8) performed considerably better in extrapolating exponentials from noisy scatterplots than graduates in the other fields (n=196; average bias=-49; t(410.68)=-4.5, p<.0001).

Other correlation analyses are available in appendix C. In particular, we found that the precision of linear extrapolation increased with mathematical knowledge, both for numerical and for pointing responses.

Putative origins of the exponential growth bias in graph perception

Why do human adults underestimate the growth of noisy exponentials? Participants did not simply prolong the exponential curves linearly: their extrapolation responses were significantly higher than the values predicted by the tangent at the last point (t(6251)=41.8, p<.0001, d=.53). This remained true even when including solely the responses from participants whose mathematical education was below the median (t(3287)=26.11, p<.0001, d=.46). Thus, participants clearly understood that exponential data are positively accelerated. Another possibility is that participants mistook exponential growth for another simpler and shallower relationship such as a quadratic one. We could reject this model, however, because its predictions were significantly lower than the actual data (t(6251)=-50.4, p<.0001, d=.64), and this was true even for responses from the least educated participants (t(3287)=-32.69, p<.0001, d=.57). Furthermore, this theory would only account for the underestimation in the noisy condition, and would fall short of explaining success or overestimation in the noiseless condition.

The latter point suggests that humans possess an intuition of exponential growth, but do not systematically apply it. We reasoned that the underestimation bias for exponential functions could arise from a subset of trials in which noise would make it hard for participants to distinguish exponentials from other functions. The high variability of responses to noisy exponentials (figures 37, 38) would arise from a mixture of trials in which participants correctly detect the exponential trend and trials in which they do not. In fact, this conclusion might be rational: noise could prevent the detection of exponentials even for an optimal observer, who would then revert to the simpler hypothesis of linear or quadratic growth.

To test this idea, we modeled a biased Bayesian ideal observer that attempted to select the best-fitting curve from three alternatives: linear, quadratic and exponential, with a special penalty against exponentials (figure 39A). The model, adapted from the one described in study 8, selects the best-fitting function according to a penalized Bayesian Information Criterion (BIC).

When penalty=0, the model is unbiased and almost always correctly detects exponential growth in our noisy stimuli (1000 trials, accuracy=91.2%). This finding validates our choice of noise level, which was sufficiently low to prevent the confusion of exponentials with other functions. However, it also shows that the unbiased optimal observer does not fit human data. Crucially, participants may not consider linear and exponential relationships as equally likely or equally complex. It seems likely that, with less and less mathematical education, participants would increasingly consider the exponential as a more exotic relationship than the linear one. By biasing the model with a single free parameter (the penalty K associated to choosing exponentials) we get a good fit to the data (figure 39). For a penalty K=3 (which cancels the difference between quadratic and exponentials, the former having one extra free

parameter), exponentials were frequently misclassified either as quadratic (22.6% of trials) or as linear (8.5%), both of which led to an underestimation of extrapolated responses.

We systematically searched for the penalty that provided the best fit to the human data, obtained by minimizing the Jensen–Shannon Distance between the 2D distributions of the model and the data in 3000 randomly selected trials. The best fit was obtained for a value of K=12.62 (figure 39B). With this value, the biased ideal observer model accounted for several additional features of the data. First, while it underestimated noisy exponentials, its average response always fell in between quadratic and exponential extrapolation, similar to humans. This finding arose because when faced with a noisy exponential graph, the model sometimes opted for a quadratic and sometimes for an exponential, but rarely for a linear function. The model also provided a plausible account of the effect of mathematical education, by assuming that less educated participants exhibit a larger prior against exponentials (figure 39C). Indeed, when we separately fitted the data from participants in the first and last quartiles of mathematical education, we found that a lower exponential penalty was required for the former than for the latter (respectively 7.1 vs 15.8; figure 39D). Importantly, these observations held only for noisy exponentials. For noiseless trials, even the heavily biased Bayesian observer (corresponding to a participant with low mathematical education) systematically opted for an exponential function when an exponential was actually present, explaining the absence of a negative bias for noiseless graphs.

Finally, we examined whether the humans and the model failed on similar trials. Using the best-fitting penalty value of 12.62, we split the noisy exponential trials according to the model's classification (Figure 39E). When the model correctly classified the function as exponential (25.5% of trials), participants' bias was not significantly different from 0 (mean=-13.1, $p>.05$) and smaller than when the model misclassified it as either quadratic (mean=-

34.5; 66.2% of trials) or linear (mean=-49.5; 8.3% of trials). The bias was significantly different in the first case than in the other two (pairwise differences in a generalized mixed effect model; linear-exponential: t(2960)=-2.62, p<.05; quadratic-exponential: t(2955)=-2.42, p<.05). Those results suggest that both humans and the model are not blind to exponential growth: bias and noise lead them to often misinterpret curves as linear or quadratic growths.

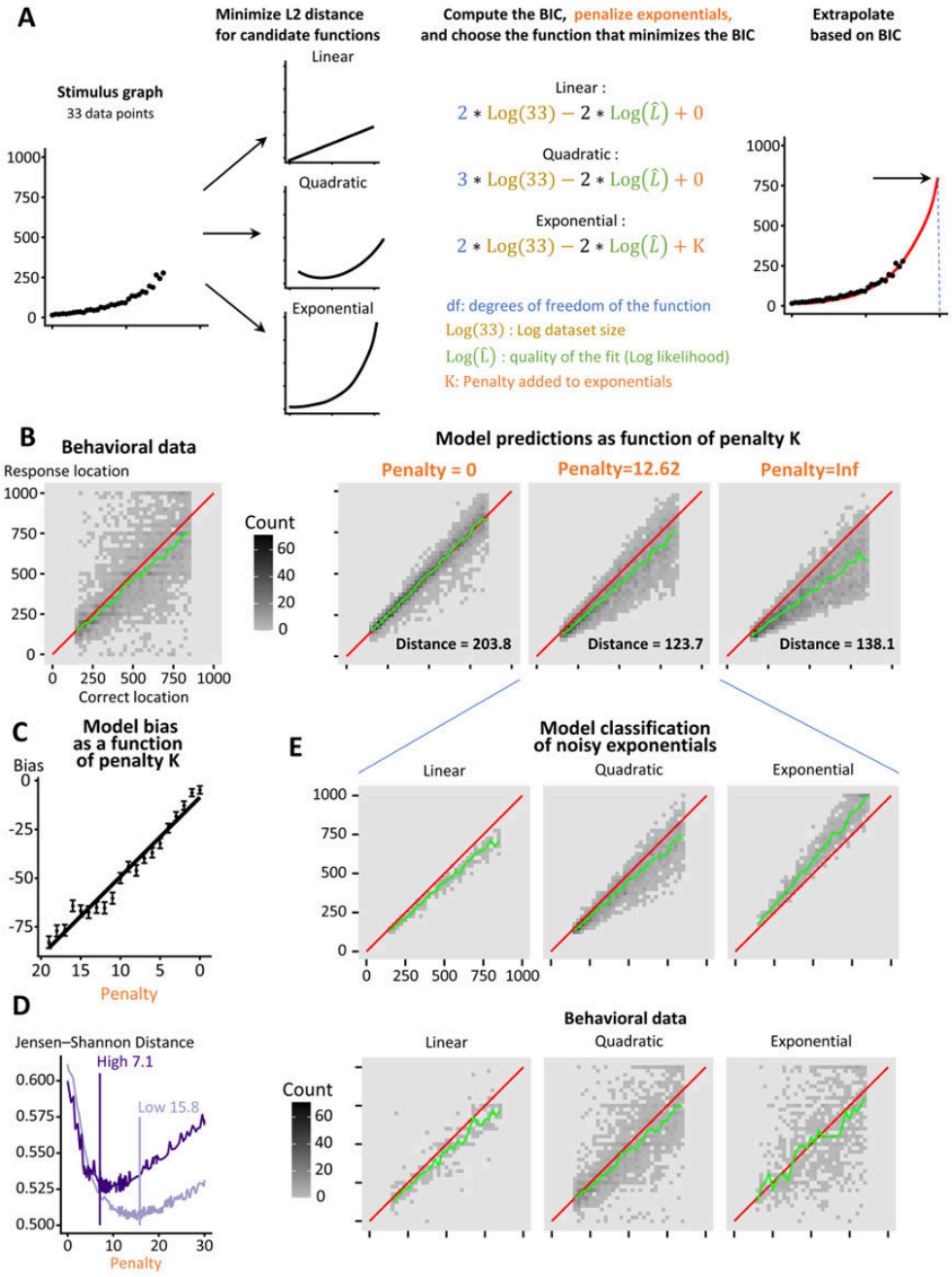


Figure 39. Ideal observer model of graph perception and extrapolation. A: architecture of the model: for a given trial, the model extrapolates the best-fitting candidate function that minimizes the Bayesian information Criterion (BIC) plus an adjustable penalty for exponentials. **B**: comparison of the distribution of answers to noisy exponentials for humans (left) and for the model with three different values of the exponential penalty (right; numbers indicate the Euclidean distance between the data and model distributions). **C**: average exponential underestimation bias of the model as a function of penalty K. **D**: modelling of participants with high and low mathematical knowledge: Jensen–Shannon Distance as a function of the penalty for the first and last quartile of math level, plus indication of the penalty K minimizing the distance. **E**: splitting of the distribution of responses according to the choice of the ideal observer, both for the model and for the behavioral data.

DISCUSSION

In a large-scale online study, we studied the factors that modulate the misperception of exponential growth in graphs. For noisy exponentials, our systematic extrapolation study replicated the exponential growth underestimation bias (Andreassen & Kraus, 1990; Levy & Tasoff, 2016; Wagenaar & Sagaria, 1975; Wagenaar & Timmers, 1979) and extended it to pointing responses. Indeed, although the bias was higher for numerical extrapolations from noisy stimuli, it was still significantly present when subjects were asked to extrapolate a curve by pointing. This finding suggests that the bias arises, at least in part, at the stage of perceiving and extrapolating the scattered noisy dots, and not just when the plot is converted into numerical values. This finding converges with prior studies showing that human adults have difficulties in interpolating or extrapolating accelerating functions (Schulz et al., 2017), including quadratics, as we showed in study 8. In agreement with previous literature (Hutzler et al., 2021), we also found that the magnitude of the bias increases as a function of the exponential growth rate.

For noiseless functions, however, we discovered that the exponential growth bias disappears. This suggests that participants do not lack an intuitive understanding of accelerating functions, but fail to deploy it properly in the presence of noisy data trends. We confirmed this suggestion by exploring the properties of an ideal-observer model of graph perception and extrapolation. The model is based on the idea that participants have to mentally chose between several candidate functions and that the presence of noise in the data may prevent the selection of the proper curve, thereby leading to a quadratic or even linear extrapolation of exponentials. We could reject an ideal observer model that does not penalize exponentials. However, we found that a biased model, with an additional penalty against exponentials, could fit the data. This analysis suggests that participants had intuitions of exponential growth

but were biased against this interpretation and failed to correctly chose an exponential trend in the presence of noise.

We also discovered an independent source of bias in human graph extrapolation, which occurred solely for numerical guesses: participants anchored their answers to round integer values and to tens and hundreds. This “rounding” behavior is intuitive and fits with the higher cross-linguistic frequency of number words corresponding to tens (Dehaene & Mehler, 1992; Sigurd, 1988). It may partially arise from an anchoring heuristic (Jacowitz & Kahneman, 1995) elicited by the values displayed on the y-axis.

One last finding merits discussion: the overestimation bias for noiseless functions. This aspect partly contradicts Hutzler et al. (2021), who found an underestimation. However, this is likely to be due to methodological differences in the studies: their stimuli did not depict genuine exponential functions, but exponentials with a temporally decreasing rate; also, their graphs were either so curved that the correct extrapolation fell way above the depicted y axis or too shallow to be perceived as exponential (which is, indeed, what often happens in the media when the exponential evolution of a pandemic is presented: the y axis is not scaled to anticipate the future number of cases). In our stimuli, we carefully chose the parameters a and b of the exponential functions in order to make sure that participants were not constrained in the range of their possible predictions. Further research will be required to understand why, in such a context, our noiseless exponentials were slightly overestimated. One possibility is that participants visually extended such curves using inappropriate but intuitive geometric operations (Dehaene et al., 2006; Sablé-Meyer et al., 2021). Another possibility is that participants knew that humans underestimate exponentials and overcorrected their responses when they were confident they saw an exponential. Indeed, the experiment was conducted one year after the beginning of the Covid-19 pandemic and

most participants could have been aware of the general underestimation in the number of new contaminated people during Covid-19 outbreak.

EVIDENCE-BASED SUGGESTIONS TO IMPROVE DATA VISUALIZATION OF EXPONENTIAL TRENDS

Many scientists recently called for an improvement in data visualization and graphical representations as a fundamental step towards a stronger public appropriation of societal, environmental, and health phenomena (Concilio et al., 2021; Dixon et al., 2021; Harold et al., 2016; Murray et al., 2020). The present results lead to the following suggestions to improve the perception of exponentials in scatterplots.

1. *Improve people's mathematical education***.** We found that, the higher the participants' self-evaluated math knowledge, the smaller their exponential underestimation bias (as well as their variability in linear extrapolation). Although correlation is not causation, it seems likely that explicitly educating people to the fast-growing mathematical properties of exponentials could mitigate their bias. Indeed, a short lecture on the exponential growth bias was shown to reduce the size of the underestimation (Wagenaar & Sagaria, 1975). Math education correlates with participants' numerosity perception (Ciccione & Dehaene, 2020; Halberda et al., 2008, 2012; Piazza et al., 2013), and intervention studies suggest a causal effect on both number acuity (N. Jordan & Dyson, 2016; Wilson et al., 2009) and intuitive mathematics (Dillon et al., 2017). Recent evidence also suggests that mathematical skills correlate with complex graph understanding (Ludewig et al., 2020). Future studies should confirm our findings with psychophysical experiments in a laboratory

context, and test causality through a randomized control trial with an education intervention on graph perception.

2. *Use pointing rather than numerical responses***.** Participants were more precise and less variable when pointing to the extrapolation location, rather than venturing a numerical response. Thus, inviting the reader to carefully locate the extrapolation point, for instance by clicking before making a numerical estimation, may help. The latter suggestion is particularly relevant for interactive visualizations: readers could click on the extrapolated location while the software would provide the corresponding number.
3. *Use a logarithmic scale, but with a high density of labels***.** A logarithmic scale makes exponential data look linear, and therefore easier to extrapolate without bias. Our data indicate that log scales have pros and cons: they remove the exponential bias, but lead to an overestimation of numerical responses. The latter effect arises from the difficulty of understanding which numbers fall in between tick marks on a log scale. As a result, subjects either select the nearest tick, or interpolate linearly between ticks. Those issues could be mitigated by providing a higher density of numerical labels (e.g., not just 10-100-1000 but also the intermediate locations for decades and hundreds). Our study also replicates and extends previous research showing that exponential extrapolations from a log scale are more variable than those from a linear scale (Menge et al., 2018; Romano et al., 2020). However, this is true only if we consider the actual numerical distance from the correct answer (see appendix C) but not if we consider the distance in the space of the graph.

   Intervention studies could be designed to foster a better understanding of log scales. For example, after performing a numerical extrapolation on a log scale, participants

could receive precise feedback about the discrepancy between their response and the correct one. In young children, such a procedure has been shown to induce a representational change in the conceptualization of the numerical scale, from logarithmic to linear (Opfer & Siegler, 2007). In passing, it is interesting and somewhat paradoxical that children and uneducated adults initially conceive of numbers as logarithmically spaced (Berteletti et al., 2010; Dehaene et al., 2008; Siegler & Opfer, 2003), then move on to a linear understanding in the first years of schooling (Booth & Siegler, 2006; Siegler & Booth, 2004; Siegler & Opfer, 2003), and finally may receive a formal mathematical training on logarithms and log scales. Future work should examine if the adult understanding of log scales rests, at least in part, on a return to the initial intuitive conceptualization of the compressive number line, or whether it constitutes an independent form of learning.

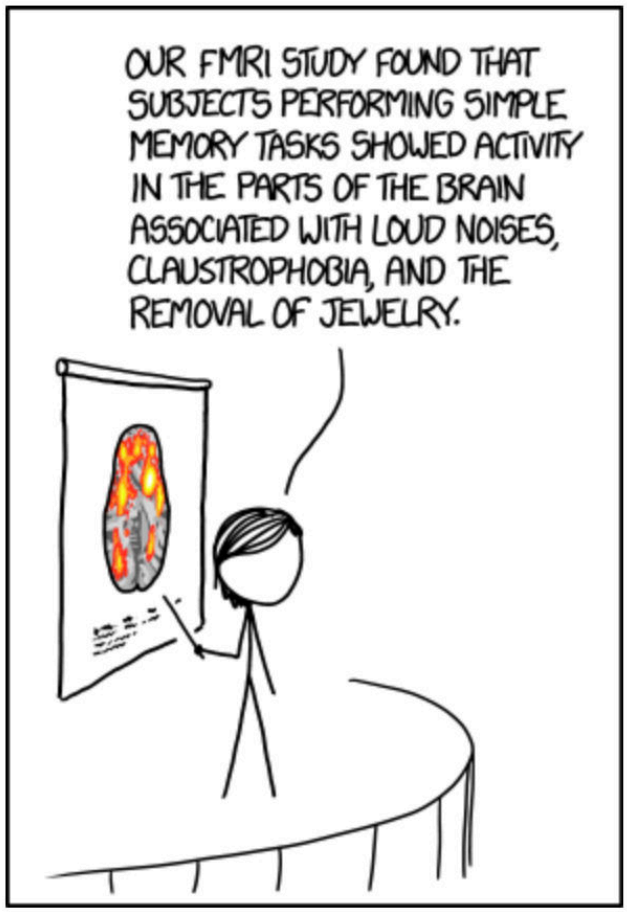
4. *Reduce the noise in the data plot***.** One of the most interesting results in our study is that the exponential underestimation bias vanished (and indeed became an overestimation) when we presented a noiseless exponential function rather than noisy data. According to our model, reducing the noise facilitates the choice of the adequate family of functions, and therefore leads to better extrapolations. This suggests that a more accurate human perception of the exponential growth in real data could be obtained by presenting a smoother graph, e.g. the best-fitting exponential curve or a moving average, instead of noisy datapoints.

We end the discussion of study 9 by noting that the last two suggestions not only eliminate the underestimation bias, but lead to overestimation. One should therefore consider the concrete implications of the conveyed message and the cost associated to either type of error. If, for example, the purpose is to alert the readers to the fast-growing progression of a

pandemic, overestimation may be less problematic than underestimation. However, overestimating the future number of contaminated people may have adverse effects as well, such as excessive sanitary measures. In sum, the preference for a data visualization over another should be based on a thorough analysis of the represented data and of the message that the graph is supposed to convey (Franconeri et al., 2021). While such decisions are beyond the scope of cognitive psychology, the present study highlights the crucial importance of informing them by psychophysical research on graph perception.

CHAPTER 5

THE NEURAL BASES OF MENTAL REGRESSION



XKCD comics

STUDY 10: FMRI STUDY ON TREND JUDGMENT

The behavioral studies presented so far provide us with some new findings about the way humans extract statistical information from noisy graphical representations. They contribute to the growing body of research on graph perception and understanding by proposing a psychophysical approach to information processing. One last question remained, to our knowledge, still never asked to date: what are the neural bases of our abilities to extract statistical summary information from noisy graphs? If we zoom out and broaden our view on all research on ensemble perception (which share many characteristics with graph perception, as described in the previous chapters), very few studies tackled the question of how our brain performs summary statistics of our visual environment (e.g., the average orientation of objects, the average emotional facial expression…) and no study specifically investigated the brain responses to graphical representations. It has been suggested that the neural bases of ensemble perception might be largely distributed and hard to capture (Whitney & Yamanashi Leib, 2018): one study in macaques found that V4 seems implicated in the extraction of high-level summary statistics of objects' textures (Okazawa et al., 2015) and one study in humans found fMRI adaptation in anterior-medial ventral visual cortex for stimuli with the same ensemble statistics (Cant & Xu, 2012). Interestingly, one EEG study found that ensemble properties might even being available before the detection of individual object properties (Epstein & Emmanouil, 2021). Whether these preliminary findings generalize to graph perception is unknown. We therefore thought that a similar neuroimaging approach could be applied to our trend judgment task: for this reason, we used 3T fMRI to investigate the neural bases of human ability to perform "mental regression" in the classic trend judgment task that we developed in our work.

Specifically, we tried to answer to the following four research questions:

1) What are the neural bases of trend judgment? In order to answer this question, we designed a blocked paradigm of the type “same stimuli, different tasks”: participants saw a series of scatterplots composed of triangles and diamonds and, for each of them, they either had to judge the graph trend (ascending or descending) or, in other blocks, the most numerous shape (triangles or diamonds). We then performed simple contrasts between the two types of blocks in order to investigate if the trend judgment task elicited a specific activation in a defined brain area.
2) Where does the brain treat the scatterplot’s parameters (i.e., noise, slope and number of points), including its t-value? Our behavioral studies showed that each of those parameters influenced participants’ accuracy and response times, which they were well subsumed by the t-value associated to the graph. We thus wondered where in the brain these parameters are treated and whether we could find brain areas that reveal an increase in activation for different levels of the t-value. To tackle this question, we designed an event-related paradigm in which participants saw a series of scatterplots (in trials separated by relatively long intervals) that varied in their slope (positive or negative), noise level (large or small), number of points (18 or 38) and the visual hemifield in which they appeared (left or right). The quasi-randomly alternating appearance on both visual hemifields was meant to control for possible retinotopic explanations of our findings: in fact, we wanted to investigate the pure effects of the scatterplots’ parameters independently from the visual area they covered. Concretely, we performed a representational similarity analysis (RSA) on regions of interest.
3) Does the cultural recycling hypothesis proposed for letters and numbers processing apply to graphical representations as well? Along the manuscript we evoked the hypothesis that brain areas devoted to object orientation detection might have been

recycled for graph perception: indeed, detecting the trend of a noisy scatterplot seems analogous to the detection of the orientation of an object but, at the same time, requires the extraction of more complex summary visual statistics. In order to investigate whether such recycling occurred, we designed a blocked paradigm of the same sort described above ("same stimuli, different tasks"), in which subjects were asked to either detect the orientation of an object (ascending or descending), or to identify its category (kitchen tool or not-kitchen tool). We then performed simple contrasts between the two types of blocks and then looked at the overlapping of the brain areas involved in object orientation with those involved in trend judgment. The "objects" trials were proposed at the end in order to avoid priming participants to conceive the graphs as objects when performing the trend judgment (in fact, if they saw objects before graphs, they might have been primed to treat graphs as objects).

4) Is trend judgment sustained by the brain areas involved in numerical cognition? Some of our previous findings suggested that numerical and mathematical cognition might play a role in our ability to correctly perform a trend judgment of noisy scatterplots or to extrapolate from non-linear functions. To investigate whether this hypothesis holds at the brain level, we asked participants to perform language and math simple tasks in order to find, for each subject, the brain areas specifically involved in numerical cognition. We then looked at the overlapping between those areas and those involved in trend judgment.

METHODS

The experimental procedure was divided into 7 runs (plus an anatomical run). Before being put into the scanner, each participant was instructed on how to perform each run and could

make a few practice trials. No feedback on correct or wrong responses was provided: the practice was intended uniquely to check for the correct execution of the task (e.g., giving only one response per trial, fixating the center of the screen, not moving the head…).

Run 1 and run 3: shape detection and trend judgment (blocked paradigm)

The experimental procedure of these runs is shown in figure 40. Participants saw one of the two possible instructions for 3000 ms. If they saw the diamond and the triangle, they had to concentrate, for the following 8 trials (consisting in noisy scatterplots made by triangles and diamonds), on the most numerous shapes in the image: if they thought there were more triangles, they had to press the right button; if they thought there were more diamonds, they had to press the left button. On the contrary, if they saw the two arrows, they had to judge the trend (descending or ascending) of the following trials and answer accordingly, analogously to the procedure described in study 1. Each stimulus remained on screen for 200 ms and then a fixation cross appeared for 1300 ms, signaling the response window. Participants were asked to give one answer within that time frame. After eight trials, the fixation cross remained on screen for 4, 6, or 8 seconds (on average for 6 seconds) and then a new block of trials started with the relevant instructions showing up, again, for 3000 ms. Run 1 and run 3 both comprised 20 blocks each (10 for shape detection and 10 for trend judgment). The order of the two tasks was, within each run, randomly determined and not known in advance by participants. The only difference between run 1 and run 3 was in the response-hand configuration: run 1 had the instructions presented as in figure 40, whereas run 3 had the opposite configuration of triangle/diamond and up arrow/down arrow, thus asking participants to respond with their right hand for diamond and left hand for triangle (shape detection task) or with their right hand for descending and left hand for ascending (trend judgment task). The duration of run 1 and 3 was of 420 seconds each.

Each scatterplot was created according to the same equation described in study 1. The experimental factors were: the slope of the generative regression (+ 0.5 or -0.5), the number of items (18 or 38), the noise level (0.05 or 0.15). Their combination resulted in 8 experimental conditions: one scatterplot per condition was presented in each block, in a random order, for a total of 4 ascending trends and 4 descending trends (participants were unaware of these distributions). Of these 8 scatterplots, each had one of the following ratios of triangles and diamonds (all ratios were presented in a given block): only triangles; 2 triangles over 18 or 4 over 38; 4 triangles over 18 or 8 over 38; 6 triangles over 18 or 12 over 38; only diamonds; 2 diamonds over 18 or 4 over 38; 4 diamonds over 18 or 8 over 38; 6 diamonds over 18 or 12 over 38 (to summarize, in each block there were 4 scatterplots with a majority of triangles and 4 with a majority of diamonds; again, participants were unaware of these distributions). A total of 160 trials was presented in each run. Each stimulus had a visual angle of 15°. Each triangle and diamond had the same number of pixels.

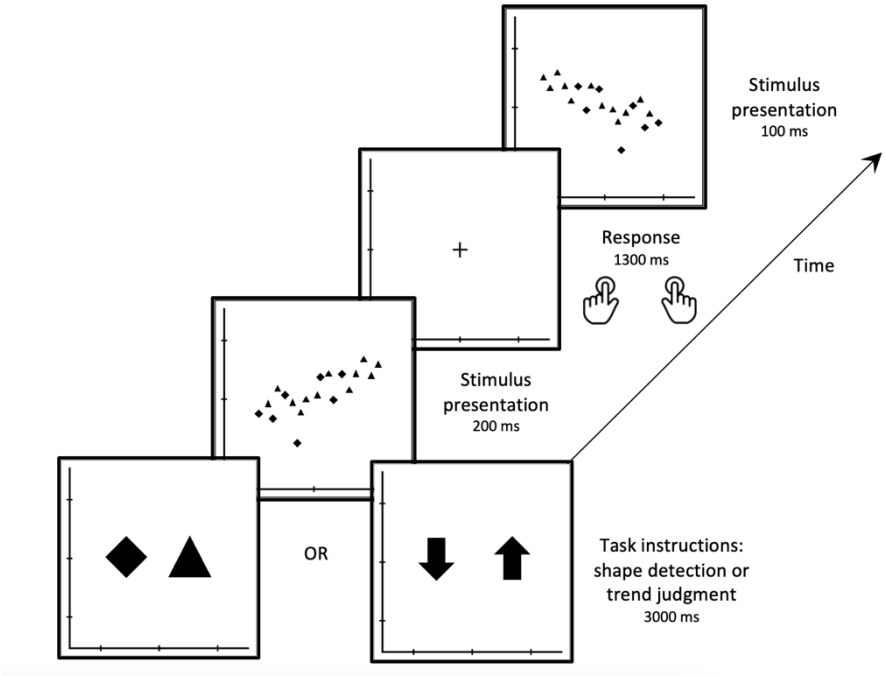


Figure 40. Experimental procedure of run 1 and run 3. Instructions were presented for 3000 ms, signaling whether participant should perform shape detection or trend judgment for the following eight trials. After a break of 6 seconds (on average), another block started again, with one of the two tasks' instructions.

Run 2 and run 4: trend judgment (event-related paradigm)

These two runs consisted uniquely in a trend judgment task (therefore, no mixture of triangles and diamonds was presented in the scatterplots, but only dots). The response-hand configuration was the same of run 1 (for run 2) and the same of run 3 (for run 4). The duration of the fixation cross' presentation was of either 3800, 5800 or 7800 ms (on average, 5800 ms). The hemifield in which the stimulus was presented could be either the left or the right. The duration of run 2 and 4 was of 774 s each.

Scatterplots were again generated according to the same algorithm of study 1. The experimental factors were: the slope of the generative function (+0.5 or -0.5); the number of points (18 or 38); the level of noise (0.05 or 0.15); the hemifield in which the scatterplot was presented (left or right). Their combination resulted in 16 experimental conditions; these 16 conditions were presented in a random order for 8 times during the run, for a total of 124 stimuli per run. Each stimulus had a visual angle of 15°.

Anatomical run

After the end of run 4, each participant was scanned for 7 minutes in order to obtain a precise anatomical image of their brains. This run was realized at this moment in order to allow participants to relax before the last 3 runs.

Run 5 and run 6: object identification and object orientation detection (blocked paradigm)

These two runs were structured identically to run 1 and run 3 but each trial consisted in the presentation of one object out of 6 possible objects (a knife, a fork, a spoon, a pen, a wrench and a brush) with 6 possible orientations (+15°, +30°, +45°, -15°, -30°, -45°). Two possible instructions were proposed to participants: they were either asked to identify the object category (if a "C" and a "A" appeared on screen, C standing for "cuisine" – *kitchen* in French; A standing for "autre" – *other* in French) or to judge the orientation of the object (if an up

arrow and a down arrow appeared on screen). Run 6 had the opposite response-hand configuration. In each block there were: in terms of object identity, 4 kitchen tools and 4 other types of tools; in terms of orientation, 4 upward and 4 downward objects (again, participants were unaware of these distributions). As for runs 1 and 3, the duration of runs 5 and 6 was of 420 seconds, comprising 160 trials.

Run 7: localizer of numerical cognition and language networks

The last run consisted in a series of language and mathematical tasks aimed at defining, for each participant, their numerical cognition and language networks. The precise procedure of this run is described by Pinel and collaborators (Pinel et al., 2004, 2007).

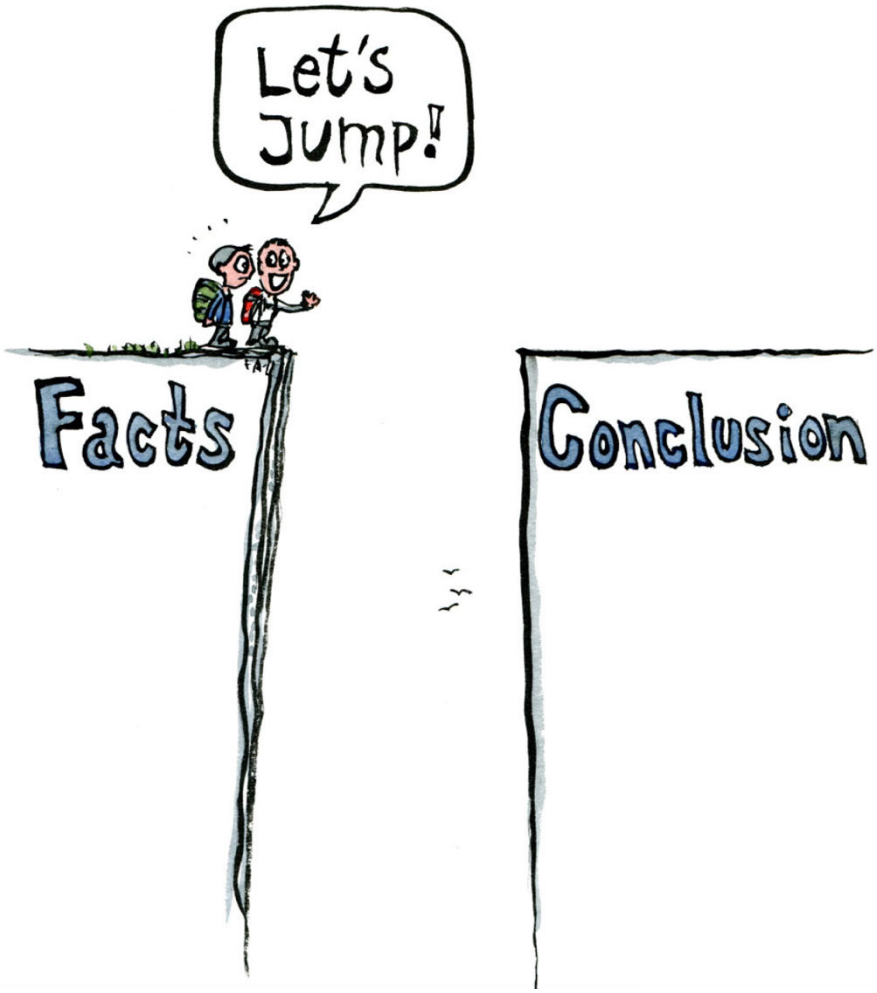
Participants

The experiment was advertised on the laboratory recruitment platform. Participants had to meet the following criteria in order to be considered eligible: being between 18 and 35 years old; being a university student or having completed at least three years of university studies (this criterion was added in order to ensure a relative homogeneity of the sample); not taking psychoactive drugs; not being pregnant; having normal or corrected to normal vision; being able to move their hands and fingers; not having mental implants in the body. 20 participants were recruited (12 women, 8 men; age: 25.2 $\pm$ 4.1). Half participants performed the experiment with the order of response-hand configurations described in the methods' section; half of them started with the opposite order (in order to control for possible effects of preferential response-hand configurations).

RESULTS

Analyses are currently being conducted: they will be presented at the defense.

CONCLUSION AND FUTURE RESEARCH DIRECTIONS



Frits Ahlefeldt

CONCLUSION

In this first section I summarize the results presented in the thesis and I briefly discuss them in relation to one another, with the aim to provide a general and summarized overview of the thesis' findings.

What are the precursors of graph perception?

I designed a novel trend judgment task over a noisy scatterplot ("does the graph goes up or down?") and I found that the probability of giving increasing responses is a sigmoid function of the t-value associated to the correlation in the graph. In other words, participants' performance was akin to an optimal statistical model based on the calculation of the entire set of parameters in the graph: the slope, the number of points and the level of noise, which are all necessary factors for the computation of the t-value formula that a statistician would calculate to determine the significance of a correlation in a scatterplot. I found that response times are also predicted by an accumulation of evidence model that takes the t-value as the decision variable. I replicated these findings in several populations: first, in a large sample of participants who performed the experiment online; second in the Himba, a Namibian people with little to no access to formal schooling; third, in 6-years-old 1st graders who never encountered a scatterplot in their school curriculum. I then operationalized each individual performance in such a trend judgment task as the slope of the psychometric function of the increasing responses over the *t-value* (what I called the "graphicacy index"). I lastly showed that such quantitative assessment of people graphicacy greatly varies in the large sample that performed the task online and tightly correlates with participants' self-evaluation of mathematical and statistical knowledge (but crucially not of native language mastery). Taken together, these findings show that the precursors of graph perception (operationalized as the

performance to a simple trend judgment task) are available early on in development and independently from formal education, but they can very likely improve with mathematical education. These findings echo the ones in the numerical cognition literature (Dehaene, 2011), showing that the number sense, while being universally available across different educational levels, ages, cultures, and even animal species, is still affects by (and predicts) mathematical achievements. In other words, they seem to provide preliminary evidence that graphical intuitions (analogously to the number sense) are not purely perceptual abilities but are somehow intertwined with mathematical understanding (that might help, for example, to better integrate the different factors of the graph). If such interrelation exists, we might wonder whether training graphical intuitions improves later mathematical performance (or, at the very least, basic statistical understanding). Investigating the potential benefits of such a training could have important implications for education, especially if we consider the relatively large proportion of the general population that does not know how to read a simple graphical representation (Galesic & Garcia-Retamero, 2011).

How accurately can humans perform a mental regression over a noisy graph?

With a novel line adjustment task, I asked participants to adjust a line over a briefly flashed scatterplot in order to investigate the precision of what I called human "mental regression". I found that participants are consistently biased (as compared with the response expected in a classic statistical framework) in their regression estimations: they do not minimize the vertical distance of the points to the fit (as it would have been predicted by a procedure akin to ordinary least squares regression) but rather the orthogonal distance (as done in Deming regression). In order to rule out the possibility that this finding derived from the experimental methodology used (i.e., the possibility that participants simply preferred moving the line

higher or lower than expected in certain conditions), I conducted a study based on an extrapolation task, in which participants were asked to locate a point beyond the presented scatterplot, as a continuation of the underlying trend. I found that participants' extrapolations perfectly mimic their line adjustments: they extrapolate points higher than expected (for positive trends) and lower than expected (for negative trends), and particularly for increasing levels of noise and larger numbers of points, thus confirming the existence of a Deming bias. As already partly discussed in chapter 2, the implications of the existence of such a bias are both applied and theoretical. First, the Deming bias might play an important role in several domains in which decisions are often based on intuitive graphical judgments (such as finance), and therefore help explaining over-optimistic and over-pessimistic behaviors (such as buying stocks too early or selling them too late) that are usually considered a default of our reasoning abilities rather than the result of our perceptual limitations (Kahneman, 2003; Kahneman et al., 1991). Second, they suggest that graphs are (at least partly) treated as objects: the Deming regression of a graph is mathematically equivalent to the principal axis of an object (which has been showed to be easily computed by humans in several behavioral tasks; e.g., Bodily et al., 2018; Lowet et al., 2018). Our remarkable ability to extract an object's principal axis might therefore have been recycled by our visual system in order to sustain the capacity to perform graphical judgments, analogously to the recycling of our object recognition system towards the development of our reading abilities (Dehaene, 2005). The cultural recycling hypothesis, if true, would also explain why certain types of graphs are so hard to grasp and, consequently, rarely used: for instance, human angle recognition system is prone to many misperceptions and, as a consequence, graphical systems that use angles to distinguish among occurrences or events in a plot (such as pie charts do) are rarely employed and often criticized (Kosara & Skau, 2016; Siirtola, 2019).

Is human mental regression resistant to outliers?

In another study, I investigated whether outliers are automatically excluded in mental regression judgments and, if not, whether paying increasing levels of attention towards them could make human statistical judgments more robust to their deviance. I found that, when participants are not informed about the presence of outliers, they spontaneously integrate them in their trend judgments and line adjustments: these findings are crucial because they suggest that graph perception operates in parallel on the entire set of observations. In fact, if it was simply another instance of ensemble perception, outliers should have been easily discarded, as several studies have shown (Avci & Boduroglu, 2021; Epstein & Emmanouil, 2021; Hochstein et al., 2018). Furthermore, I found that even when participants were informed about the presence of deviant observations in the graph, they were nevertheless biased towards them (although their bias decreased for very strong outliers). Based on these and other findings I proposed a model of outlier rejection based on the z-score of the datapoints in the scatterplot, which predicted both outlier detection and line adjustment. Overall, these findings confirm once more that intuitive judgments on noisy graphs seem to be well modeled by existing statistical models (t-value for trend judgments, orthogonal minimization for line adjustments, and z-score for outlier detection and rejection), showing how refined the perceptual precursors of our graph intuitions are.

Can humans extrapolate from non-linear trends?

I asked participants to extrapolate from noisy graphs generated from non-linear trends (namely: piecewise linear functions with early and late inflexions, sinusoids and quadratics) and I found that their performance was overall precise and, as suggested by the performance of a Bayesian ideal-observer model, likely dependent on the amount of evidence provided in

the observational range. I found that this was true with the one notable exception of quadratic functions, whose curvature was considerably underestimated. I therefore investigated another famous example of underestimation of accelerating functions: exponentials (Lammers et al., 2020; W. Wagenaar & Sagaria, 1975). By finely manipulating several conditions (i.e.: datapoints presented on linear or log scales; scatterplots being noisy or noiseless; extrapolations performed as numerical guesses or by pointing), I found that the exponential growth bias is strong but it considerably decreases when participants are asked to point instead of giving a numerical response, it disappears when data are displayed on a log scale (thus making the graph look linear), and it even reverses when the plotted function is noiseless. The bias also correlates with mathematical knowledge and, crucially, even in the pointing condition, which did not require any particular ability with numbers' representation and manipulation. This finding suggests once more, as for the trend judgment, that graph-based perceptual tasks are somehow intertwined with numerical cognition and/or mathematical understanding. A Bayesian observer showed the same underestimation bias for noisy graphs when it was modelled with a prior against exponentials; also, a larger prior better modelled the performance of participants with a low mathematical knowledge. Taken together, these biases suggest that the misperceptions described in the literature might derive from difficulties in extracting information from noise in graphical representations: I therefore proposed several evidence-based guidelines that could improve the way accelerating functions are displayed in plots.

FUTURE RESEARCH DIRECTIONS

I conclude by providing some research directions for future studies in the field of graph perception and graphicacy.

Testing the evidence-based suggestions derived from psychophysics

In this thesis I provided psychophysical results that can be used to design evidence-based guidelines to, for example, increase the detection of outliers in graphs or to better display exponential functions to the public. While these guidelines represent concrete proposals that could be easily implemented in data visualizations, the question of their utility remains. In other words, it would be important to test whether they translate into a real improvement of how people assign weight to outliers or how they forecast the evolution of exponential growths. This means that future research should both investigate whether the human statistical judgements improve when data are plotted on the basis of such guidelines and, most importantly, whether such improvement also leads to a better understanding of the underlying mathematical patterns. For example, the exponential growth bias is reduced after a short lecture on it (W. Wagenaar & Sagaria, 1975) but, when participants' performance is explored in detail, it is noticeable that they correct their bias with a compensation that is not proportional to the growth rate of the function: in other words, participants learned their bias (in a sense, they overcompensated it) but did not really extract the correct function. Graphs could be a useful tool to improve such training: for instance, correlational judgments improve after a long perceptual training on graphs (Cui et al., 2018). More broadly, it would be interesting to see whether being trained in statistical judgments on graphs extend beyond an improvement in the recognition of specific functions and statistical details and lead, for example, to a better understanding of statistics and mathematics in general.

The developmental learning trajectory of graph perception and understanding

Although children have recently been showed to perform very similarly to adults when it comes to make perceptual judgments over the physical dimensions that are needed to read and understand a graph (Panavas et al., 2022, in agreement with our findings on the trend judgment task), it has also been shown then they are quite impaired when asked to perform more complex statistical judgments over ensembles (Jones & Dekker, 2018). Future studies could investigate the evolution of graphical intuitions in children and, most importantly, what are the conceptual stages that have to be reached in order to attain a real understanding of graphs that goes beyond the perceptual/intuitive level. In other words, correctly perceiving and understanding a graph is likely to require the mastery of several non-trivial concepts, such as: oriented lines and points symbolize underlying trends; those trends are organized along lines that have a certain magnitude; each observation can concomitantly represent two quantities, as in the Cartesian plane; functional relations between dimensions allow to extrapolate future unobserved datapoints. In which order these stages are learned and mastered is an open question that would certainly merit further research work. In the same direction, it would be interesting to investigate the existence of (more or less specific) troubles of graphical understanding and whether they are separate from (or overlapping with) other learning disorders, such as dyscalculia and dyslexia.

The intuitive dictionary of mathematical functions

Our findings, together with several pieces of evidence from the literature on inductive biases (Schulz et al., 2015, 2017) and function learning (Brehmer, 1971; Lewandowsky et al., 2002; Lucas et al., 2015), defined several primitive functions that are available to human recognition and extrapolation and those, such as accelerating functions, that seem hard to grasp.

However, little is known about which and how many functions are indeed intuitively understandable by humans. Future research might thus undertake a catalogic effort to probe the limits of human intuitive dictionary of functions, at least when presented in the accessible format of graphical representation. To put it clearly, this effort would translate into asking the following questions: what is the full set of primitive functions that humans can recognize and/or learn and/or extrapolate from? What are the limits of human intuitive understanding on the number and type of combinations of such primitives into composite functions? Future studies might show, to both children and adults, all the existing mathematical functions expressed as noisy or noiseless graphs (both as primitives and in composition with one another) and ask them, for example, to draw their continuation towards a further boundary. Also, we could make (and test) the hypothesis that mathematical functions might be more or less easily recognized and learned as a function of the minimal description length (MDL) of the mental programs necessary to represent them: for example, a sinusoid would only necessitate to encode a rhythmic regular pattern having a fixed frequency and amplitude, whereas exponentials might need the representation of positive tendency, curvature and acceleration, thus making their MDL longer. These reflections are borrowed from the literature on the language of thought hypothesis, according to which music, geometry, sequences and math might be encoded and compressed based on the structure of their syntax (for a recent review: Dehaene et al., 2022).

Experimentally inducing priors in graph perception

Our comparisons of participants' performance with Bayesian optimal observers seem to suggest the existence of priors against accelerating functions in human mind. In other words, as we argued along the manuscript, people difficulty at extracting quadratics or exponentials

does not seem to depend on a lack of understanding of the function behavior itself: indeed, in studies 8 and 9 some extrapolation answers were in agreement with, respectively, quadratic and exponential growths. It rather seemed that their priors against acceleration made them less likely to select those functions when the level of noise made it hard to choose one option in their mental set of functional hypotheses. However, we did not experimentally test whether specific priors could actually affect participants' performance in trend judgment, line adjustment or extrapolation (as it has been recently done for correlation judgments: Xiong et al., 2022). Different priors could be elicited through different experimental manipulations: for example, the perception of a steeper linear regression might be induced through the indication of variables' labels that are known to be strongly correlated; or, participants could be informed beforehand about the behavior (and distribution) of non-liner functions to which they are going to be exposed; or, in the case of exponential noisy graphs, participants could be overly primed towards exponentials, possibly compensating their tendency to linearize accelerating trends.

The precursors of intuitive statistical judgments across species

In the field of numerical cognition, several studies have shown that its precursors (the so called "number sense") are available not just across several human populations but also across different animal species (K. E. Jordan et al., 2008; McComb et al., 1994; Rugani et al., 2007; Santolin et al., 2016; Versace et al., 2017). The cognitive precursors of intuitive statistical judgements might be similarly investigated in other non-human species. Indeed, there is a strong debate about the cognitive and neural mechanisms of ensemble perception and intuitive statistics (Whitney & Yamanashi Leib, 2018). One way to invigorate such debate would be to use a very simple animal model: newborn chicks of *Gallus gallus.* Immediately

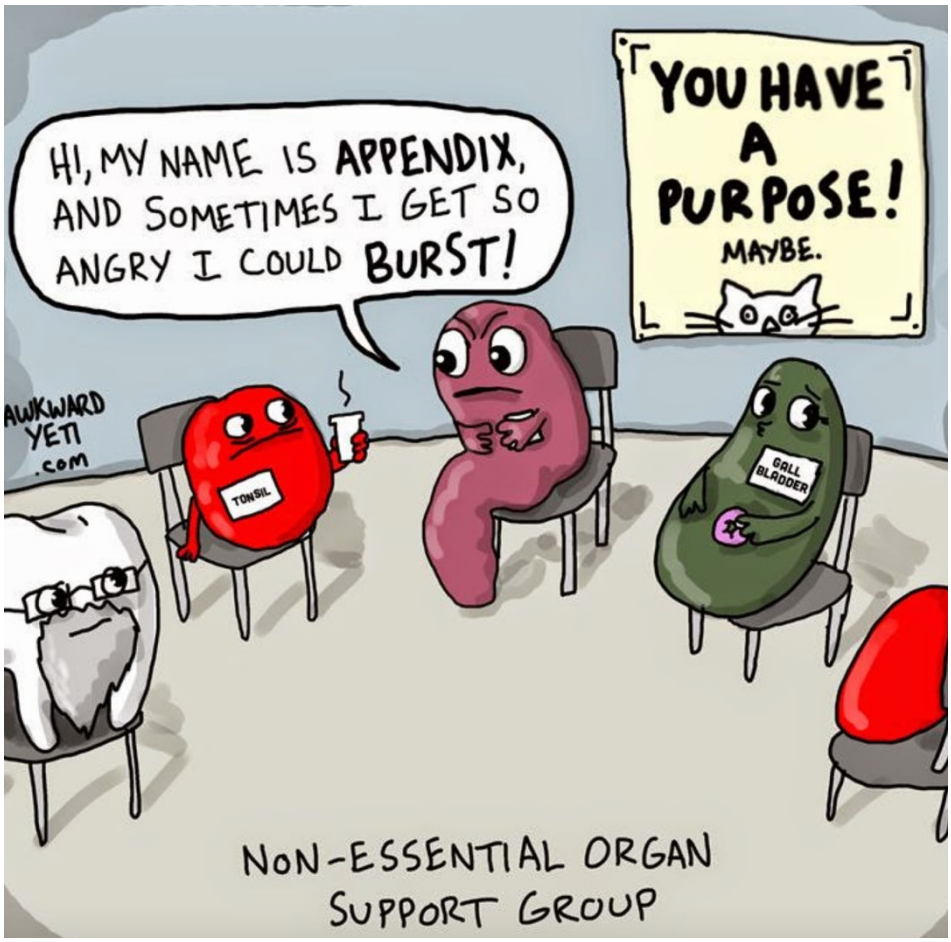
after birth (thus avoiding any possible learning mechanism) it is possible to imprint them on a given stimulus for a few days; immediately afterwards, they can be presented with two stimuli (one being the same of the imprinting phase) and test if they prefer to spend significantly more time with the stimulus they have been imprinted on; if so, this means that the chicks were able to encode and learn the stimulus' features. More interestingly, in order to test chicks' ability to learn and generalize the core features of their imprinting object, they can be presented, in the testing phase, with a stimulus deviating in terms of superficial features (such as presentation modalities, colors, position…) from the imprinting one. Using such a paradigm, newborn chicks could imprint on (and generalize from) many different stimuli, including specific numerosities and abstract patterns, thus showing that their generalization abilities go well beyond the encoding of simple and superficial stimulus' aspects. At present, however, no study has investigated whether chicks can be imprinted on (and thus generalize from) statistical summaries of ensembles, such as the regression trend underlying a series of points, as in the series of experiments presented in this thesis. Finding evidence for intuitive statistics' abilities in chicks might shed light on the phylogenetics of such remarkable skills and guide and inform research about the neural levels at which they are implemented in humans.

The meaning and interpretation of graphs

My thesis focused on the fundamental aspects of graph perception and understanding, as summarized by the three-steps distinction made in the introduction: data extraction, statistical inference and data forecasting. My findings, however, while characterizing people accuracy and bias in the statistical judgments I described, cannot extend to the final stage of graph understanding, which concerns the ability to use these statistical judgments to

interpret data and assign them a meaning. As an example, understanding the accelerating nature of the evolution of the number of cases of an epidemic illness is probably necessary but likely not sufficient to lead the reader to understand and/or memorize what such mathematical function entails. In other words, it would be interesting to investigate the degree at which graphical representations help the reader understanding, remembering and using the meaning behind the information they convey. Studies have shown what aspects attract people attention towards graphs (Borkin et al., 2016), which features and graph types improve their memorability (Borkin et al., 2013; Peña et al., 2020), and how graphs can improve decision making (Padilla et al., 2018) but, to date, no study specifically investigated whether a graph significantly improved people learning of the message derived from its interpretation. In fact, while we know that patterns are more easily recognized when shown in a graphical format, it is unclear whether such improvement in recognition translates into a better learning of the relation between the variables in the graph. Future studies could, for example, present the same piece of information in different modalities (e.g., tabular form, plain text, and graph) and then test which of them leads to the best learning performance. Importantly, graphs could take the form not just of scatterplots but also of all other types of charts and, consequently, we could also investigate the advantage in concepts/facts learning of one graph type over another.

APPENDICES



The Awkward Yeti

APPENDIX A: OPTIMAL DECISION MAKING IN THE TREND JUDGEMENT TASK (STUDY 1)

According to classical statistical theory (David & Neyman, 1938; Theil, 1971), given a dataset generated from a noisy linear function (i.e., *n* pairs of data points $\{x_i, y_i\}$, where $y_i = \alpha x_i + \varepsilon_i$ and the ε_i are independent centered Gaussian samples with standard deviation σ), the best linear unbiased estimator (BLUE) of the slope α is:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

According to the Gauss-Markov theorem, this estimator is unbiased (its mean is equal to α) and has minimum variance when compared to other possible estimators. Thus, it is the most appropriate estimator on which to base the trend judgment task. In order to optimally decide whether the trend in the graph is increasing or decreasing, an ideal observer should base its decision on whether the slope estimate $\hat{\alpha}$ is positive or negative. Assuming now that this is the decision strategy, can we predict how the associated error rate and response times should vary with experimental parameters? According to the tenets of classical signal detection theory (Green & Swets, 1966) and its extension to response times (e.g., Gold & Shadlen, 2002) the difficulty and error rate of such a decision is determined not only by the mean of this variable, but also by its distribution across trials. The standard error of the slope estimate $\hat{\alpha}$ is given by

$$s_{\hat{\alpha}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Under the null hypothesis, the ratio of the slope estimate to its standard error has a Student's t-distribution with *n-2* degrees of freedom (i.e. a distribution close to a Gaussian for large *n*). This *t-value* can also be written as

$$t = \frac{\widehat{\alpha}}{s_{\widehat{\alpha}}} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

where *r* is the Pearson coefficient of correlation, given by the covariance of the x and y values divided by the product of their standard deviations (for a comprehensive explanation, see Baguley, 2012):

$$r = \frac{cov(x,y)}{\sigma_x\ \sigma_y}$$

The t-value is the basis for the classical statistical test for significance of a linear trend: to decide if a non-null (positive or negative) trend is present, we compare the observed t-value to the Student's *t* distribution expected under the null hypothesis. Here, however, the situation is a bit different: as an experimenter, we know how the data was generated on each trial with a given slope α, which may be different from zero; and, to compute the psychometric function, we would like to know what is the probability that the decision maker will respond "the trend is increasing", assuming that the decision is based on whether $\widehat{\alpha}$ is greater than zero. Under these conditions, the t-value is no longer distributed as a Student's t-distribution (because the expected value of $\widehat{\alpha}$, being an unbiased estimator, is α). However, the following value, $t' = \frac{\widehat{\alpha}-\alpha}{s_{\widehat{\alpha}}}$, is again distributed as a Student's t-distribution. Thus, the probability of the "increasing" response is:

$$p_{increasing} = p(\hat{\alpha} > 0) = p(t' s_{\hat{\alpha}} + \alpha > 0) = p\left(t' > \frac{-\alpha}{s_{\hat{\alpha}}}\right) = p(t' > -t)$$

$$= \int_{-t}^{+\infty} Student_n(u)\, du = \int_{-\infty}^{t} Student_n(u)\, du$$

This equation indicates that the proportion of responses is an increasing function of the t-value. More specifically, it is a sigmoid-like function, namely the cumulative Student's *t*-distribution. Note that, strictly speaking, this function still depends on the number of points *n*. However, as *n* increases, it quickly becomes essentially indistinguishable from the integral of a Gaussian, and hence independent of n; it is also extremely similar to the classical sigmoid (see Gold & Shadlen, 2002).

The above theory, analogous to classical signal detection theory (SDT; Green & Swets, 1966), assumes that the decision is based on a single sample of *t*, and predicts only the psychometric function (or, equivalently, error rates) but not response times. To predict response times, we turn to a "sequential probability ratio test" variant of the above theory, according to which the decision-maker accumulates noisy samples of evidence about the sign of $\hat{\alpha}$, up to a fixed decision bound. Under such an accumulation-of-evidence model, according to the equations in (Gold & Shadlen, 2002), the psychometric response function becomes the classic sigmoid:

$$p_{increasing} = \frac{1}{1 + e^{-2B\mathrm{t}}}$$

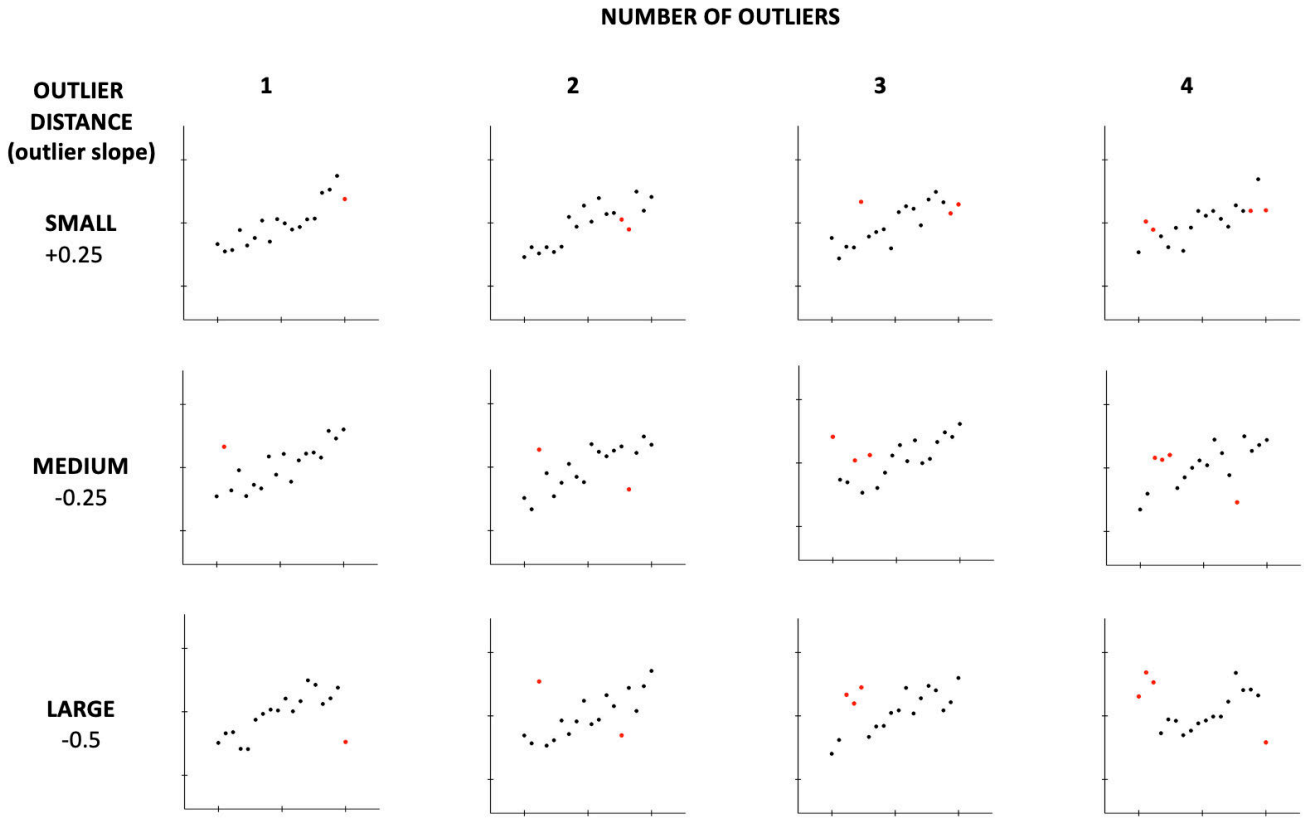
And the response time is predicted by the deviation of the absolute value of *t* from zero, according to a decreasing, convex upward function given by the equation:

$$RT = \frac{B}{|\mathrm{t}|} \tanh(B|\mathrm{t}|)$$

In those equations, *B* is a constant that jointly reflects both the sensitivity of the decision-maker (the amount of information accumulated per unit of time) and his decision threshold (controlling the speed/accuracy tradeoff). In summary, the theory predicts that decision difficulty (both error rates and RTs) should be determined by the t-value. The equation for *t*, in turn, makes it clear that the decision difficulty should depend on all manipulated graph parameters (n, σ and α), and predicts that the effects of these variables should be jointly summarized by a single effect of the t-value on behavior.

APPENDIX B: EXAMPLES OF STIMULI WITH OUTLIERS (STUDY 7)

The figure below provides examples of stimuli for the conditions with a main slope of +0.5. Outliers are signaled in red for readability purposes. Actual stimuli, as explained in the text, were white dots on a black background.



APPENDIX C: SUPPLEMENTARY STUDY DETAILS AND RESULTS (STUDY 9)

This appendix provides more details about the choice of the error measure used as dependent variable in study 9, together with additional results.

Choice of error measure

We used the difference in screen units (normalized to the range 0-1000) as an error measure since it offers a common currency for comparing conditions with different axes. Concretely, answers for stimuli presented on a log scale were translated into their corresponding values in screen units. Consider a participant making a numerical guess of 100 for an extrapolation from a log scale. If the correct answer was 1000, the numerical error would be 900. However, this corresponds, in terms of screen units, to the difference between two tick marks, i.e., to the same error that a participant would make if they answered 500 instead of a correct answer of 750 on the linear scale. Thus, in both cases, this example corresponds to an error of 250 screen units. Using screen units is natural since it assigns the same weight to the errors resulting (on a log scale) from answering 1000 instead of 100 and answering 100 instead of 10. For a log scale, this is analogous to calculating the ratio between the given answer and the correct one; however, using error ratios for extrapolations on a linear scale would be problematic since it assigns a different weight to the error resulting from a given answer of 750 instead of 500 and a given answer of 500 instead of 250: in the first case the error ratio is of 1,5 and in the second case it is of 2, although the overestimation bias is, for both cases, of 250 screen units.

An alternative error measure would be using the numerical difference between the given and the correct answer. In figure S1, results are plotted in such units. This measure is informative about absolute numerical distance from the correct value, but it makes impossible to quantitively compare errors for stimuli on different scales. When considering the actual

numbers given, answers' variability on a log scale increases relative to the linear scale: this is unsurprising given that, with a log scale, a very small change on screen can result in a very large change in number. However, the exponential underestimation bias vanishes for pointing responses (t(520)=-0.16, p=.87), and turns into an overestimation for number responses (t(520)=8.54, p<.0001; figure S1), which is reasonably low given the wide numerical extension of the log scale.

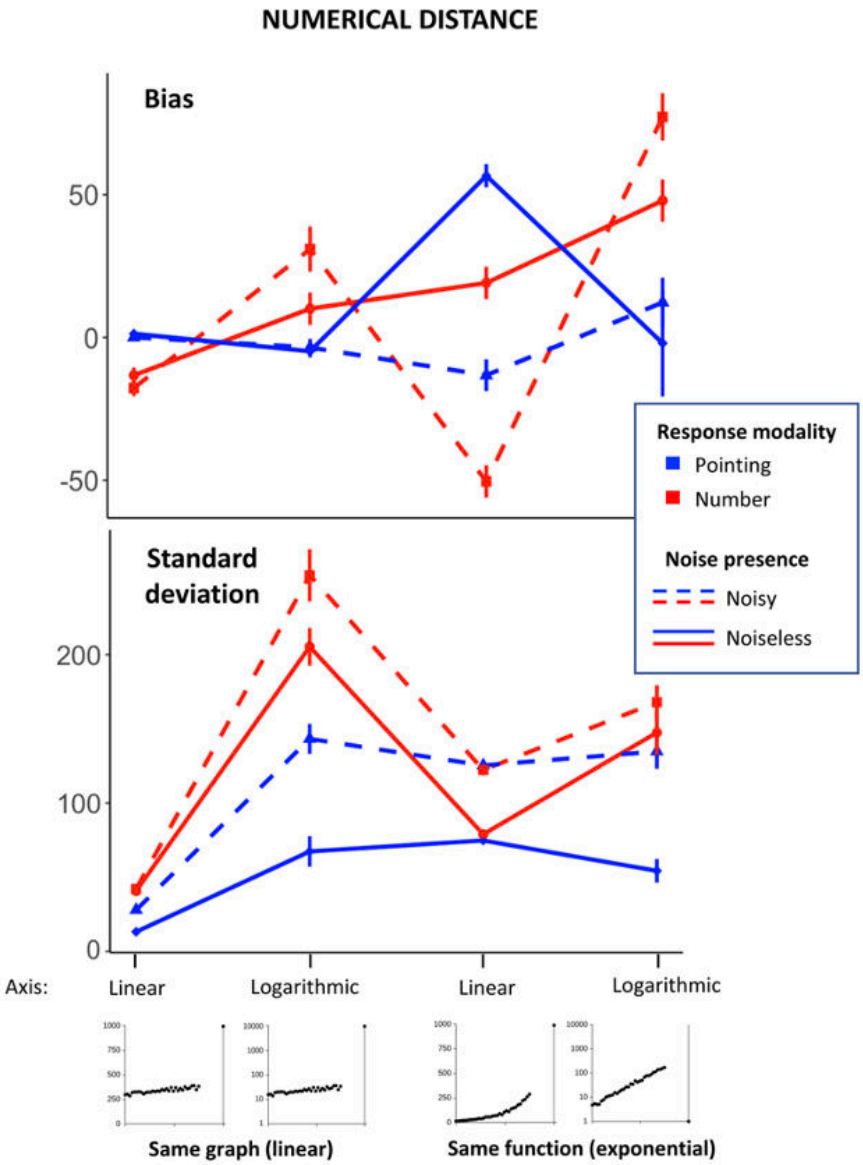


Figure S1. Mean bias and mean standard deviation of extrapolation responses, separately for experimental condition. Same format as figure 2, but the unit of measurement here is the difference between participants' **numerical** answers and the correct responses (in figure 2, we used screen units as measurement). For conditions with a linear scale, the plot therefore corresponds to figure 2. For conditions with a logarithmic scale, the higher bias and standard deviation are due to the much longer numerical interval covered by the y axis scale for log (range: 1-10000) than for linear scales (1-1000).

Anchoring effects for exponential extrapolations

For exponential functions on a linear scale, numerical responses were more likely than pointing responses to correspond to tick values (672/3126=21% versus 24/3126=1%; χ^2=676.79, df=1, p<.0001), and numerical responses corresponding to tens and hundreds were significantly more frequent than the expected frequency under the null hypothesis that each number had the same probability to be chosen by participants (binomial test, empirical proportion 2645/3126=85%, expected proportion 313/3126=10%, p<.0001). These findings replicate the anchoring effects found for linear functions. Combined, they suggest that for both linear and exponential functions, participants' numerical answers were biased towards round numbers.

Influence of mathematical knowledge on linear extrapolations

We also tested whether mathematical knowledge modulated the precision of linear extrapolations. Self-evaluated math knowledge did not correlate with participants' bias, for any experimental conditions (all p>.05), thus indicating that for linear functions, the bias remained null regardless of mathematical education. However, math knowledge negatively correlated with participants' response variability, confirming that participants with a higher mathematical education are more precise; this was true for both linear (*R*=-0.17, df=476, p<.001) and log scale (*R*=-0.13, df=476, p<.01), for both pointing (*R*=-0.09, df=476, p<.05) and numerical responses (*R*=-0.17, df=476, p<.001) and for both noiseless (*R*=-0.12, df=476, p<.01) and noisy functions (*R*=-0.2, df=476, p<.0001). Crucially, no correlation was found with self-evaluation of participants' first language skills (all related p>.05). These findings show that, even for simple linear scatterplots and even when just pointing to an extrapolated data point, participants' precision in responding is modulated by their math knowledge.

Influence of other factors on the exponential growth bias

The underestimation bias of noisy exponentials did not correlate with the highest degree obtained (in agreement with previous research: Levy & Tasoff, 2016), nor with age or gender. Other than with mathematical knowledge, the bias also correlated with familiarity with graphs and statistical knowledge and with participants' (reported) average math grade at university (all $p<.05$). Again, no correlation was found with the self-evaluated knowledge of first language.

Effect of task order

As explained in the methods' section, participants either started with the pointing task or the number task. To test for an effect of order on extrapolation performance, we conducted two ANOVAs (one on bias, the other on variability) with the task order as a between-subject factor. For the "same-graph" condition, no main effect of order was found for bias ($F[1,519]=1.04$, partial $\eta^2=.002$, $p=.31$), nor for variability ($F[1,519]=.12$,partial $\eta^2=.0002$, $p=.72$); no interaction effects of order with the other factors were found either (all $p>.05$). For the "same-function" condition, only a small interaction of order with axis scale was found for the bias ($F[1,519]=5.24$, partial $\eta^2=.01$, $p<.05$). In terms of variability, an interaction effect of order with response modality was found ($F[1,519]=13.13$, partial $\eta^2=.02$, $p<.001$): indeed, when participants started with the pointing task, their average variability (61.3) remained relatively stable for the number task (63.1); however, when they started with the number task, the average variability (69.7) shrunk to 56.5 in the pointing task. Plausibly, performing numerical guesses requires a higher level of graphical analysis and might have pushed participants to be more precise in the subsequent pointing task.

APPENDIX D: ANOVA TABLES FOR BIAS AND VARIABILITY (STUDY 9)

Effect	df	F	p	Partial η^2
Response modality	1, 520	.09	.77	.0002
Noise	1, 520	9.32	<.01	.02
Axis scale	1, 520	152.24	<.0001	.23
Response modality x Noise	1, 520	1.53	.22	.003
Response modality x Axis scale	1, 520	167.41	<.0001	.24
Noise x Axis scale	1, 520	.07	.79	.0001
Response modality x Axis scale x Noise	1, 520	.24	.63	.0005

Table 1A. Repeated measures omnibus ANOVA on extrapolation bias in the "same-graph" conditions.

Effect	df	F	P	Partial η^2
Response modality	1, 520	241.11	<.0001	.32
Noise	1, 520	72.12	<.0001	.12
Axis scale	1, 520	9.44	<.01	.02
Response modality x Noise	1, 520	51.80	<.0001	.09
Response modality x Axis scale	1, 520	6.52	.01	.01
Noise x Axis scale	1, 520	.12	.73	.0002
Response modality x Axis scale x Noise	1, 520	.002	.97	<.0001

Table 1B. Repeated measures omnibus ANOVA on extrapolation variability in the "same-graph" conditions.

Effect	df	F	p	Partial η^2
Response modality	1, 520	18.82	<.0001	.03
Noise	1, 520	229.54	<.0001	.31
Axis scale	1, 520	1.02	.31	.002
Response modality × Noise	1, 520	.002	.97	<.0001
Response modality × Axis scale	1, 520	105.4	<.0001	.17
Noise × Axis scale	1, 520	285.95	<.0001	.35
Response modality × Axis scale × Noise	1, 520	.002	.96	<.0001

Table 2A. Repeated measures omnibus ANOVA on extrapolation bias (for "same-function" conditions).

Effect	df	F	p	Partial η^2
Response modality	1, 520	20.93	<.0001	.04
Noise	1, 520	315.26	<.0001	.38
Axis scale	1, 520	1615.38	<.0001	.76
Response modality × Noise	1, 520	7.06	<.01	.01
Response modality × Axis scale	1, 520	20.06	<.0001	.04
Noise × Axis scale	1, 520	126.55	<.0001	.20
Response modality × Axis scale × Noise	1, 520	.04	.83	<.0001

Table 2B. Repeated measures omnibus ANOVA on extrapolation variability (for "same-function" conditions).

REFERENCES

Akaike, H. (1998). *Information theory and an extension of the maximum likelihood principle.* Springer, New York, NY.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003

Andreassen, P. B., & Kraus, S. J. (1990). Judgmental extrapolation and the salience of change. *Journal of Forecasting*, *9*(4), 347–372. https://doi.org/10.1002/for.3980090405

Anscombe, F. J. (1960). Rejection of Outliers. *Technometrics*, *2*(2), 124–146.

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266.

Astle, D. E., Nobre, A. C., & Scerif, G. (2010). Subliminally Presented and Stored Objects Capture Spatial Attention. *Journal of Neuroscience*, *30*(10), 3567–3571. https://doi.org/10.1523/JNEUROSCI.5701-09.2010

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, *21*(4), 547–569. https://doi.org/10.1017/S0140525X98001277

Avci, B., & Boduroglu, A. (2021). Contributions of ensemble perception to outlier representation precision. *Attention, Perception, & Psychophysics*, *83*(3), 1141–1151. https://doi.org/10.3758/s13414-021-02270-9

Ayzenberg, V., & Lourenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports*, *9*(1), 9359. https://doi.org/10.1038/s41598-019-45268-y

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences.* Macmillan International Higher.

Balchin, W. G. V., & Coleman, A. M. (1966). Graphicacy should be the fourth ace in the pack. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *3*(1), 23–28. https://doi.org/10.3138/C7Q0-MM01-6161-7315

Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, V., & Lawson, J. (2001). Are intuitive physics and intuitive psychology independent? A test with children with Asperger Syndrome. *Journal of Developmental and Learning Disorders*, *5*(1), 47–78.

Barthelmé, S., & Mamassian, P. (2009). Evaluation of Objective Uncertainty in the Visual System. *PLoS Computational Biology*, *5*(9), e1000504. https://doi.org/10.1371/journal.pcbi.1000504

Beattie, V., & Jones, M. J. (2002). The Impact of Graph Slope on Rate of Change Judgments in Corporate Reports. *Abacus*, *38*(2), 177–199. https://doi.org/10.1111/1467-6281.00104

Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalite causee par la petite verole, et des avantages de l'inoculation pour la prevenir. *Histoire et Mémoires de l'Académie Royale Des Sciences de Paris*, 1–45.

Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., & Zorzi, M. (2010). Numerical estimation in preschoolers. *Developmental Psychology*, *46*(2), 545–551. https://doi.org/10.1037/a0017887

Bertin, J. (1967). *Sémiologie graphique.* (Mouton/Gauthier-Villars).

Bjorklund, D. F. (2014). Children, childhood, and development in evolutionary perspective. *Developmental Review*, 40.

Blackman, V. H. (1919). The Compound Interest Law and Plant Growth. *Annals of Botany*, *os-33*(3), 353–360. https://doi.org/10.1093/oxfordjournals.aob.a089727

Blum, H. (1967). A transformation for extracting new descriptors of shape. *Cambridge: MIT Press*, *4*.

Blum, H. (1973). Biological shape and visual science (part I). *Journal of Theoretical Biology*, *38*(2), 205–287. https://doi.org/10.1016/0022-5193(73)90175-6

Bobko, P., & Karren, R. (1979). The perception of Pearson product moment correlations from bivariate scatterplots. *Personnel Psychology*, *32*(2), 313–325. https://doi.org/10.1111/j.1744-6570.1979.tb02137.x

Bodily, K. D., Eastman, C. K., & Sturz, B. R. (2011). Neither by global nor local cues alone: Evidence for a unified orientation process. *Animal Cognition*, *14*(5), 665–674. https://doi.org/10.1007/s10071-011-0401-x

Bodily, K. D., Sullens, D. G., Price, S. J., & Sturz, B. R. (2018). Testing principal- versus medial-axis accounts of global spatial reorientation. *Journal of Experimental Psychology: Animal*

Learning and Cognition, *44*(2), 209–215. https://doi.org/10.1037/xan0000162

Bolger, F., & Harvey, N. (1993). Context-Sensitive Heuristics in Statistical Reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *46*(4), 779–811. https://doi.org/10.1080/14640749308401039

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, *42*(1), 189–201. https://doi.org/10.1037/0012-1649.41.6.189

Borkin, M. A., Bylinskii, Z., Kim, N. W., Bainbridge, C. M., Yeh, C. S., Borkin, D., Pfister, H., & Oliva, A. (2016). Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 519–528. https://doi.org/10.1109/TVCG.2015.2467732

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What Makes a Visualization Memorable? *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2306–2315. https://doi.org/10.1109/TVCG.2013.234

Bott, L., & Heit, E. (2004). Nonmonotonic Extrapolation in Function Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 38–50. https://doi.org/10.1037/0278-7393.30.1.38

Boutsen, L., & Marendaz, C. (2001). Detection of shape orientation depends on salient axes of symmetry and elongation: Evidence from visual search. *Perception & Psychophysics*, *63*(3), 404–422. https://doi.org/10.3758/BF03194408

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science*, *24*(6), 259–260. https://doi.org/10.3758/BF03328999

Cant, J. S., & Xu, Y. (2012). Object Ensemble Processing in Human Anterior-Medial Ventral Visual Cortex. *Journal of Neuroscience*, *32*(22), 7685–7700. https://doi.org/10.1523/JNEUROSCI.3325-11.2012

Cant, J. S., & Xu, Y. (2020). One bad apple spoils the whole bushel: The neural basis of outlier processing. *NeuroImage*, *211*, 116629. https://doi.org/10.1016/j.neuroimage.2020.116629

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*(2), i–144. https://doi.org/10.1002/j.2333-8504.1963.tb00958.x

Chaloner, K., & Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, 651–659.

Cheng, K. (2005). Reflections on geometry and navigation. *Connection Science*, *17*(1–2), 5–21. https://doi.org/10.1080/09540090500138077

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Res*, *43*(4), 393–404.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Res*, *45*(7), 891–900.

Ciccione, L., & Dehaene, S. (2020). Grouping Mechanisms in Numerosity Perception. *Open Mind*, *4*, 102–118. https://doi.org/10.1162/opmi_a_00037

Ciccione, L., & Dehaene, S. (2021). Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots. *Cognitive Psychology*, *128*, 101406. https://doi.org/10.1016/j.cogpsych.2021.101406

Ciccione, L., & Dehaene, S. (2022). Graphicacy skills across ages and cultures: a new assessment tool of intuitive statistics' abilities. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).

Ciccione, L., Dehaene, G., & Dehaene, S. (2022). Outlier detection and rejection in scatterplots: Do outliers influence intuitive statistical judgments? *Journal of Experimental Psychology: Human Perception and Performance.* https://doi.org/10.1037/xhp0001065

Ciccione, L., Sablé-Meyer, M., & Dehaene, S. (2022). Analyzing the misperception of exponential growth in graphs. *Cognition*, *225*, 105112. https://doi.org/10.1016/j.cognition.2022.105112

Ciccione, L., Sable-Meyer, M., Boissin, E., Josserand, M., Potier-Watkins, C., Caparos, S., & Dehaene, S. (2022). Graphicacy across age, education, and culture: a new tool to assess intuitive graphics skills. *bioRxiv*.

Cleveland, W., Diaconis, P., & Mcgill, R. (1982). Variables on Scatterplots Look More Highly Correlated When the Scales Are Increased. *Science*, *216*(4550), 1138–1141. https://doi.org/10.1126/science.216.4550.1138

Cleveland, W., & McGill, R. (1985). Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science, New Series*, *229*(4716), 828–833.

Cohen, E. H., & Singh, M. (2006a). Perceived orientation of complex shape reflects graded part decomposition. *Journal of Vision*, *6*(8), 4–

4. https://doi.org/10.1167/6.8.4
Cohen, E. H., & Singh, M. (2006b). Perceived orientation of complex shape reflects graded part decomposition. *Journal of Vision*, *6*(8), 4. https://doi.org/10.1167/6.8.4
Concilio, G., Pucci, P., Raes, L., & Mareels, G. (Eds.). (2021). *The Data Shake: Opportunities and Obstacles for Urban Policy Making*. Springer International Publishing. https://doi.org/10.1007/978-3-030-63693-7
Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, *23*(4), 351–361. https://doi.org/10.1080/00401706.1981.10487680
Correll, M., Bertini, E., & Franconeri, S. (2020). Truncating the Y-Axis: Threat or Menace? *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376222
Correll, M., & Heer, J. (2017). Regression by Eye: Estimating Trends in Bivariate Visualizations. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 1387–1396. https://doi.org/10.1145/3025453.3025922
Cui, L., & Liu, Z. (2021). Synergy between research on ensemble perception, data visualization, and statistics education: A tutorial review. *Attention, Perception, & Psychophysics*, *83*(3), 1290–1311. https://doi.org/10.3758/s13414-020-02212-x
Cui, L., Massey, C. M., & Kellman, P. J. (2018). Perceptual Learning in Correlation Estimation: The Role of Learning Category Organization. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 7.
Cuijpers, R. H., Smeets, J. B. J., & Brenner, E. (2004). On the Relation Between Object Shape and Grasping Kinematics. *Journal of Neurophysiology*, *91*(6), 2598–2606. https://doi.org/10.1152/jn.00644.2003
Curcio, F. R. (1987). Comprehension of Mathematical Relationships Expressed in Graphs. *Journal for Research in Mathematics Education*, *18*(5), 382. https://doi.org/10.2307/749086
David, F. N., & Neyman, J. (1938). Extension of the Markoff Theorem on Least Squares. *Statistical Research Memoirs*, *2*, 105–116.
de Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, *70*(5), 789–794. https://doi.org/10.3758/PP.70.5.789
de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12. https://doi.org/10.3758/s13428-015-0567-2
Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. In *From monkey brain to human brain* (p. 33).
Dehaene, S. (2009). *Reading in the brain*. Penguin Viking.
Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: A hypothesis about human singularity. *Trends in Cognitive Sciences*, *26*(9), 751–766. https://doi.org/10.1016/j.tics.2022.06.010
Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371–396. https://doi.org/10.1037/0096-3445.122.3.371
Dehaene, S., & Cohen, L. (2007). Cultural Recycling of Cortical Maps. *Neuron*, *56*(2), 384–398. https://doi.org/10.1016/j.neuron.2007.10.004
Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006). Core knowledge of geometry in an Amazonian indigene group. *Science*, *311*, 381–384.
Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or Linear? Distinct Intuitions of the Number Scale in Western and Amazonian Indigene Cultures. *Science*, *320*(5880), 1217–1220.
Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.
Dehaene, Stanislas. (2011). *The number sense: How the mind creates mathematics*. OUP USA.
DeLosh, E. L., McDaniel, M. A., & Busemeyer, J. R. (1997). Extrapolation: The Sine Qua Non for Abstraction in Function Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968–986.
Deming, W. E. (1943). Statistical adjustment of data. *Wiley*.
Dietz, K. (1967). Epidemics and Rumours: A Survey. *Journal of the Royal Statistical Society. Series A (General)*, *130*(4), 505–528.
Dillon, M. R., Kannan, H., Dean, J. T., Spelke, E. S., & Duflo, E. (2017). Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics. *Science*, *357*(6346), 47–55. https://doi.org/10.1126/science.aal4724
Dixon, B. E., Grannis, S. J., McAndrews, C., Broyles, A. A., Mikels-Carrasco, W., Wiensch, A., Williams, J. L., Tachinardi, U., & Embi, P. J. (2021). Leveraging data visualization and a statewide health information exchange to support

COVID-19 surveillance and response: Application of public health informatics. *Journal of the American Medical Informatics Association*, ocab004. https://doi.org/10.1093/jamia/ocab004

Driver, J., Baylis, G. C., Goodrich, S. J., & Rafal, R. D. (1994). Axis-based neglect of visual shapes. *Neuropsychologia*, *32*(11), 1353–1356. https://doi.org/10.1016/0028-3932(94)00068-9

Eggleton, I. R. C. (1982). Intuitive Time-Series Extrapolation. *Journal of Accounting Research*, *20*(1), 68. https://doi.org/10.2307/2490763

Epstein, M. L., & Emmanouil, T. A. (2021). Ensemble Statistics Can Be Available before Individual Item Properties: Electroencephalography Evidence Using the Oddball Paradigm. *Journal of Cognitive Neuroscience*, *33*(6), 1056–1068. https://doi.org/10.1162/jocn_a_01704

Epstein, M. L., Quilty-Dunn, J., Mandelbaum, E., & Emmanouil, T. A. (2020). The outlier paradox: The role of iterative ensemble coding in discounting outliers. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(11), 1267–1279. https://doi.org/10.1037/xhp0000857

Espinoza, F. (2005). An analysis of the historical development of ideas about motion and its implications for teaching. *Physics Education*, *40*(2), 139–146. https://doi.org/10.1088/0031-9120/40/2/002

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659–676. https://doi.org/10.1037/0033-295X.107.4.659

Finney, D. J. (1951). Subjective Judgment in Statistical Analysis: An Experimental Study. *Journal of the Royal Statistical Society. Series B (Methodological)*, *13*(2), 284–297.

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The Science of Visual Data Communication: What Works. *Psychological Science in the Public Interest*, *22*(3), 110–161. https://doi.org/10.1177/15291006211051956

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. *Journal for Research in Mathematics Education*, *32*(2), 124. https://doi.org/10.2307/749671

Friendly, M., & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, *41*(2), 103–130. https://doi.org/10.1002/jhbs.20078

Galesic, M., & Garcia-Retamero, R. (2011). Graph Literacy: A Cross-Cultural Comparison. *Medical Decision Making*, *31*(3), 444–457. https://doi.org/10.1177/0272989X10373805

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Gleicher, M., Correll, M., Nothelfer, C., & Franconeri, S. (2013). Perception of Average Value in Multiclass Scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2316–2325. https://doi.org/10.1109/TVCG.2013.183

Godau, C., Vogelgesang, T., & Gaschler, R. (2016). Perception of bar graphs – A biased impression? *Computers in Human Behavior*, *59*, 67–73. https://doi.org/10.1016/j.chb.2016.01.036

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36(2)*, 299–308.

Gray Funkhouser, H. (1937). *Historical Development of the Graphical Representation of Statistical Data*. The University of Chicago Press.

Green, D., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Krieger Publishing Company.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception & Psychophysics*, *72*(7), 1825–1838. https://doi.org/10.3758/APP.72.7.1825

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences*, *109*(28), 11116–11120. https://doi.org/10.1073/pnas.1200196109

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668. https://doi.org/10.1038/nature07246

Harold, J., Lorenzoni, I., Shipley, T. F., & Coventry, K. R. (2016). Cognitive and psychological science insights to improve climate change data visualization. *Nature Climate Change*, *6*(12), 1080–1089. https://doi.org/10.1038/nclimate3162

Haroz, S., Kosara, R., & Franconeri, S. L. (2016). The Connected Scatterplot for Presenting Paired Time Series. *IEEE Transactions on Visualization and Computer Graphics*, *22*(9), 2174–2186. https://doi.org/10.1109/TVCG.2015.2502587

Harrison, L., Yang, F., Franconeri, S., & Chang, R. (2014). Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1943–1952.

https://doi.org/10.1109/TVCG.2014.2346979
Harvey, N., Ewart, T., & West, R. (1997). Effects of Data Noise on Statistical Judgement. *Thinking & Reasoning*, *3*(2), 111–132. https://doi.org/10.1080/135467897394383
Hawkins, D. M. (1980). *Identification of outliers*. London: Chapman and Hall.
Heer, J., & Bostock, M. (2010). Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. *ACM Human Factors in Computing Systems (CHI)*.
Hochstein, S., Pavlovskaya, M., Bonneh, Y. S., & Soroker, N. (2018). Comparing set summary statistics and outlier pop out in vision. *Journal of Vision*, *18*(13), 12. https://doi.org/10.1167/18.13.12
Hong, M.-H., Witt, J. K., & Szafir, D. A. (2021). The Weighted Average Illusion: Biases in Perceived Mean Position in Scatterplots. *ArXiv:2108.03766 [Cs]*. http://arxiv.org/abs/2108.03766
Hsieh, P.-J., Colas, J. T., & Kanwisher, N. (2011). Pop-Out Without Awareness: Unseen Feature Singletons Capture Attention Only When Top-Down Attention Is Available. *Psychological Science*, *22*(9), 1220–1226. https://doi.org/10.1177/0956797611419302
Hutzler, F., Richlan, F., Leitner, M. C., Schuster, S., Braun, M., & Hawelka, S. (2021). Anticipating trajectories of exponential growth. *Royal Society Open Science*, *8*(4), rsos.201574, 201574. https://doi.org/10.1098/rsos.201574
Izard, V., Pica, P., Spelke, E. S., & Dehaene, S. (2011). Flexible intuitions of Euclidean geometry in an Amazonian indigene group. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(24), 9782–9787. https://doi.org/10.1073/pnas.1016686108
Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166.
Jenkinson, J. W. (1906). On the Relation Between the Symmetry of the Egg and the Symmetry of the Embryo in the Frog (Rana Temporaria). *Biometrika*, 22.
Jones, P. R., & Dekker, T. M. (2018). The development of perceptual averaging: Learning what to do, not just how to do it. *Developmental Science*, *21*(3), e12584. https://doi.org/10.1111/desc.12584
Jordan, K. E., MacLean, E. L., & Brannon, E. M. (2008). Monkeys match and tally quantities across senses. *Cognition*, *108*(3), 617–625. psyh. https://doi.org/10.1016/j.cognition.2008.05.006
Jordan, N., & Dyson, N. (2016). Catching Math Problems Early: Findings From the Number Sense Intervention Project. *Continuous Issues in Numerical Cognition, Academic Press*, 59–79.
Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review*, *114*(3), 678–703. https://doi.org/10.1037/0033-295X.114.3.678
Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, *93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392
Kahneman, D., Knetsch, J., & Thaler, R. (1991). Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal of Economic Perspectives*, *5.1*, 193–206.
Kalish, M. L. (2013). Learning and extrapolating a periodic function. *Memory & Cognition*, *41*(6), 886–896. https://doi.org/10.3758/s13421-013-0306-9
Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294. https://doi.org/10.3758/BF03194066
Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General*, *131*(2), 287–297. https://doi.org/10.1037/0096-3445.131.2.287
Kelly, D. M., & Durocher, S. (2011). Comparing geometric models for orientation. *Communicative & Integrative Biology*, *4*(6), 710–712.
Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, *18*(9), 23. https://doi.org/10.1167/18.9.23
Kinchla, R. A. (1977). The role of structural redundancy in the perception of visual targets. *Perception & Psychophysics*, *22*(1), 19–30. https://doi.org/10.3758/BF03206076
Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics*, *63*(8), 1421–1455. https://doi.org/10.3758/BF03194552
Knuth, K. H. (2019). Optimal data-based binning for histograms and histogram-based probability density models. *Digital Signal Processing*, *95*, 102581. https://doi.org/10.1016/j.dsp.2019.102581
Kosara, R., & Skau, D. (2016). Judgment Error in Pie Chart Variations. *EuroVis 2016 - Short Papers*, 5 pages.

https://doi.org/10.2312/EUROVISSHORT.20161167
Kosslyn, S. M., & Kosslyn, S. M. (2006). *Graph Design for the Eye and Mind*. Oxford University Press, USA.
Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: A critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 857–875. https://doi.org/10.1098/rstb.2007.2093
Kovacs, I., & Julesz, B. (1994). Perceptual sensitivity maps within globally defined visual shapes. *Nature*, *370*(6491), 644–646. https://doi.org/10.1038/370644a0
Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive Physics: Current Research and Controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759. https://doi.org/10.1016/j.tics.2017.06.002
Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research*, *33*(2), 188–229. https://doi.org/10.1177/0049124103262065
Kurtz, A. K., & Edgerton, H. A. (1939). *Statistical Dictionary of Terms and Symbols* (Hafner Publishing Company).
Lammers, J., Crusius, J., & Gast, A. (2020). Correcting misperceptions of exponential coronavirus growth increases support for social distancing. *Proceedings of the National Academy of Sciences*, *117*(28), 16264–16266.
Lane, D. M., Anderson, C. A., & Kellam, K. L. (1985). Judging the Relatedness of Variables: The Psychophysics of Covariation Detection. *Journal of Experimental Psychology: Human Perception and Performance*, *11(5)*, 640.
Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25years. *International Journal of Forecasting*, *22*(3), 493–518. https://doi.org/10.1016/j.ijforecast.2006.03.007
Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, *43*(2), 172–187. https://doi.org/10.1016/0749-5978(89)90049-6
Levy, M. R., & Tasoff, J. (2016). Exponential-Growth Bias and Lifecycle Consumption. *Journal of the European Economic Association*, *14*(3), 545–583.
Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General*, *131*(2), 163–193. https://doi.org/10.1037/0096-3445.131.2.163
Linnet, K. (1998). Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry*, *44*(5), 1024–1031. https://doi.org/10.1093/clinchem/44.5.1024
Little, D. R., & Shiffrin, R. M. (2009). Simplicity Bias in the Estimation of Causal Functions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 7.
Liu, T., Li, X., Bao, C., Correll, M., Tu, C., Deussen, O., & Wang, Y. (2021). Data-Driven Mark Orientation for Trend Estimation in Scatterplots. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445751
Lowet, A. S., Firestone, C., & Scholl, B. J. (2018). Seeing structure: Shape skeletons modulate perceived similarity. *Attention, Perception, & Psychophysics*, *80*(5), 1278–1289. https://doi.org/10.3758/s13414-017-1457-8
Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215. https://doi.org/10.3758/s13423-015-0808-5
Ludewig, U., Lambert, K., Dackermann, T., Scheiter, K., & Möller, K. (2020). Influences of basic numerical abilities on graph reading performance. *Psychological Research*, *84*(5), 1198–1210. https://doi.org/10.1007/s00426-019-01144-y
Ma, J. (2020). Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, *5*, 129–141. https://doi.org/10.1016/j.idm.2019.12.009
Maino, J. L., & Kearney, M. R. (2015). Testing mechanistic models of growth in insects. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1819), 20151973. https://doi.org/10.1098/rspb.2015.1973
Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, *26*(1), 147–155. https://doi.org/10.1017/S0952523808080905
Mamassian, P., Landy, M., & Maloney, L. T. (2001). Bayesian Modelling of Visual Perception. *Probabilistic Models of the Brain: Perception and Neural Function*, 21.
Martin, R. F. (2000). General Deming Regression for Estimating Systematic Bias and Its Confidence Interval in Method-Comparison Studies. *Clinical Chemistry*, *46*(1), 100–104. https://doi.org/10.1093/clinchem/46.1.100
McCloskey, M. (1983). Intuitive Physics. *Scientific*

American, 11.
McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 636–649. https://doi.org/10.1037/0278-7393.9.4.636
McComb, K., Packer, C., & Pusey, A. (1994). Roaring and numerical assessment in contests between groups of female lions, Panthera leo.. *Animal Behaviour*, *47*(2), 379–387. psyh. https://doi.org/10.1006/anbe.1994.1052
Mcdaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*(1), 24–42.
McElroy, F. W. (1967). A Necessary and Sufficient Condition That Ordinary Least-Squares Estimators Be Best Linear Unbiased. *Journal of the American Statistical Association*, *62*(320), 1302–1304. https://doi.org/10.1080/01621459.1967.10500935
Menge, D. N. L., MacPherson, A. C., Bytnerowicz, T. A., Quebbeman, A. W., Schwartz, N. B., Taylor, B. N., & Wolf, A. A. (2018). Logarithmic scales in ecological data presentation may cause misinterpretation. *Nature Ecology & Evolution*, *2*(9), 1393–1402. https://doi.org/10.1038/s41559-018-0610-7
Meyer, J., & Shinar, D. (1992). Estimating Correlations from Scatterplots. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*(3), 335–349. https://doi.org/10.1177/001872089203400307
Meyer, J., Taieb, M., & Flascher, I. (1997). Correlation Estimates as Perceptual Judgments. *Journal of Experimental Psychology: Applied*, *3(1)*(3).
Micallef, L., Palmas, G., Oulasvirta, A., & Weinkauf, T. (2017). Towards Perceptual Optimization of the Visual Design of Scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, *23*(6), 1588–1599. https://doi.org/10.1109/TVCG.2017.2674978
Morgan, M., Chubb, C., & Solomon, J. A. (2008). A "dipper" function for texture discrimination based on orientation variance. *Journal of Vision*, *8*(11), 9–9. https://doi.org/10.1167/8.11.9
Mosteller, F., Siegel, A. F., Trapido, E., & Youtz, C. (1981). Eye-Fitting of Straight Lines. *The American Statistician*, *2*, 150–152.
Murray, C. J. L., Alamro, N. M. S., Hwang, H., & Lee, U. (2020). Digital public health and COVID-19. *The Lancet Public Health*, *5*(9), e469–e470. https://doi.org/10.1016/S2468-2667(20)30187-0
Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2(2)*, 175–220.
Nisbett, R. E., & Krantz, D. H. (1983). The Use of Statistical Heuristics in Everyday Inductive Reasoning. *Psychological Review*, *90*(4), 339–363.
Okazawa, G., Tajima, S., & Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proceedings of the National Academy of Sciences*, *112*(4). https://doi.org/10.1073/pnas.1415146112
Opfer, J., & Siegler, R. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, *55*(3), 169–195. https://doi.org/10.1016/j.cogpsych.2006.09.002
Orr, J. M., Sackett, P. R., & Dubois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology*, *44*(3), 473–486. https://doi.org/10.1111/j.1744-6570.1991.tb02401.x
Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, *3*(1), 29. https://doi.org/10.1186/s41235-018-0120-9
Panavas, L., Worth, A. E., Crnovrsanin, T., Sathyamurthi, T., Cordes, S., Borkin, M. A., & Dunne, C. (2022). Juvenile Graphical Perception: A Comparison between Children and Adults. *CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3491102.3501893
Park, J., & Brannon, E. M. (2013). Training the Approximate Number System Improves Math Proficiency. *Psychological Science*, *24*(10), 2013–2019. https://doi.org/10.1177/0956797613482944
Parrott, S., Guzman-Martinez, E., Ortega, L., Grabowecky, M., Huntington, M. D., & Suzuki, S. (2014). Spatial Position Influences Perception of Slope from Graphs. *Perception*, *43*(7), 647–653. https://doi.org/10.1068/p7758
Pastore, M., Lionetti, F., & Altoè, G. (2017). When One Shape Does Not Fit All: A Commentary Essay on the Use of Graphs in Psychological Research. *Frontiers in Psychology*, *8*, 1666. https://doi.org/10.3389/fpsyg.2017.01666
Peña, A., Ragan, E., & Harrison, L. (2020). Memorability of Enhanced Informational Graphics.

Interdisciplinary Journal of Signage and Wayfinding, *4*(1). https://doi.org/10.15763/issn.2470-9670.2020.v4.i1.a54

Perez, J., & Feigenson, L. (2021). Stable individual differences in infants' responses to violations of intuitive physics. *Proceedings of the National Academy of Sciences*, *118*(27), e2103805118. https://doi.org/10.1073/pnas.2103805118

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29–46. https://doi.org/10.1037/h0024722

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392–424. https://doi.org/10.1037/a0039980

Piazza, M., De Feo, V., Panzeri, S., & Dehaene, S. (2018). Learning to focus on number. *Cognition*, *181*, 35–45. https://doi.org/10.1016/j.cognition.2018.07.011

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

Piazza, M., Pica, P., Izard, V., Spelke, E. S., & Dehaene, S. (2013). Education Enhances the Acuity of the Nonverbal Approximate Number System. *Psychological Science*, *24*(6), 1037–1043. https://doi.org/10.1177/0956797612464057

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and Approximate Arithmetic in an Amazonian Indigene Group. *Science*, *306*(5695), 499–503. pbh.

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004). Distributed and Overlapping Cerebral Representations of Number, Size, and Luminance during Comparative Judgments. *Neuron*, *41*(6), 983–993. https://doi.org/10.1016/S0896-6273(04)00107-2

Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J.-B., & Dehaene, S. (2007). Fast reproducible identification and large-scale databasing of individual functional cognitive networks. *BMC Neuroscience*, *8*(1), 91. https://doi.org/10.1186/1471-2202-8-91

Poulton, E. C. (1985). Geometric illusions in reading graphs. *Perception & Psychophysics*, *37*(6), 543–548. https://doi.org/10.3758/BF03204920

Puntanen, S., & Styan, G. P. H. (1989). The Equality of the Ordinary Least Squares Estimator and the Best Linear Unbiased Estimator. *The American Statistician*, *43*(3), 153–161. https://doi.org/10.1080/00031305.1989.10475644

Pylyshyn, Z. (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*(3), 341–365. https://doi.org/10.1017/S0140525X99002022

Ramachandran, G. (1986). Exponential Model Of Fire Growth. *Fire Safety Science*, *1*, 657–666. https://doi.org/10.3801/IAFSS.FSS.1-657

Reimann, D., Ram, N., & Gaschler, R. (2022). Lollipops Help Align Visual and Statistical Fit Estimates in Scatterplots with Nonlinear Models. *IEEE Transactions on Visualization and Computer Graphics*, 1–1. https://doi.org/10.1109/TVCG.2022.3158093

Rensink, R. A. (2021). Visualization as a stimulus domain for vision science. *Journal of Vision*, *21*(8), 3. https://doi.org/10.1167/jov.21.8.3

Rensink, R. A., & Baldridge, G. (2010). The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, *29*(3), 1203–1210. https://doi.org/10.1111/j.1467-8659.2009.01694.x

Ricciardi, V., & Simon, H. (2000). What is behavioral finance? *Business, Education & Technology Journal*, *2(2)*, 1–9.

Riener, C., Proffitt, D. R., & Salthouse, T. (2005). A psychometric approach to intuitive physics. *Psychonomic Bulletin & Review*, *12*(4), 740–745. https://doi.org/10.3758/BF03196766

Romano, A., Sotis, C., Dominioni, G., & Guidi, S. (2020). The scale of COVID-19 graphs affects understanding, attitudes, and policy preferences. *Health Economics*, *29*(11), 1482–1494. https://doi.org/10.1002/hec.4143

Rugani, R., Regolin, L., & Vallortigara, G. (2007). Rudimental numerical competence in 5-day-old domestic chicks (Gallus gallus): Identification of ordinal position. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(1), 21–31. https://doi.org/10.1037/0097-7403.33.1.21

Sablé-Meyer, M., Fagot, J., Caparos, S., van Kerkoerle, T., Amalric, M., & Dehaene, S. (2021). Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. *Proceedings of the National Academy of Sciences*, *118*(16), e2023123118. https://doi.org/10.1073/pnas.2023123118

Santolin, C., Rosa-Salva, O., Regolin, L., & Vallortigara,

G. (2016). Generalization of visual regularities in newly hatched chicks (Gallus gallus). *Animal Cognition*, *19*(5), 1007–1017. https://doi.org/10.1007/s10071-016-1005-2

Schonger, M., & Sele, D. (2020). How to better communicate the exponential growth of infectious diseases. *PLOS ONE*, *15*(12), e0242839. https://doi.org/10.1371/journal.pone.0242839

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79. https://doi.org/10.1016/j.cogpsych.2017.11.002

Schulz, E., Tenenbaum, J., Reshef, D., Speekenbrink, M., & Gershman, S. (2015). Assessing the Perceived Predictability of Functions. *CogSci*, 6.

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*(4), 1241–1260. https://doi.org/10.3758/s13428-016-0783-4

Siegler, R. S., & Booth, J. L. (2004). Development of Numerical Estimation in Young Children. *Child Development*, *75*(2), 428–444. https://doi.org/10.1111/j.1467-8624.2004.00684.x

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. https://doi.org/10.1111/1467-9280.02438

Sigurd, B. (1988). Round numbers. *Language in Society*, *17*(2), 243–252.

Siirtola, H. (2019). The Cost of Pie Charts. *2019 23rd International Conference Information Visualisation (IV)*, 151–156. https://doi.org/10.1109/IV.2019.00034

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, 100306. https://doi.org/10.1016/j.cosrev.2020.100306

Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, *10*(14), 19–19. https://doi.org/10.1167/10.14.19

Spear, M. E. (1952). *Charting Statistics* (New York: McGraw-Hill).

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Spence, I. (2006). William Playfair and the Psychology of Graphs. *Proceedings of the American Statistical Association, Section on Statistical Graphics*, 2426–2436.

Starkey, G. S., & McCandliss, B. D. (2014). The emergence of "groupitizing" in children's numerical cognition. *Journal of Experimental Child Psychology*, *126*, 120–137. https://doi.org/10.1016/j.jecp.2014.03.006

Stokes, D. (2013). Cognitive Penetrability of Perception. *Philosophy Compass*, *8*(7), 646–663.

Strahan, R. F., & Hansen, C. J. (1978). Underestimating Correlation from Scatterplots. *Applied Psychological Measurement*, *2*(4), 543–550. https://doi.org/10.1177/014662167800200409

Sunday, M. A., Patel, P. A., Dodd, M. D., & Gauthier, I. (2019). Gender and hometown population density interact to predict face recognition ability. *Vision Research*, *163*, 14–23. https://doi.org/10.1016/j.visres.2019.08.006

Surber, C. (1986). Model Testing Is Not Simple: Comments on Lane, Anderson, and Kellam. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(1), 108–109.

Szafir, D. A., Haroz, S., Gleicher, M., & Franconeri, S. (2016). Four types of ensemble coding in data visualizations. *Journal of Vision*, *16*(5), 11. https://doi.org/10.1167/16.5.11

Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley and Sons.

Tufte, E. (2001). *The visual display of quantitative information* (Cheshire, CT: Graphics press., Vol. 2).

Tukey, J. W. (1977). *Exploratory Data Analysis* (Reading, MA: Addison-Wesley).

Turvey, M. T., Burton, G., Pagano, C. C., Solomon, H. Y., & Runeson, S. (1992). Role of the inertia tensor in perceiving object orientation by dynamic touch. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 714–727. https://doi.org/10.1037/0096-1523.18.3.714

Van Opstal, F., de Lange, F. P., & Dehaene, S. (2011). Rapid parallel semantic processing of numbers without awareness. *Cognition*. https://doi.org/10.1016/j.cognition.2011.03.005

Vanderplas, S., Cook, D., & Hofmann, H. (2020). Testing Statistical Charts: What Makes a Good Graph? *Annual Review of Statistics and Its Application*, *7*(1), 61–88. https://doi.org/10.1146/annurev-statistics-031219-041252

Versace, E., Spierings, M. J., Caffini, M., ten Cate, C., & Vallortigara, G. (2017). Spontaneous generalization of abstract multimodal patterns in young domestic chicks. *Animal Cognition*, *20*(3), 521–529. https://doi.org/10.1007/s10071-017-1079-5

Villagra, P. L., Preda, I., & Lucas, C. G. (2018). Data Availability and Function Extrapolation. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 7.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228–243. https://doi.org/10.1037/a0027127

Wagenaar, W. A., & Timmers, H. (1978). Extrapolation of exponential time series is not enhanced by having more data points. *Perception & Psychophysics*, *24*(2), 182–184. https://doi.org/10.3758/BF03199548

Wagenaar, W., & Sagaria, S. (1975). Misperception of exponential growth. *Perception & Psychophysics*, *18*(6), 416–422. https://doi.org/10.3758/BF03204114

Wagenaar, W., & Timmers, H. (1979). The pond-and-duckweed problem; three experiments on the misperception of exponential growth. *Acta Psychologica*, *43*, 239–251.

Wang, Y., Chen, X., Ge, T., Bao, C., Sedlmair, M., Fu, C.-W., Deussen, O., & Chen, B. (2019). Optimizing Color Assignment for Perception of Class Separability in Multiclass Scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 820–829. https://doi.org/10.1109/TVCG.2018.2864912

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology*, *69*, 105–129.

Wilson, A. J., Dehaene, S., Dubois, O., & Fayol, M. (2009). Effects of an Adaptive Game Intervention on Accessing Number Sense in Low-Socioeconomic-Status Kindergarten Children. *Mind, Brain, and Education*, *3*(4), 224–234. https://doi.org/10.1111/j.1751-228X.2009.01075.x

Xiong, C., Stokes, C., Kim, Y.-S., & Franconeri, S. (2022). Seeing What You Believe or Believing What You See? Belief Biases Correlation Estimation. *IEEE Transactions on Visualization and Computer Graphics*, Article arXiv:2208.04436. http://arxiv.org/abs/2208.04436

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015. https://doi.org/10.1073/pnas.0704450105

Yang, F., Harrison, L. T., Rensink, R. A., Franconeri, S. L., & Chang, R. (2019). Correlation Judgment and Visualization Features: A Comparative Study. *IEEE Transactions on Visualization and Computer Graphics*, *25*(3), 1474–1488. https://doi.org/10.1109/TVCG.2018.2810918

Yoghourdjian, V., Dwyer, T., Klein, K., Marriott, K., & Wybrow, M. (2018). Graph Thumbnails: Identifying and Comparing Multiple Graphs at a Glance. *IEEE Transactions on Visualization and Computer Graphics*, *24*(12), 3081–3095. https://doi.org/10.1109/TVCG.2018.2790961

RÉSUMÉ

Les graphiques sont un produit culturel, ce qui signifie qu'ils sont une invention humaine avec des règles et une syntaxe définies. À cet égard, ils sont très similaires aux mots écrits et aux chiffres, probablement les deux inventions culturelles les plus célèbres. Toutefois, contrairement à ces derniers, les graphiques ont été inventés beaucoup plus récemment et ne se sont répandus qu'au cours des deux derniers siècles. En outre, la psychologie cognitive n'a accordé que peu ou pas d'attention à la “graphicacy", c'est-à-dire à la capacité de lire et de comprendre des graphiques. Dans cette thèse, je présente des nouvelles découvertes sur la capacité humaine à extraire intuitivement des statistiques et des relations mathématiques à partir de représentations graphiques. Plus précisément, je montre que : les intuitions des graphiques sont disponibles très tôt dans le développement, indépendamment de l'éducation formelle, et sont corrélées avec les connaissances statistiques et mathématiques ; les humains sont biaisés dans leur régression mentale, estimant des pentes plus raides que prévu ; ils ne sont pas robustes à la présence de valeurs aberrantes, étant largement affectés par celles-ci dans leurs jugements statistiques intuitifs ; ils peuvent extrapoler des fonctions mathématiques non linéaires, à l'exception notable des courbes quadratiques et exponentielles. Sur la base de ces résultats, je propose également des suggestions concrètes pour améliorer la visualisation des données.

MOTS CLÉS

Graphiques ; statistiques intuitives ; psychophysique ; perception visuelle ; fonctions mathématiques.

ABSTRACT

Graphs are a cultural product, meaning that they are a human invention with defined rules and syntax. In this respect, they are very similar to written words and numbers, probably the two most famous cultural inventions. However, unlike them, graphs have been invented much more recently and they became widespread only in the last two centuries. Furthermore, graphicacy, the ability to read and understand graphs, has received little to no attention from cognitive psychology. In this thesis I present some new findings about the human ability to intuitively extract statistics and mathematical relations from graphical representations. Specifically, I show that: graphics’ intuitions are available early on in development, independently from formal education, and correlate with statistical and mathematical knowledge; humans are biased in their mental regression, estimating steeper slopes than expected; they are not robust to the presence of outliers, being largely affected by them in their intuitive statistical judgments; they can extrapolate non-linear mathematical patterns, with the notable exception of quadratic and exponential functions. Based on these findings I also propose concrete suggestions to improve data visualization.

KEYWORDS

Graphs; intuitive statistics; psychophysics; visual perception; mathematical functions.