"Piecing Data Connections Together Like a Puzzle": Effects of Increasing Task Complexity on the Effectiveness of Data Storytelling Enhanced Visualisations

Mikaela Elizabeth Milesi*
Monash University
Melbourne, Australia
mikaela.milesi@monash.edu

Paola Mejia-Domenzain*
Machine Learning for Education
EPFL
Lausanne, Switzerland
paola.mejia@epfl.ch

Laura Brandl
Psychology
LMU Munich
Munich, Germany
l.brandl@psy.lmu.de

Vanessa Echeverria
Department of Human Centred Computing
Monash University
Clayton, Australia
Escuela Superior Politécnica del Litoral
Guayaquil, Ecuador
vanessa.echeverria@monash.edu

Yueqiao Jin
Monash University
Clayton, Victoria, Australia
ariel.jin@monash.edu

Dragan Gasevic
Faculty of Information Technology
Monash University
Clayton, Victoria, Australia
dragan.gasevic@monash.edu

Yi-Shan Tsai
Department of Human Centred Computing
Monash University
Melbourne, Australia
yi-shan.tsai@monash.edu

Tanja Käser
Machine Learning for Education
EPFL
Lausanne, Switzerland
tanja.kaeser@epfl.ch

Roberto Martinez-Maldonado
Faculty of Information Technologies
Monash University
Melbourne, Victoria, Australia
roberto.martinezmaldonado@monash.edu

Abstract

The emerging concept of *data storytelling* (DS) suggests that enhancing visualisations with annotations and narratives can make complex data more insightful than conventional visualisations. Previous works found that DS-enhanced visualisations are more effective than conventional visualisations for simple tasks like identifying key data points or the main message. However, no previous work has explored the extent to which DS enhancements influence task completion across different levels of cognitive complexity. We address this gap by presenting the results of a study where 128 participants completed tasks based on four visualisations (two line charts and two choropleth maps, either with or without DS elements) spanning a range of complexity based on Bloom's taxonomy, which has been applied in data visualisation to categorise tasks hierarchically from lower to higher-order thinking. Results suggest that while DS-enhanced visualisations effectively support lower-order tasks (finding data points and understanding insights), they don't necessarily aid the correct completion of higher-order tasks (*application*, *analysis*, *evaluation* and *creation*). However, DS enhancements improve how efficiently participants complete complex tasks.

*These two authors contributed equally to this work.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714270

CCS Concepts

• **Human-centered computing** → **Visualization design and evaluation methods**.

Keywords

data storytelling, information visualisation, annotated visualisations, bloom's taxonomy

ACM Reference Format:
Mikaela Elizabeth Milesi, Paola Mejia-Domenzain, Laura Brandl, Vanessa Echeverria, Yueqiao Jin, Dragan Gasevic, Yi-Shan Tsai, Tanja Käser, and Roberto Martinez-Maldonado. 2025. "Piecing Data Connections Together Like a Puzzle": Effects of Increasing Task Complexity on the Effectiveness of Data Storytelling Enhanced Visualisations. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3714270

1 Introduction

Data has become an integral and pervasive component of modern society [21, 88], influencing sectors ranging from personal decision-making to large-scale governmental policies [43, 54, 59, 78]. Tasks involving the incorporation of data insights can vary in complexity, from laypersons using data-driven information to learn about new topics [26, 88] or about themselves [43, 54, 59, 61, 86], to higher-order analysis by data scientists generating insights and visualisations [55, 84], and governments evaluating data to create policies [78]. However, the ability to interpret and extract data insights, particularly in tasks of varying cognitive complexity, remains a challenge for many [30, 66, 73, 84]. As both the availability and

complexity of data continue to increase, it is important to support individuals – particularly those who are not data savvy – in navigating and interpreting the data [58, 65].

One of the key roles of data visualisation is to simplify the presentation of data into clear, actionable visuals, enabling people to perform tasks of varying complexity [79]. Data visualisations are often intended as a communication tool to enable *'visual discovery'* of insights [70, 80]. Some visualisations are designed and used to enable data *exploration*, where individuals are expected to independently interpret the information and extract insights for analysis [45, 85]. However, this approach relies either on the patterns and insights extracted from the data being readily apparent [85], or on individuals possessing the necessary data analysis skills to interpret the information [22]. These are therefore commonly referred to as *exploratory* visualisations, as they are primarily designed for users with data analysis expertise who are exploring unfamiliar datasets to uncover insights [12, 37, 47, 59]. In contrast, *explanatory* visualisations serve as an alternative approach whereby judicious design choices and visualisation techniques are applied in advance to enable non-data savvy individuals to make data-driven decisions by directly communicating insights through a data story [45, 70]. A technique that enables this explanatory approach is *data storytelling* (DS), whereby information is compressed by adding visual or textual enhancements to conventional visualisations [25, 70, 71].

Theoretical researchers and visualisation experts [10, 11, 18, 31, 71] have argued that DS helps users more effectively (with greater accuracy) and efficiently (in lesser time) extract insights from data compared to conventional methods [23, 31, 34, 45, 49, 71, 87]. Sensemaking refers to the process of interpreting complex data to uncover patterns, relationships, and insights that might otherwise remain hidden [15]. DS has been claimed to facilitate this process by using explanatory narratives and visual cues to guide the user [59, 75]. Yet, despite the theoretical arguments, empirical findings offer mixed results. Some studies have shown that DS can improve long-term recall [7], while others find no significant differences in memory retention compared to conventional visualisations [87]. Evidence supporting the *effectiveness* of DS in comprehension tasks comes from Shao et al. [73], who found that DS visualisations indeed improves insight comprehension, especially for single insights. Yet, findings related to *efficiency* are similarly inconsistent. While small-scale eye-tracking studies [19, 68] (with fewer than 30 participants) suggest that DS effectively captures attention, and Echeverria et al. [22] found that DS-enhanced visualisations reduced interpretation time for educators, a large-scale study [73] showed no improvement in speed. These inconsistencies suggest the need for further research to explore how DS impacts both effectiveness and efficiency in a wide range of tasks that might require different cognitive skills. The primary goal of DS is to draw focus towards salient data points and make key takeaways from data accessible [70], thus existing studies predominantly focus on lower-order tasks. However, in practice, DS is commonly applied to complex topics that require higher-order thinking, such as politics, economics, or history [64, 72]. This leaves a gap in understanding the impact that DS-enhancements could have on interpreting complex information and completing tasks that require higher-order thinking skills.

In this study, we aim to examine how DS-enhanced visualisations compare to conventional data visualisations in influencing users' understanding of tasks with varying levels of complexity. To explore this, we conducted a within-subjects controlled experiment where users are presented with tasks spanning the full range of Bloom's taxonomy [48], from low-order thinking skills (*Remember/Identify*, *Understand*) to higher-order tasks (*Apply*, *Analyse*, *Evaluate*, *Create*). Each participant was presented with four visualisations: two line charts and two choropleth maps, both with and without DS elements. An important methodological contribution of this work is the introduction of a novel *Human-in-the-Loop-AI* task generation process. This approach integrates the automation capabilities of large language models with human validation to generate question-based tasks across different levels of Bloom's taxonomy. By systematically aligning cognitive complexity with task design, this process provides a scalable framework for evaluating the cognitive impact of visualisations and can be used by future research to study visualisation efficacy across diverse contexts. A total of 128 users participated in the study where we addressed the following research questions: To what extent do DS-enhanced visualisations improve ***effectiveness*** *(**RQ1**) and* ***efficiency*** *(**RQ2**) of tasks involving data point identification, understanding, applying, analysing, and evaluating insights compared to conventional visualisations? To what extent do DS-enhanced visualisations improve the* ***quality of a creation task*** *output compared to conventional visualisations? (**RQ3**); and how are visualisations with data storytelling elements* ***perceived*** *in comparison to conventional visualisations in terms of preferences, usefulness and utility? (**RQ4**)*

Our findings contribute to the growing body of research on data storytelling by providing empirical evidence on users' performance in a wide range of cognitive tasks. Key insights include:

- DS-enhanced visualisations can influence how effectively lower-order tasks, such as finding data points and understanding insights, are completed. However, this influence does not necessarily translate to higher-order tasks where analysis, critical evaluation, or the ability to synthesise information into a novel form are necessary.
- For tasks at a higher cognitive level (such as those involving analysis, evaluation, or creation), DS-enhanced visualisations can positively impact the efficiency with which tasks are completed.
- Complex writing tasks that were written with DS-enhanced visualisations had less sentences categorised at the *evaluation* level of Bloom's taxonomy. This suggests that individuals may be less inclined to critically engage with the narrative presented by DS-enhanced visualisations.
- Participants praised DS-enhancements for providing necessary context and clarity to the data visualisations. However, some individuals still preferred the minimalism and simplicity of conventional visualisations.

2 Background

2.1 Data Visualisation & Data Storytelling

Sensemaking is a key process through which humans interpret and ascribe meaning to information [15] relying on mental models that serve as functional internal representations of external systems. These mental models preserve structural, behavioural, and data-related properties, enabling reasoning and problem-solving through

simulation in working memory [53]. Klein et al. [44] describe sensemaking as a "*deliberate effort to understand events*" with key functions including forming explanations, predicting future states, and identifying relationships. Data visualisations, such as line charts or bar charts, support sensemaking by enabling *exploratory* analysis to uncover patterns and insights [24, 45, 88]. These exploratory visualisations are often minimalistic and tailored for audiences familiar with the data context or equipped with analytical skills [59, 75].

Existing research has recognised the potential role that narratives can have on the sensemaking process [27, 67], particularly in aligning external visual representations with users' mental models [53]. To address diverse levels of visualisation literacy and facilitate cognitive offloading [35, 53], *explanatory* visualisations have evolved, incorporating storytelling elements to communicate insights effectively [27]. These author-driven "*data stories*" guide audiences through structured narratives that emphasise key findings [22, 35, 71]. Data storytelling (DS) combines data, visual elements, and narratives to contextualise and share insights [21, 70]. Principles outlined by Echeverria et al. [22] and Martinez-Maldonado et al. [59] emphasise guiding audience attention through intentional visualisation, narrative choices, and alignment with specific goals. Complementing this, Hullman and Diakopoulos [35] argues for the use of rhetorical devices such as framing, annotations, and metonymy guide interpretation, focusing the audience's attention on key insights to maintain a visual narrative flow [60].

DS can take various formats, such as data comics [5, 6, 62], narrative slideshows [17], data videos [3, 81], and annotated charts [22]. Among these, narrative-enhanced visualisations stand out for their effectiveness. Visual features of this format often align with Tufte [79]'s "*data-ink ratio*," emphasising components that directly support insight delivery [45, 87]. Design features proposed by Knaflic [45] include:

(1) ***Highlighting important data points*** by using visual attributes like colour, shading, or shapes to direct attention to critical insights [35, 70].
(2) ***Explanatory titles*** that guide the interpretation of the visualisation by directly communicating the main insight intended by the creator [9, 22].
(3) ***Annotations*** that provide context or explain key takeaways, focusing viewers on significant aspects [35, 45].

Additionally, DS-enhanced visualisations often employ ***decluttering*** to simplify the communication of potentially complex ideas by removing non-essential elements, thereby minimising cognitive load and clarifying the narrative [35, 45].

2.2 Data Storytelling Effectiveness and Efficiency

Theoretical researchers and visualisation experts, such as Segel and Heer [71], Gershon and Page [31], and Daradkeh [18] have argued that combining narratives with graphics makes visualisations more intuitive and should, in theory, improve comprehension and decision-making. An explanation for this is that DS leverages pre-attentive processing, using visual attributes like colour and shape to guide attention and facilitate quicker understanding of key data points [34, 45, 49]. However, empirical researchers have found mixed evidence regarding the effectiveness and effectiveness of DS-enhanced visualisations.

Regarding the effectiveness in low cognitive tasks like remembering, Bateman et al. [7] found that visual embellishments, such as colour and textual pointers, did not impact short-term recall compared to minimalist visualisations but did significantly enhance long-term recall. In contrast, Zdanovic et al. [87] found no significant differences in memory retention between DS-enhanced and conventional visualisations in either short-term or long-term contexts. Further evidence supporting the effectiveness of DS comes from Shao et al. [73], who conducted a study to explore how DS-enhancements affect information retrieval and insight comprehension. Their findings suggest that DS visualisations improve the effectiveness of comprehension tasks, particularly those involving single insights, compared to conventional visualisations. Moreover, Stokes et al. [77] found that charts with heavy textual annotations were preferred by users and led to better comprehension, particularly when the annotations highlighted statistical or relational components. Similarly, Kong et al. [46] found that framing the titles of visualisations influenced users' perceptions and recall, indicating that even subtle DS elements like title framing can affect how insights are derived. It is noteworthy that these studies were limited to tasks requiring low levels of cognition like finding specific data points or understanding the main take away message of the visualisation.

The efficiency of DS-enhanced visualisations — how quickly individuals can extract insights — has also been the subject of investigation to some extent. Small-scale eye-tracking studies [19, 68] suggest that DS elements attract users' attention more effectively, yet they do not provide direct evidence of whether such elements facilitate quicker comprehension. In an educational context, Echeverria et al. [22] found that DS elements can reduce the time required to process information, where teachers were able to interpret data faster with DS-enhanced visualisations compared to conventional ones. However, the only large-scale empirical study on efficiency comparing conventional visualisations against those enhanced with DS elements [73] found that while DS elements improved the effectiveness of insight comprehension, they did not necessarily make the process faster.

2.3 Data Storytelling for Supporting Tasks of Varying Complexity

The relationship between task complexity and the potential impact that it can have on the effectiveness of visualisations has already been explored in the literature. Many of these prior studies have used Bloom's taxonomy, either partially or fully, as a framework to understand this effect. Bloom's taxonomy is a hierarchical framework that emerged from educational literature [8] and it has recently been adapted and adopted in information visualisation literature [13, 51]. It organises tasks by cognitive complexity, ranging from basic levels of thinking, such as identifying and understanding, to higher-order cognitive processes including analysing, evaluating, and creating. In the revised version of Bloom's taxonomy [48], the category previously referred to as *Knowledge* was renamed *Remember* to reflect the focus on retrieving relevant knowledge. In the context of information visualisation (InfoVis), however, the first

level is often referred to as *Identify*, as it better captures the process of recognising and selecting information from visual data [73]. Accordingly, throughout this paper, we will use the term *Identify* to refer to the first level of Bloom's taxonomy.

The first level of Bloom's taxonomy was the key focus of the Zdanovic et al. [87] exploration into the effects that DS has on the ability to recall information delivered by visualisations. They found that there were no significant differences in recall between conventional and DS-enhanced visualisations.

In their investigation into the effect of DS on information retrieval and insight comprehension, Shao et al. [73] found that there were no significant differences in how efficiently participants were able to successfully answer multiple-choice questions based on DS-enhanced visualisations compared to conventional visualisations. Conversely, they found that the addition of storytelling elements can significantly influence how effectively individuals comprehend data insights, as measured by the number of correct responses. However, this study was limited to the first two levels of Bloom's taxonomy: *Remember/Identify data points* and *Understand*.

In contrast to the aforementioned studies that focus primarily on the lower levels of Bloom's taxonomy, Burns et al. [13] provided a framework adapted from the full extent of the original taxonomy to evaluate the levels to which participants understood visualisations. Although the visualisations do not explicitly incorporate the design features of DS from Knaflic [45], Burns et al. [13] highlight that the design decisions made when creating visualisations can influence a reader's ability to correctly interpret information. This motivates the exploration of the effectiveness and efficiency of visualisation with DS elements, not only at lower levels of cognitive complexity but also to understand the extent to which the design enhancements impact higher-level tasks.

Recently, the revised Bloom's taxonomy from Krathwohl [48] has been used by Adar and Lee [1] and Lee-Robbins et al. [51] as part of their research into an explanatory approach called communicative visualisations. Specifically, Bloom's taxonomy inspired the structure of learning objectives, a method of framing the intention of the visualisation in order to guide the audience, as well as evaluate the "success" of the visualisation's design [1]. However, while the effectiveness of the visualisation design is discussed using metrics such as memorability, readability, and user preference, there is minimal discussion on the relationship between the complexity of the task (as measured by Bloom's taxonomy) and the design of the visualisations.

3 Method

To address the research questions, this paper presents a study comparing pairs of visualisations: conventional versus DS-enhanced, across a range of topics. Tasks were developed based on Bloom's taxonomy to assess different levels of cognitive complexity. A within-subjects design was used, where participants completed both multiple-choice and open-ended question-based tasks for each visualisation pair. The rest of this section outlines the key elements of the study, including: i) the dataset and materials used; ii) the creation and transformation process for conventional and DS-enhanced visualisations; iii) the development of question-based tasks aligned with Bloom's taxonomy to evaluate the visualisations; iv) the experimental design and procedure; v) participant recruitment and demographics; and vi) the analysis performed for each research question.

3.1 Dataset and Materials

In order to compare conventional visualisations with those containing DS-enhancements (RQs 1-4), we first needed to identify an appropriate dataset that could be used to extract a set of data stories to communicate.

For this study, we aimed to find a dataset that would be unfamiliar enough to the general public that they would require the visualisations to complete tasks of different complexity, but not so inaccessible that they would be unable to answer questions at all. To this end, we selected open-source data from the *Our World in Data* website. The data stories presented in this study were based on the interconnected topics of *Colonisation*, *Territory Control*, *State Capacity*, and *Tax as a percentage of GDP*, as explained by Herre et al. [33][1]. These topics were chosen over alternative datasets because the *Our World in Data* page already contained well-documented insights, reducing the risk of introducing bias or fabricating narratives. Moreover, the topics offered a diverse set of visual representations, including both maps and line charts, alongside detailed endnotes, which provided a rich context for DS-enhancements.

The visualisations in this study (to be fully described below, in Section 3.2) were created using Python's *Matplotlib* and *Geopandas* library. Additional visual embellishments, such as annotations, were added using the graphic design platform *Canva* [2].

3.2 Visualisation Design and Transformation Process

The design process to create the pairs of visualisations for this study followed the *Data Storytelling Framework* developed by Zdanovic et al. [87] and the visual enhancements described in the data storytelling and visualisation literature [22, 45, 70, 77, 79]. Zdanovic et al. [87]'s framework is divided into three distinct phases – 1) *Explore the data*, 2) *Craft the visualisation*, and 3) *Tell the story* – each with individual steps that are necessary to guide the visualisation creation process. While generating conventional visualisations generally involves going through the first two phases – data exploration and visualisation creation – generating data stories extends this process by performing a more nuanced crafting of the visualisation in the second phase and incorporating storytelling elements in the third phase, enriching the narrative and enhancing the communication of insights. The full details about the visualisation design, transformation process, and implementation, are available at this link. The resulting pairs of visualisations are depicted in Figures 1-4.

The conventional visualisations were crafted using the original aesthetics and visualisation types from the respective *Our World in Data* data stories. To ensure a fair comparison between tasks solved with the conventional and DS-enhanced visualisations, we applied the same data filtering process (as outlined in Zdanovic et al. [87]) to include identical data points in both visualisations. Consequently, some conventional visualisations differ slightly from the original visualisation from *Our World in Data*. For instance, in conventional

[1]Dataset: https://ourworldindata.org/state-capacity
[2]Canva: https://www.canva.com/

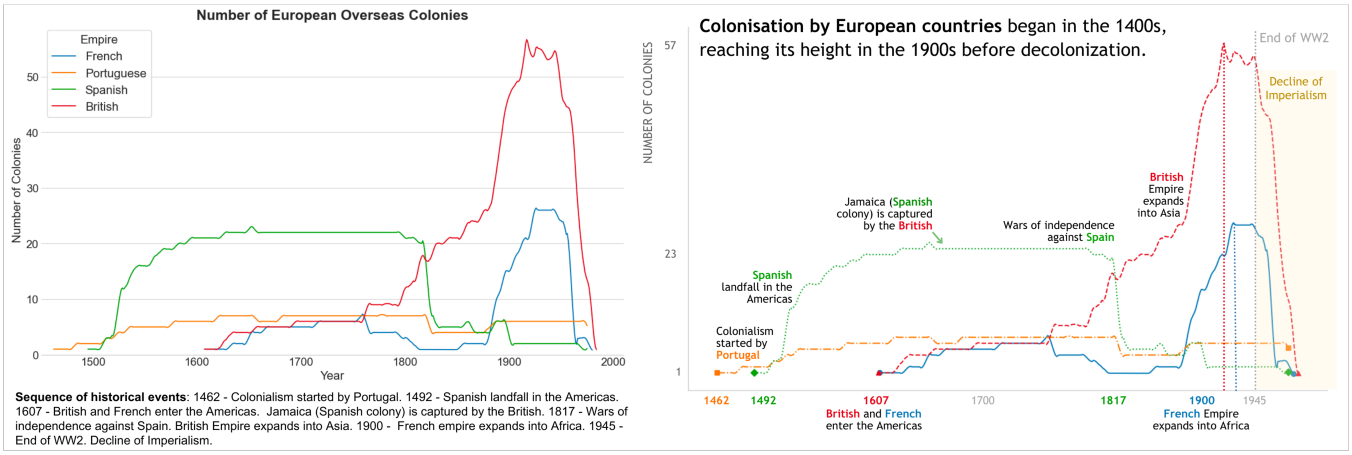

Figure 1: Visualisation A based on the topic of Colonialism. The pair consists of the conventional visualisation A-CV [left] and the DS-enhanced visualisation A-DS [right].

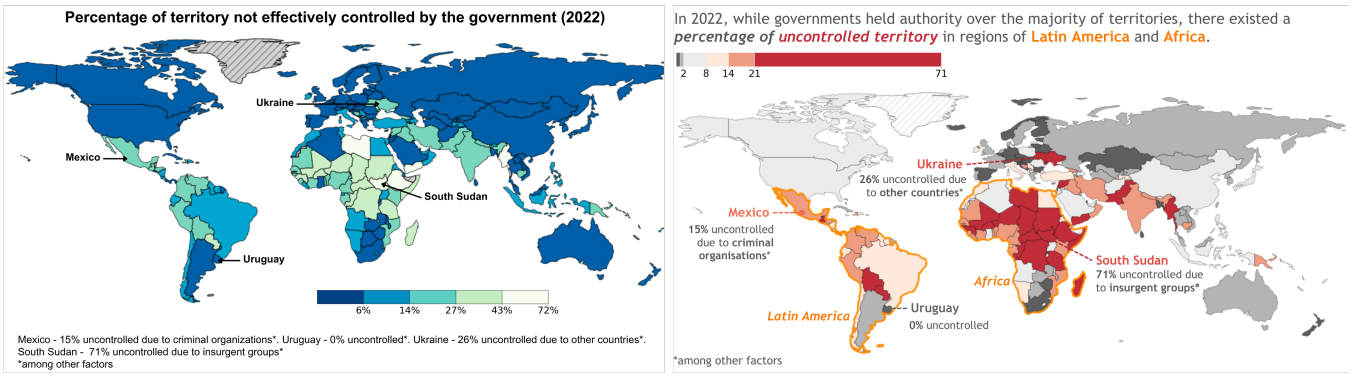

Figure 2: Visualisation B based on the topic of Territory Control. The pair consists of the conventional visualisation B-CV [left] and the DS-enhanced visualisation B-DS [right].

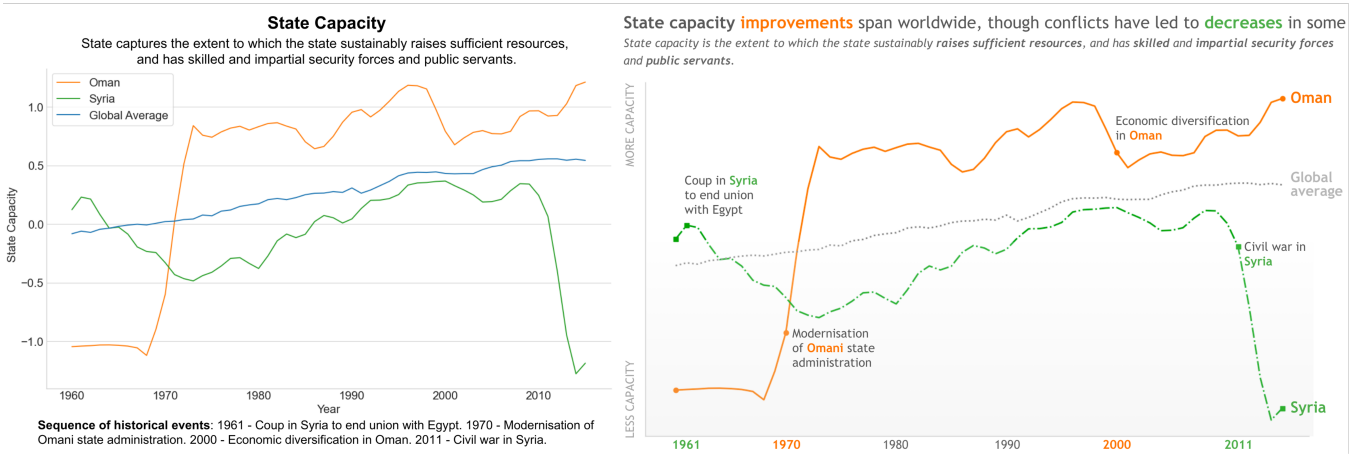

Figure 3: Visualisation C based on the topic of State Capacity. The pair consists of the conventional visualisation C-CV [left] and the DS-enhanced visualisation C-DS [right].

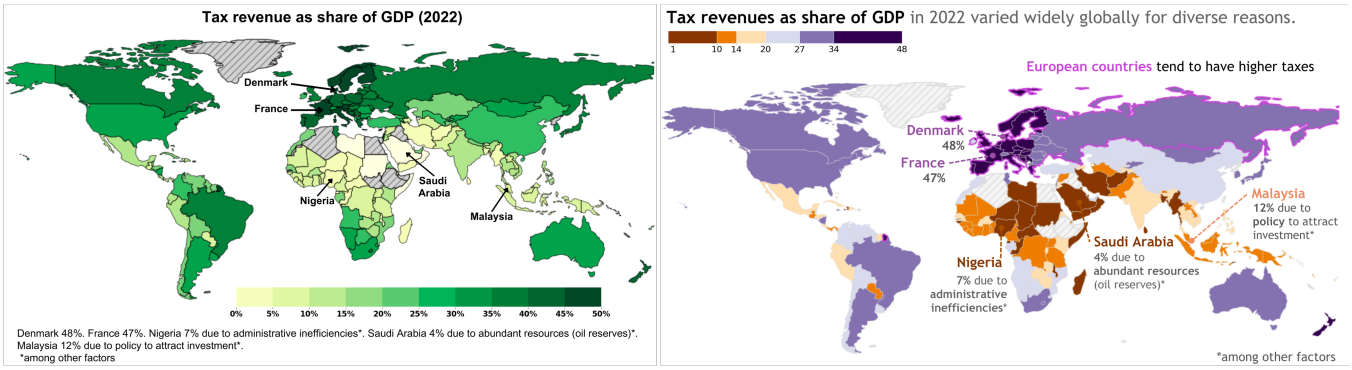

Figure 4: Visualisation D based on the topic of Tax Revenue as a share of GDP. The pair consists of the conventional visualisation D-CV [left] and the DS-enhanced visualisation D-DS [right].

visualisation A (Figure 1), certain European empires included in the original visualisation were excluded. Both the original *Our World in Data* visualisations and the conventional visualisations from this study can be viewed via this link.

In the DS-enhanced visualisations, text from Herre et al. [33] was adapted to the form of annotations and explanatory titles. Using the suggestions by Kim et al. [42] and Stokes et al. [77], annotations were carefully placed to guide the audience towards the data points they were describing. Pre-attentive attributes for text, such as bolding particular terms or using colour to highlight to further draw attention towards key terms or data-points [45]. Additionally, a footnote was added to the bottom of the conventional visualisations to provide participants with the same contextual information, ensuring that differences in responses were due to how the information was presented (i.e., DS-enhanced vs. conventional) rather than the amount of information provided.

3.3 Human-in-the-Loop-AI Question-Based Task Generation Process

To generate question-based tasks aligned with the various levels of Bloom's taxonomy and ensure comprehensive coverage of the selected dataset, we adopted a *Human-in-the-Loop-AI* collaborative process [32]. This process involves instructing AI systems to generate question-based tasks on a large scale, followed by human review and refinement of the content [4, 32, 89]. Our question-based task generation process included four steps:

Step 1: Generation with LLMs. Motivated by the latest research in automated item generation [4, 32, 89], GPT-4 was used to create an initial pool of questions. The rationale behind using GPT-4 was to efficiently generate a large set of potential questions that aligned with both the content of the visualisations and the specific levels of Bloom's taxonomy. This aligns with emerging approaches, such as those used by Duolingo for generating assessment items, which leverage LLMs to efficiently create questions at varying levels of complexity while minimising human bias and keeping a consistent style and tone [4]. For the first five levels of Bloom's taxonomy, the conventional visualisations were uploaded alongside specific prompts based on definitions from Krathwohl [48]. The conventional visualisations were chosen to be uploaded as they contained the least amount of explanatory information to ensure fair comparisons. For example, to generate questions for the *Understand* level, GPT-4 was instructed as follows: "*You are an expert learning scientist. Given the provided graph, generate 10 close-ended questions (e.g., multiple-choice, ranking questions, checkboxes) for the second level of Bloom's taxonomy: Understand. Determining the meaning of instructional messages, including oral, written, and graphic communication (e.g., Explaining)*".

Step 2: Validation with BloomBERT. A total of 89 questions were generated by GPT-4 which were then validated using BloomBERT [50], a fine-tuned DistilBERT model on over 6K questions for 40 epochs achieving a classification accuracy of 91% for Bloom's taxonomy levels. The 46 questions that BloomBERT accurately categorised into the intended Bloom's level were retained for further refinement.

Step 3: Adaptation by Authors. Following validation by BloomBERT, the resulting 46 questions were reviewed by two of the authors, who jointly refined them for clarity, relevance, and alignment with the visualisations. Low quality questions were discarded leaving 36 viable questions.

Step 4: Validation by Human Annotators. Finally, the questions were validated by another two of the authors who were not involved in the previous steps. They categorised each question according to Bloom's taxonomy levels as described by Burns et al. [13] and Krathwohl [48]. The inter-rater agreement between the human annotators was evaluated using Cohen's Kappa, resulting in an initial agreement score of 0.89, indicating almost perfect agreement. Any questions where the annotators disagreed were discarded. An example of questions for visualisation A are available in appendix A.

3.4 Study Design

We conducted a controlled study to investigate the impact of DS-enhancements in visualisations using a within-subjects design. Participants were exposed to both conditions: conventional visualisations and DS-enhanced visualisations. Figure 5 details the structure of the survey. The survey structure, managed via Qualtrics, handled the randomisation process to minimise order effects. The full questionnaire is available here.

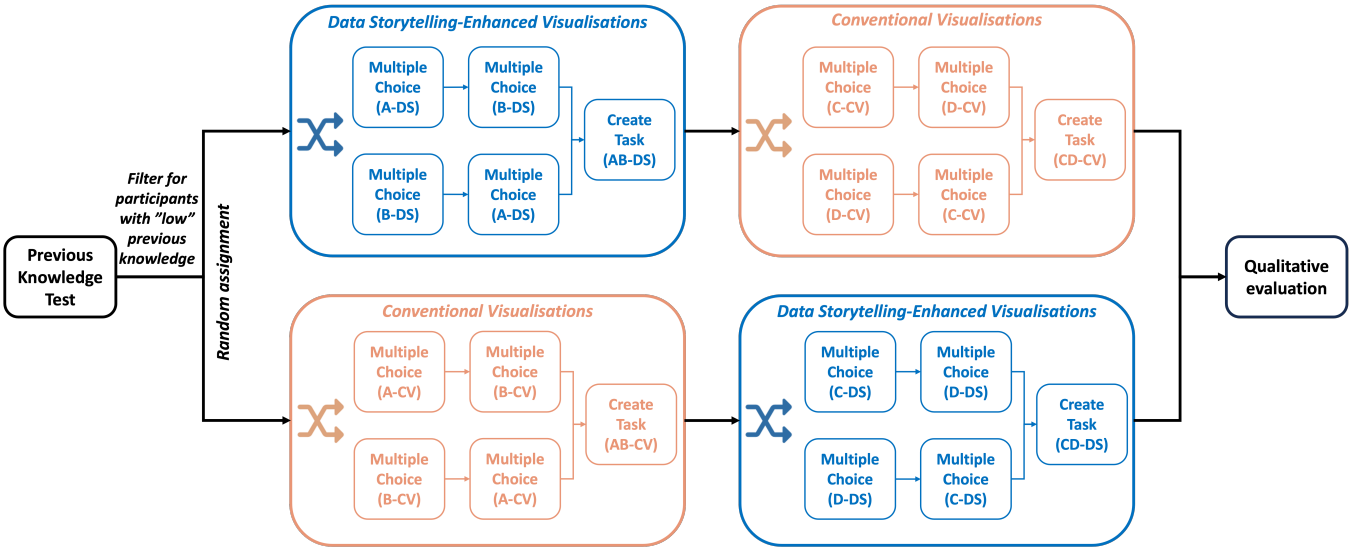

Figure 5: Outline of the study design comparing conventional and DS-enhanced visualisations. Participants, after a pre-test, were randomly assigned to view both types of visualisations and answer MCQ for each visualisation and an open task linking each pair. Lastly, participants provided a qualitative evaluation of their preferences.

Prior Knowledge Questions. To ensure responses were based on information from visualisations rather than prior knowledge, inspired by Zdanovic et al. [87], participants completed a self-assessment (five-point Likert scale) and a multiple-choice question on each topic. Participants who self-reported as *Very Knowledgeable* or *Extremely Knowledgeable* and answered correctly were excluded and compensated for their time with 1 USD.

RQ1: Measuring Effectiveness & RQ2: Efficiency. Participants answered eight multiple-choice questions per visualisation, covering the first five levels of Bloom's taxonomy, with an "*I am not sure*" option to discourage guessing [66, 73]. Visualisations `A` and `B` and `C` and `D` were presented alternately with and without DS enhancements (`AB-DS/CD-CV` or `AB-CV/CD-DS`), ensuring exposure to one conventional and one DS-enhanced line chart and map per participant.

RQ3: Measuring Quality of Response to Create Tasks. Next, participants completed an open-ended "Create" task requiring them to integrate insights from visualisations. For example, participants wrote compositions combining information from `A` and `B` (e.g., `A-CV` and `B-CV`) and `C` and `D` (e.g., `C-DS` and `D-DS`).

RQ4: Measuring Perception of Data Storytelling. Lastly, participants compared DS-enhanced and conventional visualisations of the same type (e.g., line charts). They indicated preferences and justified their choices. Additionally, participants clicked on visual elements they found useful to simulate eye-tracking insights.

3.5 Participants

A priori power analysis using G*Power 3.1 determined that a sample size of 128 participants was needed to achieve sufficient statistical power ($\alpha = 0.05$, power $= 0.95$) using an effect size of 0.3 based on Shao et al. [73]. Three pilot studies preceded the main study: the first refined the survey design with feedback from four HCI experts; the second, involving five participants, estimated study duration; and the third, with 25 participants, tested randomisation and analysis. Participants who answered questions that changed during pilots (e.g., changing the heatmap questions to allow multiple clicks) had their results excluded from the final dataset.

Participants for the main study were recruited through CloudConnect. The study lasted a median of 51 minutes (IQR = 30), and participants were compensated 9 USD. Pre-screening ensured fluency in English and a balanced sex ratio. Of 157 initial participants, 128 valid responses remained after excluding eight overly knowledgeable individuals, ten who failed attention checks, and 21 suspected of using generative AI for written tasks. Written responses with a calculated word-per-minute rate of over 50 [20] were manually checked for attributes that are characteristic of GPT-4, such as being overly verbose or formal in language [40].

The final sample included 63 males (49%) and 65 females (51%), aged 18–64 years ($\mu = 24.14, \sigma = 7.02$). Educational backgrounds varied: 41% had some college education, 37% held bachelor's or associate degrees, and 14% had high school diplomas. Participants were mostly from the US (109), with smaller numbers from the UK, Canada, Australia, New Zealand, and Ireland. Most participants (53–54%) expressed confidence in understanding and interpreting data, with fewer than 10% lacking confidence. Ethics approval was obtained from Monash University. Informed consent was secured from each participant.

3.6 Metrics

Effectiveness (RQ1) was measured as the percentage of correct responses, calculated separately for each visualisation type and cognitive category (e.g., *Identify*). For a set S, effectiveness was determined by averaging the binary correctness of responses, where each correct answer scored 1 and incorrect answers scored 0.

Efficiency (RQ2) assessed the time participants took to answer questions correctly, measured from when a question was displayed to when the correct answer was submitted. Efficiency was calculated as the average time spent on correctly answered questions for each participant, separated by visualisation type and cognitive levels in Bloom's taxonomy.

To compare text compositions across visualisation conditions, the *Percentage Change* was used, reflecting the proportional difference in the number of sentences normalised by the larger value. This symmetric metric bounded changes between −1 and 1. For example, the percentage change between DS-enhanced and conventional conditions was calculated as:

$$\text{Percentage Change}_{\text{Overall}} = \frac{DS - enhanced - conventional}{\max(DS - enhanced, conventional)} \times 100$$

3.7 Data Analysis

We used the Shapiro-Wilk W-test to evaluate the normality of the *effectiveness* and *efficiency* scores. The results indicated that these variables deviated significantly from a normal distribution, thus we used non-parametric tests for the analysis. Wilcoxon signed-rank test was used to compute differences between paired observations; and consequently used the matched pairs rank biserial correlation to calculate the effect size [41]. We report the absolute value of the effect size.

For the *effectiveness* (RQ1) and *efficiency* (RQ2), we calculated the score for each set of visualisations (DS-enhanced and conventional) for each participant, both overall and at each level of Bloom's taxonomy. We then applied the Wilcoxon signed-rank test, suitable for paired observations, to evaluate within-subject differences. This analysis was conducted both overall and across each level of Bloom's taxonomy, with corrections for multiple comparisons applied using the Benjamini-Hochberg (BH) procedure.

The *Create* task (RQ3) involved sentence-level rhetorical structure analysis [57]. Several frameworks could be employed for this analysis including Rhetorical Structure Theory [14], semantic content levels for accessible visualisations [56] and Bloom's taxonomy [1, 38, 39]. Among these, we selected Bloom's taxonomy for its hierarchical and systematic approach to evaluating cognitive processes involved in text comprehension and creation [38, 39] and its previous use as a taxonomy for visualisation learning objectives [1, 51, 73]. The dataset consisted of 1818 sentences coming from 256 texts (two texts per participant). Four researchers developed a rubric for annotation (see Appendix B), with two independently labelling 32 sentences to validate the rubric, achieving an inter-rater reliability of 0.78 (Cohen's Kappa). Discrepancies were resolved through discussion, and remaining sentences were annotated individually, including Bloom's taxonomy levels and uncertainty. The annotators then met to discuss the sentences where they were not certain to reach a consensus.

The total number of sentences and their distribution across Bloom's levels were calculated separately for DS-enhanced and conventional conditions. Tasks from visualisation pairs (`AB-DS/CD-DS` and `AB-CV/CD-CV`) were aggregated to account for differences in visualisation sets, and potential effects of the presentation sequence, as explored in Hullman et al. [36]. A Wilcoxon signed-rank test, with Benjamini-Hochberg adjustments, was used to assess within-subject differences between conditions.

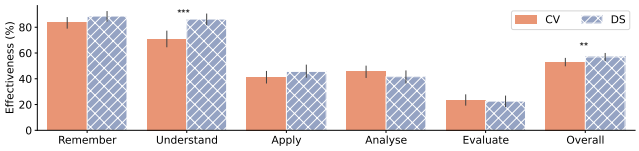

Figure 6: *Effectiveness* of participants in answering questions across different cognitive levels of Bloom's taxonomy for DS-enhanced and conventional visualisations. The effectiveness is measured as the percentage of correct answers. Error bars represent a 95% confidence interval, $*$ indicates significance at $p < 0.05$, $**$ at $p < 0.01$, and $***$ at $p < 0.001$.

User preferences (RQ4) were evaluated based on the frequency of DS-enhanced visualisation selection (overall preference rate). Clickstream interactions generated heatmaps highlighting useful elements (e.g., titles, annotations). In addition, we calculated the number of clicks on specific elements of focus including the title, legend, axis labels, specific data points within the graph, annotations on data points on the graph, and contextual information on data points or as a footnote. Thematic analysis of open-ended responses identified key themes, reviewed and verified by two researchers. Finally, the frequency of these themes was calculated for each group (DS-enhanced vs conventional).

4 Results

In this section, we present results for each of our research questions.

4.1 RQ1: Effectiveness

In the first analysis, we examined participants' *effectiveness* when answering questions about the DS-enhanced and conventional visualisations for the first five levels of Bloom's taxonomy. As seen on the *Overall* bar in Figure 6, we observed that generally participants performed better in the DS-enhanced set of questions compared to the conventional set. The overall *effectiveness* for the DS-enhanced and conventional visualisation sets were 57% and 53%, respectively. A Wilcoxon signed-rank test [3] revealed a statistically significant difference with a moderate effect size ($W = 2217$, $p = 0.004$, $r = 0.31$)

Transitioning to a more detailed analysis of the effectiveness in tasks of different complexity, Table 1 and Figure 6 show the fine-grained effectiveness scores per Bloom's taxonomy level. In the first three levels: *Identify*, *Understand*, and *Apply* – the DS-enhanced condition outperformed the conventional visualisations. For *Identify* ($W = 667$, $p = 0.96$, $r = 0.22$) and *Apply* ($W = 1826$, $p = 0.16$, $r = 0.16$), the DS-enhanced effectiveness was 5% higher than the conventional condition, though these differences were not statistically significant and the effect size was small. The most substantial difference occurred at the *Understand* level ($W = 336$, $p < 0.001$, $r = 0.59$), where the DS-enhanced effectiveness was 15% higher than conventional visualisations, with a mean score of 86% compared to 71% for the conventional condition – a statistically

[3]The non-parametric test was performed after significant Shapiro tests, $W = 0.978$, $p = 0.038$ and $W = 0.966$, $p = 0.003$

significant difference and a moderate effect. Different from the lower cognitive levels where DS-enhanced effectiveness was higher than the conventional scores, at the higher cognitive levels –*Analyse* ($W = 1398$, $p = 0.17$, $r = 0.13$) and *Evaluate* ($W = 954$, $p = 0.70$, $r = 0.05$) – conventional visualisations performed slightly better by 4% and 1%, respectively, though these differences were not statistically significant.

In summary, participants generally performed better with DS-enhanced visualisations, achieving a significantly higher overall effectiveness (57%) compared to conventional visualisations (53%). The most notable improvement for DS-enhanced visualisations was at the *Understand* level, where effectiveness was 15% higher than conventional visualisations

4.2 RQ2: Efficiency

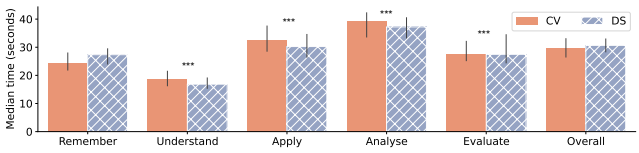

Figure 7: *Efficiency* of participants in answering questions across different cognitive levels of Bloom's Taxonomy for the DS-enhanced and conventional visualisations. The efficiency is measured as the average correct answer time. Error bars represent a 95% confidence interval, $*$ indicates significance at $p < 0.05$, $**$ at $p < 0.01$, and $***$ at $p < 0.001$.

Regarding RQ2, we examined participants' *efficiency* in answering the questions correctly. The efficiency data presented in Table 1 and Figure 7 indicates how quickly participants arrived at the correct answers in both sets of questions. The table shows the median seconds of participants and the IQR. In general, there was no significant difference in the overall *efficiency* between the two visualisation methods ($W = 3644$, $p = 0.2$, $r = 0.11$). However, interesting trends emerge at specific levels of Bloom's taxonomy. At the lowest level, *Identify*, participants were able to identify the correct answers faster when using the conventional visualisations, with a median time of 24.2 seconds, which was more than 3 seconds faster than the median time for the DS-enhanced condition (27.4 seconds). Despite this, the within-subjects difference was not statistically significant ($W = 3944$, $p = 0.7$, $r = 0.03$).

In contrast, for the higher levels of Bloom's taxonomy –*Understand* ($W = 2471$, $p < 0.001$, $r = 0.18$), *Apply* ($W = 2482$, $p < 0.001$, $r = 0.02$), *Analyse* ($W = 2381$, $p =< 0.001$, $r = 0.04$), and *Evaluate* ($W = 691$, $p < 0.001$, $r = 0.16$) – participants were more efficient when using the DS-enhanced visualisations. The median time per participant for the DS-enhanced visualisations was consistently lower than the conventional visualisations across these levels, and the differences within subjects were statistically significant. This suggests that participants were significantly more efficient at arriving at the correct answers when engaging with visualisations that included data storytelling elements at these higher cognitive levels.

To summarise, participants showed no significant difference in overall efficiency between DS-enhanced and conventional visualisations. However, participants were significantly more efficient with DS-enhanced visualisations at higher cognitive levels (*Understand*, *Apply*, *Analyse*, and *Evaluate*) with a low effect size.

4.3 RQ3: Quality of Complex Writing Creation Tasks

To investigate cognitive differences in the *Create* task, we analysed participants' sentence-level responses. The focus was to determine whether visualisations with data storytelling elements (DS-enhanced) influenced participants' text composition compared to conventional visualisations (conventional). Overall, the average sentences for both conditions was 7.1 and there were no within-subject differences ($W = 1774$, $p = 0.958$, $r < 0.01$). The most frequent cognitive level across conditions was *Identify*, with participants writing an average of 2.6 sentences, followed by *Analysis* (1.7 sentences) and *Understand* (0.9 sentences). The least frequent levels were *Evaluate* (0.7 sentences), *Create* (0.4 sentences), *Other* (0.36 sentences), and finally *Apply*, which was the least frequent category with an average of 0.32 sentences.

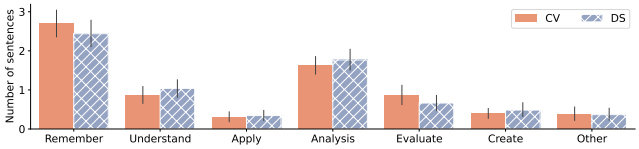

Figure 8: Distribution of sentences across different cognitive levels of Bloom's taxonomy for DS-enhanced and conventional visualisations. Error bars represent a 95% confidence interval.

Figure 8 visualises the distribution of sentences across cognitive categories for each condition. Analysis of the data, as shown in Table 2, revealed no significant within-subject differences in text composition between the DS-enhanced and conventional visualisations. However, some trends were notable: participants who viewed conventional visualisations tended to produce more sentences at the *Identify* ($W = 2152$, $p = 0.19$, $r = 0.15$) and *Evaluate* level ($W = 867$, $p = 0.12$, $r = 0.22$).

As seen in Table 2, *Evaluate* had the largest percentage change between the conditions involving DS-enhanced and conventional visualisations. When the participants composed a text using the conventional visualisations, they wrote on average 29% more *Evaluate* sentences. In contrast, those who used DS-enhanced visualisations wrote 21% more sentences in the *Understand* level ($W = 1188$, $p = 0.1$, $r = 0.21$), 9% more at the *Analysis* level ($W = 1818$, $p = 0.55$, $r = 0.07$), and 24% more at the *Create* level ($W = 622$, $p = 0.39$, $r = 0.13$).

To sum up, although not statistically significant, the most notable differences between the tasks with conventional and DS-enhanced visualisations were that participants using DS-enhanced wrote, on average, 21% more *Understand* sentences and 24% more *Create* sentences, but 29% fewer *Evaluate* sentences

4.4 RQ4: User Perception

To explore how visualisations with data storytelling elements are perceived compared to conventional visualisations, we analysed

Table 1: Comparison of *Effectiveness* and *Efficiency* between conventional and DS-enhanced visualisations at different cognitive levels. *Effectiveness* is measured as percentage of correct answers and *Efficiency* as the median correct answering time in seconds

Level	*Effectiveness*			*Efficiency*		
	conventional	DS-enhanced	*p*-value	conventional	DS-enhanced	*p*-value
Identify	84%	**89%**	0.096	**24.2** (*IQR*=21.9)	27.4 (*IQR*= 19.9)	0.661
Understand	71%	**86%**	<0.001	18.6 (*IQR*=15.1)	**16.9** (*IQR*= 12.8)	<0.001
Apply	41%	**46%**	0.156	32.7 (*IQR*=31.0)	**30.2** (*IQR*=22.6)	<0.001
Analyse	**46%**	42%	0.145	39.3 (*IQR*=25.5)	**37.2** (*IQR*=27.6)	<0.001
Evaluate	**24%**	23%	0.699	27.6 (*IQR*=26.5)	**27.4** (*IQR*=19.5)	<0.001
Overall	53%	**57%**	0.004	**29.7** (*IQR*=19.7)	30.7 (*IQR*=20.5)	0.249

Table 2: Comparison of the average number of sentences participants wrote across cognitive levels when using DS-enhanced and conventional visualisations.

Level	DS-enhanced	conventional	DS - CV	Percentage Change (%)	*p*-value
Identify	2.44	**2.70**	-0.26	-11%	0.19
Understand	**1.04**	0.86	**0.18**	21%	0.10
Apply	**0.34**	0.30	**0.04**	16%	0.57
Analysis	**1.77**	1.63	**0.14**	9%	0.55
Evaluate	0.66	**0.86**	-0.20	-29%	0.12
Create	**0.48**	0.39	**0.09**	24%	0.39
Other	0.37	0.36	0.01	2%	0.99

participants' preferences and justifications. The results showed that the DS-enhanced visualisations were favoured over the conventional visualisations 60% of the time across both graph types.

When participants chose the DS-enhanced visualisation over the conventional visualisation (60% of the time), we found that in the justification of their choice, the most commonly mentioned element was the placement of contextual information directly next to relevant data points (e.g., next to country names) cited 84% of participants. In particular, one participant stated: *"I think the blurbs of text that explain the trends in the lines give some much-needed context in a visually pleasing format"*. Another noted that the DS-enhancements permitted them to "*piece data connections together like a puzzle*". Additionally, 17% specifically mentioned annotations on data points on the graph, specifically the display of exact values (such as percentages) on the visualisation. One participant noted the direct percentages: *"The percentages were up with the country name and made it easier to process, requiring less time to find percentages"*. Furthermore, 16% of participants mentioned that the high contrast between colours, particularly among different categories, made distinctions clearer; while 11% mentioned the different dashed line patterns in the line charts, for example, one participant stated: *"I think the difference in the pattern of the lines in the second graph also made the countries more differentiable"*. Lastly, 5% praised the focus on important areas. As one participant stated: *I appreciate that it focused on specific places and didn't color the other ones that were not of focus — it helped keep the perspective and focus on the areas of interest instead of being overwhelming"*.

In contrast, when participants preferred the conventional visualisation (40% of the time), 56% of the justification cited a preference for minimal clutter, favouring charts that were not overloaded with information. Furthermore, 21% referred to the map's legend preferring a legend where the intervals between different categories are consistent and evenly distributed and lastly, 13% expressed a preference for a monotonic colour scale.

Complementing participants' written justifications, Figure 11 and Figure 12 display heatmaps of participants' clicks for visualisation B. In both versions, conventional (Figure 12) and DS-enhanced (Figure 11), the most popular areas were the annotations on data points on the graph (like the country names) and the legends. Moreover, the heatmaps shown in Figure 9 (DS-enhanced) and Figure 10 (conventional) show the areas perceived as useful from visualisation A.

In Figure 9 (DS-enhanced) the contextual information on data points were clicked on 170 times whereas in Figure 10 (conventional) the contextual information as a footnote received three times fewer clicks (only 46 clicks). Instead, in Figure 10 (conventional), the most highlighted element is the legend which was clicked on 75 times. Interestingly, both charts have the same y-axis label yet it is one of the main chosen elements in Figure 10 (conventional) clicked on it by 29 participants and it was less frequently chosen in Figure 9 (DS-enhanced) where it was only selected by 8 participants.

In summary, participants generally favoured DS-enhanced visualisations over conventional ones 60% of the time, citing the placement of contextual information next to relevant data points (84%), clear colour contrast (16%), and distinct line patterns (11%) as key reasons. In contrast, those who preferred conventional visualisations (40%) valued minimal clutter (56%), evenly spaced legends (21%), and monotonic colour scales (13%).

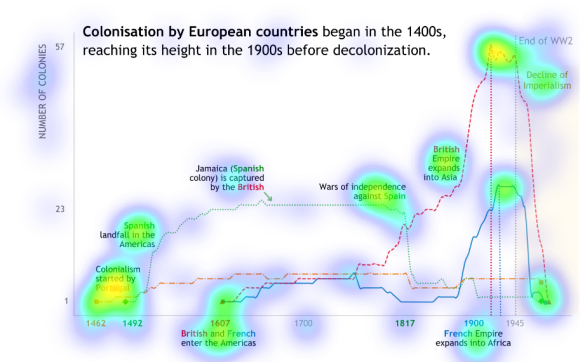

Figure 9: Heatmap showing the aspects that participants found useful in A–DS.

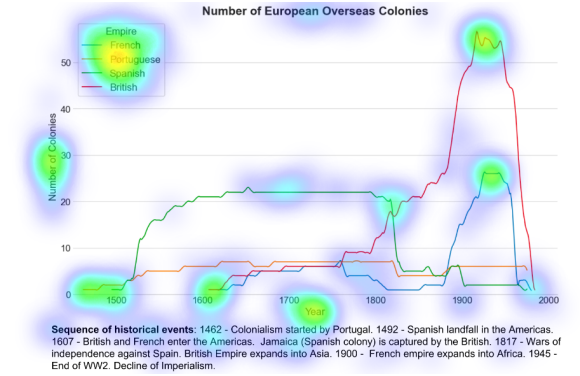

Figure 10: Heatmap showing the aspects that participants found useful in A–CV.

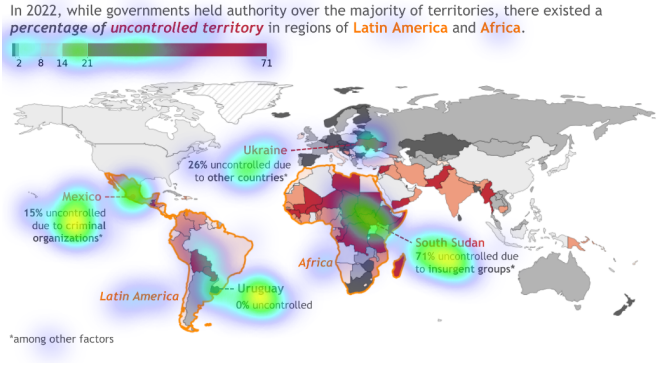

Figure 11: Heatmap showing the elements that participants found useful in B–DS.

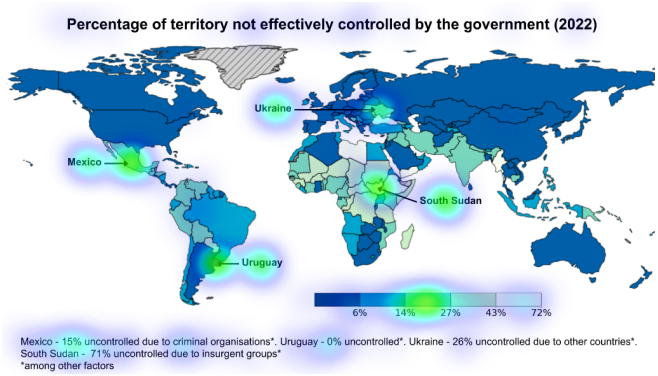

Figure 12: Heatmap showing the aspects that participants found useful in B–CV.

5 Discussion

In this section, we present a summary of the key findings per research questions, provide the implications for both research and practice, and address the limitations of our study. Additionally, we outline potential directions for future research.

5.1 Summary of Results and Research Questions

5.1.1 RQ1: Effectiveness. Participants demonstrated higher *effectiveness* scores on the DS-enhanced visualisations compared to the conventional visualisations, with a statistically significant difference and a small effect size. Interestingly, the *effectiveness* scores were higher for the first three levels of Bloom's taxonomy—*Identify*, *Understand*, and *Apply*—when using DS-enhanced visualisations. These results are consistent with those of Shao et al. [73], who also observed that DS significantly improved participants' ability to identify and comprehend key data insights in the first two levels of cognition. However, unlike previous work [73, 87], we studied the effectiveness at further levels of Bloom's taxonomy. For tasks requiring higher-order cognitive skills, such as *Analyse* and *Evaluate*, the *effectiveness* scores were higher for the conventional visualisations. Although the differences were not significant, this may suggest that DS elements are most effective in simpler tasks, where narrative elements help structure information, but less so when participants must engage in more abstract thinking, where the narrative may not provide additional value or could even distract from critical thinking processes. The more structured nature of DS-enhanced visualisations, which often highlight specific data points, might direct users' attention and thinking, whereas conventional visualisations offer more flexibility for individual interpretation.

5.1.2 RQ2: Efficiency. Regarding *efficiency*, participants showed no significant overall difference between DS-enhanced and conventional visualisations. However, at higher cognitive levels (*Understand*, *Apply*, *Analyse*, *Evaluate*), participants were significantly more efficient with DS-enhanced visualisations while for conventional visualisations they were slightly quicker at the *Identify* level. Our findings from the *Identify* level are in line with Shao et al. [73] where at tasks of lower cognitive levels, like *Identify*, the DS elements on visualisations made participants take longer to answer questions correctly. One possible explanation is that the additional narrative and contextual elements introduced in DS-enhanced visualisations can distract or delay participants when the task requires simple recall or recognition. In contrast, conventional visualisations, being more straightforward and minimalistic, allow for quicker retrieval of known information.

Another possible explanation may be related to the hierarchical ordering of the tasks. All the participants were presented tasks following the order Bloom's taxonomy, meaning that the first question was an *Identify* question followed by *Understand* and higher cognitive levels. When participants were exposed to the DS-enhanced visualisations, they may have taken longer in the *Identify* tasks than in the conventional visualisation condition as they were reviewing all of the information available in the visualisation. Once they had understood the story they were faster at answering correctly the questions from all the higher levels (*Understand*, *Apply*, *Analyse*, *Evaluate*). This is consistent with the discussion from Liu et al. [52] and Galati et al. [28] who suggest that individuals will take time to engage in sensemaking behaviours with visualisations in order to understand the contents before they complete any tasks using the visualisation's information. The findings from this study also resonates with the argument put forth by Ryan [70] who believes that individuals will typically notice visual attributes, such as colour, before focusing on the deeper meaning of the visualisation. It is possible that participants took the time to perceive the artistic visual elements of the DS-enhancements before reviewing the intended message of the data story.

5.1.3 RQ3: Quality of Complex Creation Tasks. Participants' text composition showed no significant overall difference between DS-enhanced and conventional visualisations. However, conventional visualisations prompted more sentences at the *Identify* and *Evaluate* levels, while DS-enhanced visualisations led to more sentences in the *Understand*, *Analyse*, and *Create* categories. The pattern of engaging in different levels of cognition depending on the added elements on the visualisation resembles the findings from Stokes et al. [77], where participants engaged in deeper semantic content takeaways when the text was placed near relevant data points on the chart, rather than on the title. In our study, contextual information on data points in the DS-enhanced visualisations may have contributed to participants engaging more with higher-level cognitive tasks, providing further evidence of the relationship between text placement and cognitive depth in visualisation-based tasks.

One possible explanation for the higher number of *Identify* sentences in the conventional visualisation condition is that participants may have felt the need to restate and emphasise information that was less visually highlighted, as opposed to DS-enhanced where key values were already emphasised. This could have led participants in the DS-enhanced condition to move beyond simple recall and engage more frequently at the *Understand* level and reach higher cognitive levels, such as *Apply*, *Analyse*, and *Create*. Interestingly, participants using conventional visualisations produced more *Evaluate* sentences. Similar to the discussion on RQ1, this could be attributed to conventional visualisations being more open-ended, thus encouraging interpretation and critical thinking. In contrast, DS-enhanced visualisations, which are designed to guide users toward a specific message or narrative, may reduce the need for critical evaluation, such as identifying missing elements that could influence interpretation. This hypothesis aligns with the work by Kong et al. [46] on the effect of visualisation titles which found that participants were often unaware of the potential bias in visualisations leading to misconceptions of information. Our work extends their findings suggesting that data storytelling elements make the visualisation more convincing and there is less awareness of the limitations or biases of the information presented and the way it is presented. Rogha et al. [69] hypothesised that the passive data interpretation could be due to the cognitive load induced by the narrative elements. In the work of Garretón et al. [29], visualisations are effective in eliciting emotional impact and driving attitude change. Building on this, we argue that while the persuasive power of such visualisations can be instrumental in driving action, such as addressing urgent issues like the water crisis [29], this same persuasiveness may have unintended consequences. Specifically, these visualisations could inadvertently limit users' ability to question the presented narrative, potentially leading to over-reliance on the guided interpretation provided.

5.1.4 RQ4: User Perception. Lastly, regarding user preferences, the majority of participants preferred DS-enhanced over conventional visualisations for the placement of contextual information and clear distinctions between countries, while conventional visualisations were preferred for their simplicity and minimal clutter. This aligns with the findings by Yen et al. [86] where participants perceived visualisations as useful when the text was presented in shorter segments that focused on one aspect at a time rather than all together. The idea that DS enhancements, such as annotations and explanatory titles, help to contextualise information was shared by the participants from Echeverria et al. [22], Milesi and Martinez-Maldonado [63], Stokes and Hearst [76] who praised textual elements for their ability to make insights understandable and accessible. Similarly, Stokes et al. [77] found that participants preferred charts with a narrative or story to a text-only version of the data or visualisations without text annotations. A limitation of their work was the use of synthetically generated line charts, which lacked the complexity of real-world examples. Our study addresses this gap by providing empirical validation of Stokes and Hearst [76], Stokes et al. [77] observations with real-world datasets, particularly emphasising the importance of placing contextual information near relevant data points. Additionally, the results are congruent with Bateman et al. [7] where visualisations with DS elements were perceived as more attractive, enjoyable, and easier to read and remember. Lastly, it is worth noting that there are individual preferences for information presentation or omission [56]. Lundgard and Satyanarayan [56] advocate for adaptable descriptions that balance elemental, perceptual, and contextual information to meet diverse user needs in natural language descriptions of visualisations. We propose that this principle also applies to contextual annotations in DS-enhanced visualisations.

5.2 Implications for Research

In the literature, DS-enhancements have been a focal point for their perceived ability to facilitate the sensemaking process and more effectively and efficiently guide individuals towards complex insights compared to conventional data visualisations [10, 18, 71]. This study, along with the work of Shao et al. [73], has suggested that with respect to effectiveness, this idea holds for tasks that require a low level of cognitive processing. However, our work indicates that for tasks that require higher order thinking DS-enhancement may be less effective at facilitating the sensemaking process such that individuals can uncover patterns and insights on their own [15].

This indicates that the complexity of a task can significantly impact the *effectiveness* of DS-enhanced visualisations. This presents an interesting direction for future works in establishing whether the nature of the task (e.g., written, verbal, etc.) can impact the effectiveness just as much as the complexity of the task.

In terms of efficiency, while Shao et al. [73] observed no improvements in speed between DS-enhanced and conventional visualisations, this study indicated that DS-enhancements can significantly impact how quickly individuals are able to extract insights from complex data. This is consistent with the findings from De Simone et al. [19], Echeverria et al. [22], and Pozdniakov et al. [68] who argue that DS-enhancements enable a deeper understanding of the presented narrative and thus facilitate faster interpretation of the material. This research could be expanded to investigate the specific DS-enhancements that aid in efficient extraction of insights or examine whether the improvements in efficiency are also observed different storytelling genres [71], such as data videos [3, 81] or data comics [5, 62].

This paper contributes to the InfoVis and HCI literature by building on prior frameworks, such as those proposed by Adar and Lee [1] and Shao et al. [73], to systematically assess the role of DS in supporting cognitive tasks across the full spectrum of complexity, as classified Bloom's taxonomy. Building on this theoretical foundation, we demonstrate its applicability to real-world tasks requiring both lower-order and higher-order skills through a large-scale empirical evaluation of DS-enhanced visualisations. Our study spans all levels of Bloom's taxonomy, and provides insights into how DS can support higher-order cognitive levels – *Apply* to *Create* – which have been underexplored in the literature but are critical for communicating complex ideas in practice [64, 72]. Unlike the work of Lee-Robbins et al. [51], which focuses primarily on design specifications and user preferences, this study evaluates how these design principles translate into measurable improvements in task performance across varying levels of cognitive complexity. Additionally, we introduce a human-in-the-loop-AI task generation process, which combines large language models, domain-specific validation tools, and human oversight to create tasks aligned with Bloom's taxonomy. This approach provides a framework for designing cognitively diverse tasks, and similar to Yan et al. [82, 83], highlights the potential for human-AI collaboration in the HCI and InfoVis community.

5.3 Implications for Design and Practice

The different effects of DS-enhanced and conventional visualisations across various cognitive levels present distinct challenges and opportunities for researchers, practitioners, and designers. DS elements can significantly improve the effectiveness **(RQ1)** in understanding tasks and improve the efficiency **(RQ2)** in higher-order cognitive tasks requiring interpretation and critical thinking. Educators and trainers could consider adding DS elements based on specific learning objectives and the expected cognitive engagement levels of their audience. This could lead to higher learning efficiency and effectiveness, particularly in data-driven disciplines.

Despite the advantages of DS-enhanced visualisations in improving effectiveness in the *Understanding* tasks and efficiency across most cognitive levels, it is noteworthy that in the *Create* task, where users composed text based on the visualisations, participants using DS-enhanced produced fewer *Evaluate* sentences (**RQ3**). Although this difference was not statistically significant, it warrants further investigation. This finding suggests a potential risk of DS-enhanced visualisations in contexts such as news or sensitive topics, where presenting a single narrative may inadvertently discourage critical thinking and reduce users' inclination to question underlying assumptions or consider missing information not presented in the graph. Understanding this effect more deeply could highlight an important limitation of data storytelling, especially when fostering critical evaluation is essential [75]. This is consistent with the concerns raised by Milesi and Martinez-Maldonado [63], who discussed the potential for critical data points or narratives to be inadvertently obscured by the inclusion of DS enhancements.

Based on the results of this study, we recommend the following:

- For complex cognitive tasks, DS-enhancements did not support the correctness (**RQ1**) but tended to facilitate more *efficient* information retrieval (**RQ2**). Consequently, conventional visualisations should be enhanced with DS elements in scenarios where the primary objective is to complete complex tasks more efficiently.
- The relatively low level of critical evaluation in the DS-enhanced condition compared to the conventional visualisations (**RQ3**) suggests that while DS-enhancements can be leveraged to guide attention, they may lead to individuals interpreting the data more passively compared to those who explore the visualisations and discover the insights themselves. Therefore, depending on the context, it is important to balance the implementation of DS-enhancements to ensure that the intended message is still clear while also permitting critical evaluation of the data. For example, by using explicit elicitation techniques like contrastive narratives [69] or interactive elements [64, 74].
- Given that not all participants preferred DS-enhanced visualisations over conventional visualisations (**RQ4**), designers should consider creating adaptable visualisation tools that can modify the level of narrative complexity and data storytelling elements based on user preferences [52], feedback, or the intended cognitive load of the task [16]. As noted by Figueiras [27] if a user is proficient in the task, DS-enhanced visualisations may be perceived as "*boring*". Conversely, users who are not proficient in the task may find highly exploratory visualisations difficult to engage with or comprehend [27]. Therefore, the integration of user control features, such as the ability to toggle between more and less complex narrative features, may help accommodate a wider range of users and use cases.

5.4 Limitations and Future Works

A limitation of this study is the focus on line charts and maps as visualisation types. This choice was made to minimise potential fatigue due to the survey's length [2] and to maintain consistency with the real-world context of *Our World in Data*, whose preexisting data-driven narratives constrained the design to align with the information and presentation of the original visualisations. As a result, factors such as variations in contextual annotations (explored

in Stokes et al. [77]) or DS-enhancements were not comprehensively explored. Instead, this study prioritised observing differences in task-solving performance across varying complexities within a defined DS-enhancement framework. Future studies should expand on these findings by exploring a broader range of visualisations (e.g., bar charts, scatter plots) and storytelling components. This study validates findings from studies that used synthetic data (e.g., Stokes et al. [77]) with real-world visualisations, serving as a foundation for future research to explore the generalisability of these results in more realistic contexts, such as full articles [69] or interactive narratives [64, 74]. For example, including breaking the fourth wall techniques, such as the *Golden Hook* for capturing interest and the *Magic Mirror* for tailoring insights as done in Shi et al. [74]; as well as visual narrative flow elements including story layout and progression [60], or sequencing techniques [36].

While participants were instructed to not use external resources to answer questions and "*I am not sure*" was given as an option for the multiple-choice questions, the online nature of this study means that it is possible that participants used search engines or GenAI to answer the questions. Despite efforts to exclude participants suspected of using GenAI in the written tasks, based on indicators such as words per minute, specific vocabulary, and grammatical structures, it remains difficult to completely verify the use of external sources. Consequently, conducting this study in a controlled in-person setting could mitigate the risk of cheating and obtain a richer qualitative dataset as participants work through the questions.

Another limitation of the study is the open-ended nature of the *Create* task. No specific instructions were provided to participants outside of the general prompt asking them to write about the relationship between the visualisations. This left both the interpretation of the question and the content of their response to the discretion of the participants. As there is variation in the participants' demographics (age, background, country of origin, etc.) this may have influenced the responses to RQ3.

6 Conclusion

This study compared data storytelling enhanced visualisations to conventional visualisations across tasks of varying complexity based on Bloom's taxonomy. In particular, we focused on line charts and choropleth maps to communicate the data stories and support task completion. The results show that the effectiveness and efficiency are not significantly affected in the simpler tasks like extracting values from the graphs (*Identify*). Whereas, it does improve the effectiveness in *Understanding* tasks and the efficiency in higher-thinking tasks ranging from *Understanding* to *Evaluating*. Interestingly, DS-enhanced visualisations reduced the number of *Evaluate* sentences in creation tasks, suggesting that they may decrease critical thinking by making users overly confident in the presented data. This highlights potential risks when using DS-enhanced visualisations in sensitive contexts, such as news or policy discussions, where critical evaluation is crucial. Overall, most participants preferred DS-enhanced visualisations for their clarity and contextual information, while some favoured the simplicity of conventional visualisations. These findings suggest that visualisation design should be adaptable to user preferences, task complexity, and task goals.

References

[1] Eytan Adar and Elsie Lee. 2021. Communicative Visualizations as a Learning Problem. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb 2021), 946–956. https://doi.org/10.1109/TVCG.2020.3030375

[2] Fereshteh Amini, Matthew Brehmer, Gordon Bolduan, Christina Elmer, Benjamin Wiederkehr, Nathalie Henry Riche, Sheelagh Carpendale, Christophe Hurter, Nicholas Diakopoulos, Nathalie Henry Riche, Sheelagh Carpendale, Christophe Hurter, and Nicholas Diakopoulos. 2018. Evaluating Data-Driven Stories and Storytelling Tools. In *Data-Driven Storytelling* (1 ed.). Vol. 1. CRC Press, United Kingdom, 249–286.

[3] Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Christophe Hurter, and Pourang Irani. 2015. Understanding Data Videos: Looking at Narrative Visualization through the Cinematography Lens. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1459–1468. https://doi.org/10.1145/2702123.2702431

[4] Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A Von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence* 5 (2022), 903077.

[5] Benjamin Bach, Nathalie Henry Riche, Sheelagh Carpendale, and Hanspeter Pfister. 2017. The Emerging Genre of Data Comics. *IEEE Computer Graphics and Applications* 37, 3 (May 2017), 6–13. https://doi.org/10.1109/MCG.2017.33

[6] Benjamin Bach, Zezhong Wang, Matteo Farinella, Dave Murray-Rust, and Nathalie Henry Riche. 2018. Design Patterns for Data Comics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173612

[7] Scott Bateman, Regan L. Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 2010. Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2573–2582. https://doi.org/10.1145/1753326.1753716

[8] Benjamin S. Bloom. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Vol. Handbook 1: Cognitive Domain. Longmans, Green and Co., London.

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

[10] Jeremy Boy, Francoise Detienne, and Jean-Daniel Fekete. 2015. Storytelling in Information Visualizations: Does it Engage Users to Explore Data?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1449–1458. https://doi.org/10.1145/2702123.2702452

[11] Jeremy Boy, Anshul Vikram Pandey, John Emerson, Margaret Satterthwaite, Oded Nov, and Enrico Bertini. 2017. Showing People Behind Data: Does Anthropomorphizing Visualizations Elicit More Empathy for Human Rights Data?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5462–5474. https://doi.org/10.1145/3025453.3025512

[12] Andrea Bravo and Anja M Maier. 2020. Immersive visualisations in design: Using augmented reality (AR) for information presentation. In *Proceedings of the Design Society: DESIGN Conference*, Vol. 1. Cambridge University Press, Cambridge University Press, 1215–1224. https://doi.org/10.1017/dsd.2020.33

[13] Alyxander Burns, Cindy Xiong, Steven Franconeri, Alberto Cairo, and Narges Mahyar. 2020. How to evaluate data visualizations across different levels of understanding. In *2020 IEEE Workshop on Evaluation and Beyond - Methodological Approaches to Visualization (BELIV)*. 19–28. https://doi.org/10.1109/BELIV51497.2020.00010

[14] Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems* 18, 1 (2003), 32–39.

[15] Fabio C. Campos, June Ahn, Daniela K. DiGiacomo, Ha Nguyen, and Maria Hays. 2021. Making Sense of Sensemaking: Understanding How K–12 Teachers and Coaches React to Visual Analytics. *Journal of Learning Analytics* 8, 3 (2021), 60–80. https://doi.org/10.18608/jla.2021.7113

[16] Giuseppe Carenini, Cristina Conati, Enamul Hoque, Ben Steichen, Dereck Toker, and James Enns. 2014. Highlighting interventions and user differences: informing adaptive information visualization support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1835–1844.

https://doi.org/10.1145/2556288.2557141

[17] Qing Chen, Zhen Li, Ting-Chuen Pong, and Huamin Qu. 2019. Designing Narrative Slideshows for Learning Analytics. In *2019 IEEE Pacific Visualization Symposium (PacificVis)*. Institute of Electrical and Electronics Engineers, Bangkok, Thailand, 237–246. https://doi.org/10.1109/PacificVis.2019.00036

[18] Mohammad Kamel Daradkeh. 2021. An empirical examination of the relationship between data storytelling competency and business performance: The mediating role of decision-making quality. *Journal of Organizational and End User Computing* 33 (2021). Issue 5. https://doi.org/10.4018/JOEUC.20210901.oa3

[19] Flavia De Simone, Roberta Presta, Federica Protti, et al. 2014. Evaluating data storytelling strategies: A case study on urban changes. *IARIA Cognitive* (2014), 250–255.

[20] Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. Observations on Typing from 136 Million Keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174220

[21] Brent Dykes. 2020. *Effective Data Storytelling: How to Drive Change with Data, Narrative, and Visuals* (1st edition ed.). Wiley, Hoboken, New Jersey.

[22] Vanessa Echeverria, Roberto Martinez-Maldonado, Simon Buckingham Shum, Katherine Chiluiza, Roger Granda, and Cristina Conati. 2018. Exploratory versus Explanatory Visual Learning Analytics: Driving Teachers' Attention through Educational Data Storytelling. *Journal of Learning Analytics* 5, 3 (2018), 73—97. https://doi.org/10.18608/jla.2018.53.6

[23] Micheline Elias, Marie-Aude Aufaure, and Anastasia Bezerianos. 2013. Storytelling in Visual Analytics Tools for Business Intelligence. In *Human-Computer Interaction – INTERACT 2013*, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 280–297.

[24] Stephanie D. H. Evergreen. 2020. *Effective Data Visualization: The Right Chart for the Right Data* (2 ed.). SAGE Publications, Inc., Thousand Oaks, California.

[25] Gloria Milena Fernandez-Nieto, Roberto Martinez-Maldonado, Vanessa Echeverria, Kirsty Kitto, Dragan Gašević, and Simon Buckingham Shum. 2024. Data Storytelling Editor: A Teacher-Centred Tool for Customising Learning Analytics Dashboard Narratives. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (Kyoto, Japan) *(LAK '24)*. Association for Computing Machinery, New York, NY, USA, 678–689. https://doi.org/10.1145/3636555.3636930

[26] Marta Ferreira, Miguel Coelho, Valentina Nisi, and Nuno Jardim Nunes. 2021. Climate Change Communication in HCI: a Visual Analysis of the Past Decade. In *Proceedings of the 13th Conference on Creativity and Cognition* (Virtual Event, Italy). Association for Computing Machinery, New York, NY, USA, Article 5, 16 pages. https://doi.org/10.1145/3450741.3466774

[27] Ana Figueiras. 2014. Narrative Visualization: A Case Study of How to Incorporate Narrative Elements in Existing Visualizations. In *2014 18th International Conference on Information Visualisation*. 46–52. https://doi.org/10.1109/IV.2014.79

[28] Alexia Galati, Riley Schoppa, and Aidong Lu. 2021. Exploring the SenseMaking Process through Interactions and fNIRS in Immersive Visualization. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2714–2724. https://doi.org/10.1109/TVCG.2021.3067693

[29] Manuela Garretón, Francesca Morini, Pablo Celhay, Marian Dörk, and Denis Parra. 2024. Attitudinal Effects of Data Visualizations and Illustrations in Data Stories. *IEEE Transactions on Visualization and Computer Graphics* 30, 7 (July 2024), 4039–4054. https://doi.org/10.1109/TVCG.2023.3248319

[30] Lily W. Ge, Maryam Hedayati, Yuan Cui, Yiren Ding, Karen Bonilla, Alark Joshi, Alvitta Ottley, Benjamin Bach, Bum Chul Kwon, David N. Rapp, Evan Peck, Lace M. Padilla, Michael Correll, Michelle A. Borkin, Lane Harrison, and Matthew Kay. 2024. Toward a More Comprehensive Understanding of Visualization Literacy. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 494, 7 pages. https://doi.org/10.1145/3613905.3636289

[31] Nahum Gershon and Ward Page. 2001. What storytelling can do for information visualization. *Commun. ACM* 44, 8 (aug 2001), 31–37. https://doi.org/10.1145/381641.381653

[32] Jiangang Hao, Alina A von Davier, Victoria Yaneva, Susan Lottridge, Matthias von Davier, and Deborah J Harris. 2024. Transforming assessment: The impacts and implications of large language models and generative ai. *Educational Measurement: Issues and Practice* 43, 2 (2024), 16–29.

[33] Bastian Herre, Pablo Arriagada, and Max Roser. 2023. State Capacity. *Our World in Data* (2023). https://ourworldindata.org/state-capacity.

[34] Martin Hicks. 2009. *Perceptual and Design Principles for Effective Interactive Visualisations*. Springer London, London, 155–174. https://doi.org/10.1007/978-1-84800-269-2_7

[35] Jessica Hullman and Nick Diakopoulos. 2011. Visualization Rhetoric: Framing Effects in Narrative Visualization. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec 2011), 2231–2240. https://doi.org/10.1109/TVCG.2011.255

[36] Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. 2013. A Deeper Understanding of Sequence in Narrative Visualization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec 2013), 2406–2415. https://doi.org/10.1109/TVCG.2013.119

[37] Noah Iliinsky and Julie Steele. 2011. *Designing data visualizations: Representing informational Relationships*. O'Reilly Media, Inc.

[38] Sehrish Iqbal, Mladen Rakovic, Guanliang Chen, Tongguang Li, Jasmine Bajaj, Rafael Ferreira Mello, Yizhou Fan, Naif Radi Aljohani, and Dragan Gasevic. 2024. Towards Improving Rhetorical Categories Classification and Unveiling Sequential Patterns in Students' Writing. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (Kyoto, Japan) *(LAK '24)*. Association for Computing Machinery, New York, NY, USA, 656–666. https://doi.org/10.1145/3636555.3636927

[39] Sehrish Iqbal, Mladen Rakovic, Guanliang Chen, Tongguang Li, Rafael Ferreira Mello, Yizhou Fan, Giuseppe Fiorentino, Naif Radi Aljohani, and Dragan Gasevic. 2023. Towards automated analysis of rhetorical categories in students essay writings using Bloom's taxonomy. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. 418–429.

[40] Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2024. Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 935, 17 pages. https://doi.org/10.1145/3613904.3642596

[41] Dave S Kerby. 2014. The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology* 3 (2014), 11–IT.

[42] Dae Hyun Kim, Vidya Setlur, and Maneesh Agrawala. 2021. Towards Understanding How Readers Integrate Charts and Captions: A Case Study with Line Charts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 610, 11 pages. https://doi.org/10.1145/3411764.3445443

[43] Young-Ho Kim, Bongshin Lee, Arjun Srinivasan, and Eun Kyoung Choe. 2021. Data@Hand: Fostering Visual Exploration of Personal Data on Smartphones Leveraging Speech and Touch Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 462, 17 pages. https://doi.org/10.1145/3411764.3445421

[44] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. A data–frame theory of sensemaking. In *Expertise out of context*. Psychology Press, 118–160.

[45] Cole Nussbaumer Knaflic. 2015. *Storytelling with Data: A Data Visualization Guide for Business Professionals*. John Wiley & Sons, Inc., Hoboken, New Jersey.

[46] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2018. Frames and Slants in Titles of Visualizations on Controversial Topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174012

[47] Robert Kosara and Jock Mackinlay. 2013. Storytelling: The Next Step for Visualization. *Computer* 46, 5 (May 2013), 44–50. https://doi.org/10.1109/MC.2013.36

[48] David R. Krathwohl. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice* 41, 4 (2002), 212–218. https://doi.org/10.1207/s15430421tip4104_2

[49] Randy Krum. 2013. *Cool infographics: Effective communication with data visualization and design*. John Wiley & Sons.

[50] Ryan Lau. 2023. BloomBERT: A Task Complexity Classifier. https://github.com/RyanLauQF/BloomBERT.

[51] Elsie Lee-Robbins, Shiqing He, and Eytan Adar. 2022. Learning Objectives, Insights, and Assessments: How Specification Formats Impact Design. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan 2022), 676–685. https://doi.org/10.1109/TVCG.2021.3114811

[52] Zhengliang Liu, R. Jordan Crouser, and Alvitta Ottley. 2020. Survey on Individual Differences in Visualization. *Computer Graphics Forum* 39, 3 (2020), 693–712. https://doi.org/10.1111/cgf.14033 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14033

[53] Zhicheng Liu and John Stasko. 2010. Mental Models, Visual Reasoning and Interaction in Information Visualization: A Top-down Perspective. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (Nov 2010), 999–1008. https://doi.org/10.1109/TVCG.2010.177

[54] Meagan B. Loerakker, Jasmin Niess, Marit Bentvelzen, and Paweł W. Woźniak. 2024. Designing Data Visualisations for Self-Compassion in Personal Informatics. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 169 (jan 2024), 22 pages. https://doi.org/10.1145/3631448

[55] Joy Lowe and Machdel Matthee. 2020. Requirements of Data Visualisation Tools to Analyse Big Data: A Structured Literature Review. In *Responsible Design, Implementation and Use of Information and Communication Technology*, Marié Hattingh, Machdel Matthee, Hanlie Smuts, Ilias Pappas, Yogesh K. Dwivedi, and Matti Mäntymäki (Eds.). Springer International Publishing, Cham, 469–480.

[56] Alan Lundgard and Arvind Satyanarayan. 2022. Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan 2022), 1073–1083. https://doi.org/10.1109/TVCG.2021.3114770

[57] William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information

Sciences Institute Los Angeles.
[58] Kim. Marriott, Falk. Schreiber, Tim. Dwyer, Karsten. Klein, Nathalie Henry. Riche, Takayuki. Itoh, Wolfgang. Stuerzlinger, and Bruce H. Thomas. 2018. *Immersive Analytics* (1st ed. 2018. ed.). Springer International Publishing, Cham.
[59] Roberto Martinez-Maldonado, Vanessa Echeverria, Gloria Fernandez Nieto, and Simon Buckingham Shum. 2020. From Data to Insights: A Layered Storytelling Approach for Multimodal Learning Analytics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376148
[60] S. McKenna, N. Henry Riche, B. Lee, J. Boy, and M. Meyer. 2017. Visual Narrative Flow: Exploring Factors Shaping Data Visualization Story Reading Experiences. *Computer Graphics Forum* 36, 3 (2017), 377–387. https://doi.org/10.1111/cgf.13195 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13195
[61] Paola Mejia-Domenzain, Eva Laini, Seyed Parsa Neshaei, Thiemo Wambsganss, and Tanja Käser. 2023. Visualizing Self-Regulated Learner Profiles in Dashboards: Design Insights from Teachers. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*. Springer Nature Switzerland, Cham, 619–624.
[62] Mikaela E Milesi, Riordan Alfredo, Vanessa Echeverria, Lixiang Yan, Linxuan Zhao, Yi-Shan Tsai, and Roberto Martinez-Maldonado. 2024. "It's Really Enjoyable to See Me Solve the Problem like a Hero": GenAI-enhanced Data Comics as a Learning Analytics Tool. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.1145/3613905.3651111
[63] Mikaela Elizabeth Milesi and Roberto Martinez-Maldonado. 2024. Data Storytelling in Learning Analytics? A Qualitative Investigation into Educators' Perceptions of Benefits and Risks. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (Kyoto, Japan) *(LAK '24)*. Association for Computing Machinery, New York, NY, USA, 167–177. https://doi.org/10.1145/3636555.3636865
[64] Eric Mörth, Stefan Bruckner, and Noeska N. Smit. 2023. ScrollyVis: Interactive Visual Authoring of Guided Dynamic Narratives for Scientific Scrollytelling. *IEEE Transactions on Visualization and Computer Graphics* 29, 12 (Dec 2023), 5165–5177. https://doi.org/10.1109/TVCG.2022.3205769
[65] Ifeanyi Glory Ndukwe and Ben Kei Daniel. 2020. Teaching analytics, value and tools for teacher data literacy: A systematic and tripartite approach. *International Journal of Educational Technology in Higher Education* 17, 1 (2020), 1–31.
[66] Carolina Nobre, Kehang Zhu, Eric Mörth, Hanspeter Pfister, and Johanna Beyer. 2024. Reading Between the Pixels: Investigating the Barriers to Visualization Literacy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 197, 17 pages. https://doi.org/10.1145/3613904.3642760
[67] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.
[68] Stanislav Pozdniakov, Roberto Martinez-Maldonado, Yi-Shan Tsai, Vanessa Echeverria, Namrata Srivastava, and Dragan Gasevic. 2023. How Do Teachers Use Dashboards Enhanced with Data Storytelling Elements According to Their Data Visualisation Literacy Skills?. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (Arlington, TX, USA) *(LAK2023)*. Association for Computing Machinery, New York, NY, USA, 89–99. https://doi.org/10.1145/3576050.3576063
[69] Milad Rogha, Subham Sah, Alireza Karduni, Douglas Markant, and Wenwen Dou. 2024. The Impact of Elicitation and Contrasting Narratives on Engagement, Recall and Attitude Change With News Articles Containing Data Visualization. *IEEE Transactions on Visualization and Computer Graphics* 30, 7 (July 2024), 4375–4389. https://doi.org/10.1109/TVCG.2024.3355884
[70] Lindy Ryan. 2016. *The Visual Imperative: Creating a Visual Culture of Data Discovery*. Elsevier Inc., Cambridge, MA.
[71] E. Segel and J. Heer. 2010. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1139–1148. https://doi.org/10.1109/TVCG.2010.179
[72] Doris Seyser and Michael Zeiller. 2018. Scrollytelling – An Analysis of Visual Storytelling in Online Journalism . In *2018 22nd International Conference Information Visualisation (IV)*. IEEE Computer Society, Los Alamitos, CA, USA, 401–406. https://doi.org/10.1109/iV.2018.00075
[73] Hongbo Shao, Roberto Martinez-Maldonado, Vanessa Echeverria, Lixiang Yan, and Dragan Gasevic. 2024. Data Storytelling in Data Visualisation: Does it Enhance the Efficiency and Effectiveness of Information Retrieval and Insights Comprehension?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 195, 21 pages. https://doi.org/10.1145/3613904.3643022
[74] Yang Shi, Tian Gao, Xiaohan Jiao, and Nan Cao. 2023. Breaking the Fourth Wall of Data Stories through Interaction. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan 2023), 972–982. https://doi.org/10.1109/TVCG.2022.3209409
[75] Arjun Srinivasan, Steven M. Drucker, Alex Endert, and John Stasko. 2019. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan 2019), 672–681. https://doi.org/10.1109/TVCG.2018.2865145
[76] Chase Stokes and Marti Hearst. 2022. Why More Text is (Often) Better: Themes from Reader Preferences for Integration of Charts and Text. arXiv:2209.10789 [cs.HC] https://arxiv.org/abs/2209.10789
[77] Chase Stokes, Vidya Setlur, Bridget Cogley, Arvind Satyanarayan, and Marti A. Hearst. 2023. Striking a Balance: Reader Takeaways and Preferences when Integrating Text and Charts. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan 2023), 1233–1243. https://doi.org/10.1109/TVCG.2022.3209383
[78] Taewoo Nam Sungsoo Hwang and Hyunsang Ha. 2021. From evidence-based policy making to data-driven administration: proposing the data vs. value framework. *International Review of Public Administration* 26, 3 (2021), 291–307. https://doi.org/10.1080/12294659.2021.1974176 arXiv:https://doi.org/10.1080/12294659.2021.1974176
[79] Edward R Tufte. 2001. *The visual display of quantitative information*. Vol. 2. Graphics press Cheshire, CT. https://doi.org/10.4135/9781071812082.n670
[80] Zezhong Wang, Lovisa Sundin, Dave Murray-Rust, and Benjamin Bach. 2020. Cheat Sheets for Data Visualization Techniques. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376271
[81] Xian Xu, Aoyu Wu, Leni Yang, Zheng Wei, Rong Huang, David Yip, and Huamin Qu. 2023. Is It the End? Guidelines for Cinematic Endings in Data Videos. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 171, 16 pages. https://doi.org/10.1145/3544548.3580701
[82] Lixiang Yan, Roberto Martinez-Maldonado, Yueqiao Jin, Vanessa Echeverria, Mikaela Milesi, Jie Fan, Linxuan Zhao, Riordan Alfredo, Xinyu Li, and Dragan Gašević. 2024. From Data Stories to Dialogues: A Randomised Controlled Trial of Generative AI Agents and Data Storytelling in Enhancing Data Visualisation Comprehension. *arXiv preprint arXiv:2409.11645* (2024).
[83] Lixiang Yan, Linxuan Zhao, Vanessa Echeverria, Yueqiao Jin, Riordan Alfredo, Xinyu Li, Dragan Gaševi'c, and Roberto Martinez-Maldonado. 2024. VizChat: enhancing learning analytics dashboards with contextualised explanations using multimodal generative AI chatbots. In *International Conference on Artificial Intelligence in Education*. Springer, 180–193.
[84] Yalong Yang, Wenyu Xia, Fritz Lekschas, Carolina Nobre, Robert Krüger, and Hanspeter Pfister. 2022. The Pattern is in the Details: An Evaluation of Interaction Techniques for Locating, Searching, and Contextualizing Details in Multivariate Matrix Visualizations. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 84, 15 pages. https://doi.org/10.1145/3491102.3517673
[85] Nathan Yau. 2013. *Data Points: Visualization That Means Something* (1 ed.). Wiley, Newark.
[86] Yu-Chun Grace Yen, Joy O. Kim, and Brian P. Bailey. 2020. Decipher: An Interactive Visualization Tool for Interpreting Unstructured Design Feedback from Multiple Providers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376380
[87] Dominyk Zdanovic, Tanja Julie Lembcke, and Toine Bogers. 2022. The Influence of Data Storytelling on the Ability to Recall Information. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. ACM, New York, NY, USA, 67–77. https://doi.org/10.1145/3498366.3505755
[88] Jiayi Eris Zhang, Nicole Sultanum, Anastasia Bezerianos, and Fanny Chevalier. 2020. DataQuilt: Extracting Visual Elements from Images to Craft Pictorial Visualizations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376172
[89] Jiyun Zu, Ikkyu Choi, and Jiangang Hao. 2023. Automated distractor generation for fill-in-the-blank items using a prompt-based learning approach. *Psychological Testing and Assessment Modeling* 65, 2 (2023), 55–75.

A Example Questions Aligned with Bloom's Taxonomy

Table 3: Examples of questions aligned with Bloom's Taxonomy levels used in the study. The examples are drawn from visualisation A, with questions ranging from multiple-choice formats to open-ended tasks for the *Create* level.

Level & Definition	Example Question	Possible Answers
Identify: Retrieving relevant knowledge (e.g., Recognising)	Which Empire had overseas colonies as early as the 1460s, as shown in the graph?	French; Spanish; **Portuguese**; British
Understand: Determining the meaning of instructional messages, including oral, written, and graphic communication (e.g., Explaining)	Which is the most probable explanation for the observed trend in the Spanish Empire line from 1800-1850?	Increased military conquests by Spain in Africa; **Wars of independence in Latin America**; Napoleonic Wars that led to the cession of colonies to France; The liberal principles in the 1812 Constitution were implemented throughout the Spanish Empire
Apply: Carrying out or using a procedure in a given situation. (e.g.,Implementing)	How would the graph change if the British Empire had captured half of Spain's colonies in 1700? Select all that apply.	**The distance between the French plot line and the British plot line would increase in 1700**; The distance between the British plot line and the Spanish plot line would decrease in 1700; The British plot line would reach its maximum in 1700; The Portuguese plot line would be greater than the Spanish plot line in 1700; The distance between the Portuguese plot line and the Spanish plot line would increase in 1700
Analyse: Breaking material into its constituent parts and detecting how the parts relate to one another and to an overall structure or purpose. (e.g, Differentiating, Organizing)	According to the graph, how does the colonisation pattern of the French Empire compare to that of the British Empire? Select all that apply.	The British Empire always had more colonies than the French Empire; The British Empire did not lose colonies before 1900, while the French Empire did; **The British and French Empires started to decline simultaneously**; The British and French Empires reach their peak number of colonies in the same year; The British Empire gained overseas colonies years before the French Empire
Evaluate: Making judgments based on criteria and standards. (e.g., Critiquing)	What other pieces of information would you need to evaluate the validity of the following sentence: "*The Spanish Empire impacted negatively more people than the French Empire*". Select only the relevant options.	The duration of colonialism under the French and Spanish Empires; The number of colonies over time for the British and French Empires; **The population size of the colonies under the French and Spanish Empires**; **Economic and cultural policies implemented by the French and Spanish Empires in their colonies**; The information on the graph is sufficient
Create: Putting elements together to form a novel, coherent whole or make an original product. (e.g., Generating, Planning, Producing)	Based on the visualisations, write a text on the relationship between the topics presented in the graphs. Your response should be 8 to 12 sentences long and should reference specific data points or trends from the graphs to support your ideas.	**[Open Ended Task]** Example of an extract of an answer: *The two figures can be linked by the historical trends. Figure 1 shaped the data shown in Figure 2. For example, Figure 1 shows the rise and fall of Spanish colonialism. We see likely downstream consequences of this reflected in Figure 2, with certain countries having higher rates of uncontrolled territory, as with Mexico, Peru, Colombia, Venezuela.*

B Create Task Labelling Rubric

Table 4: Rubric used to categorise sentences from participants' responses in the *Create* task, based on the cognitive levels of Bloom's Taxonomy.

Level	Description	Example
Identify	Factual information from a graph without interpretation. Reference to values and information in one graph.	For example, the Spanish started colonizing the Americas in the 1400s and the British and French in the 1600s as can be seen from the first graph.
Understand	Explaining only one graph and interpreting the values.	The factors that affect GDP of a country including civil wars, politics, government policies, resources, investments and local and international trade.
Apply	Applying information from a graph to a new situation or hypothetical scenario about future trends.	For Syria, the civil war is still ongoing, so it is likely their state capacity is still low in 2022.
Analysis	Explanation linking both topics (reference to two graphs). Identifying potential relationships between the two topics.	The length of time that former colonies have been independent correlates with how much control those countries have over their territory.
Evaluate	Critical assessment of information and underlying assumptions in the graphs. Highlighting discrepancies, outliers or missing elements that affect interpretation.	More information is needed to validate the relationship as the first graph provides information until 2010 and the second graph provides information of 2022.
Create	Extending or extrapolating beyond the information provided in the graph to draw new conclusions, or propose alternative interpretations in the form of a new argument.	At the same time, rather than stability, countries with high rates of controlled land (that were post-colonial), such as South Africa, might not necessarily be a positive thing as colonialism imposed cultural, social, political norms, perhaps they had less ability for self-determination (versus other countries that had more ability to do so, even if it means higher rates of uncontrolled territory).
Other	Information not directly related to the graph to introduce another sentence (for example: topic sentences). Sentences of previous knowledge.	It's interesting to compare these visualisations between these two countries.