Where's My Data? Evaluating Visualizations with Missing Data

Hayeong Song & Danielle Albers Szafir

Fig. 1: We measured factors influencing response accuracy, data quality, and confidence in interpretation for time series data with missing values. We found that visualizations that highlight missing values have higher perceived data quality while those that break visual continuity decrease these perceptions and can bias interpretation.

Abstract—Many real-world datasets are incomplete due to factors such as data collection failures or misalignments between fused datasets. Visualizations of incomplete datasets should allow analysts to draw conclusions from their data while effectively reasoning about the quality of the data and resulting conclusions. We conducted a pair of crowdsourced studies to measure how the methods used to impute and visualize missing data may influence analysts' perceptions of data quality and their confidence in their conclusions. Our experiments used different design choices for line graphs and bar charts to estimate averages and trends in incomplete time series datasets. Our results provide preliminary guidance for visualization designers to consider when working with incomplete data in different domains and scenarios.

Index Terms—Information Visualization, Graphical Perception,Time Series Data, Data Wrangling, Imputation

1 Introduction

Visualizations allow people to analyze and interpret data to understand current phenomena and guide informed decision-making. However, analysts often must make decisions using imperfect datasets. These datasets may be missing datapoints due to factors such as failures in the data collection pipeline or fusing data at different granularities. As part of the data wrangling process, visualizations have several choices for dealing with missing data, including not encoding missing elements or *imputing* new data (calculating substitute values) based on existing data. Prior studies show that the ways we represent data influence how accurately people interpret data and change their confidence in their data and results [16, 20, 37, 47]. We hypothesize that the ways we impute and visualize missing data may also bias analysts perceptions of that data. This study aims to provide a deeper empirical understanding of visualization for missing data.

We measure how imputation and visualization techniques influence perceived confidence, data quality, and accuracy for visualizing incomplete datasets. We explore how four different categories of visualization designs employed in prior systems might manipulate perceived data quality: highlighting imputed data (e.g., making data more salient, as in highlighting), downplaying imputed data (e.g., making the data less salient, as in alpha blending), annotation imputed values (e.g., adding additional information about the imputation outcomes, such as error bars), and visually removing information (Fig. 2). We measure effect of existing techniques corresponding to these four categories of these visual attributes on perceived data quality, result confidence, and response accuracy in two common visualizations: line graphs and bar charts. While this categorization is not exhaustive, we use this categorization as a scaffold for exploring a subset of techniques used in existing visualization systems.

We also explore how methods of imputing missing values might additionally shift perceptions of data quality and bias responses. Systems use imputation to compute values that approximate missing datapoints to support analysis. As missing data is itself a type of data (it indicates no values are available), imputation allows systems to indicate where data is unexpectedly absent and provide principled approximations to avoid potential misinterpretation of absent data values [7]. Imputing values also allows systems to indicate potential threats to data quality by providing visual anchors analysts can use to readily enumerate and contextualize quality errors [5, 49]. We focus on three common imputation methods encountered in current visualization systems: ad-hoc zero-filling, local linear interpolation, and marginal means (Fig. 3).

While we commonly expect that missing data should optimally degrade perceived quality, there are many cases that run counter to this assumption. For example, we may not wish to degrade perceived quality when we can closely approximate missing values or when quality may interfere with decision speed in low-risk scenarios. We therefore evaluate how visualizations manipulate confidence relative to other

- *Hayeong Song is with the University of Colorado Boulder. E-mail: hayeong.song@colorado.edu.*
- *Danielle Albers Szafir is with the University of Colorado Boulder. E-mail: danielle.szafir@colorado.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

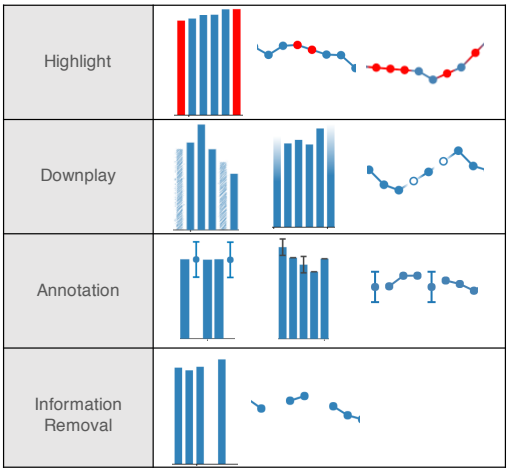


Fig. 2: We examined three distinct categories of visualizations that could encode imputed values: **highlight** and **downplay** encodings that manipulate attention, **annotation** encodings that provide additional information, and encodings that use **information removal**.

techniques to provide designers an empirical basis for visualizing missing data. We compare imputation and visualization methods in four crowdsourced studies measuring the effects of these factors on analysts' accuracy with and without imputed data, confidence in their conclusions, credibility, and perceived data quality. We found that highlighting and annotating imputed values lead to higher perceived data quality and more accurate interpretation. Downplaying imputed values or removing information associated with missing values significantly degraded perceived quality. Our findings suggest ways visualizations might leverage imputation and visualization to appropriately manipulate perceived data quality in different scenarios.

2 Related Work

Missing data is typically a challenge associated with "dirty data"—datasets containing missing data, incorrect data, misalignments, and other such anomalies that may lead to erroneous conclusions [38]. Missing data can occur throughout the data lifecycle and has significant implications for analysts' trust in data [45]. These implications can be especially problematic for data visualizations as little empirical understanding exists to guide how visualizations can balance between indicating the presence of dirty data and not distracting from or biasing of the rest of the data [35]. Our research builds such knowledge by measuring the influence of various designs for visualizing missing data.

2.1 Methods for Analyzing Incomplete Data

Missing data can arise at all points in the data lifecycle, including during data capture, storage, updates, transmission, deletions, and purges [38]. A scraping process might fail due to an interrupted script, packet loss, or memory errors. Subsets of data may be withheld due to privacy considerations [21]. Part of the process of data wrangling [25, 35] is locating missing data and deciding how to manage it. In many cases, systems choose to *impute*—estimate a substitute value for—missing data to address potential anomalies affecting dataset coverage [42].

A broad variety of methods exist for data imputation (see Little & Rubin [40] and Lajeunesse [39] for surveys). For example, hot-deck imputation samples substitute values from the current signal while cold-deck imputation uses values from other sources, such as related datasets [27] or domain heuristics [31]. Interpolation methods use weighted combinations of available data to infer missing values using methods like linear interpolation, regression, and adaptive interpolation [26]. More complex imputation methods can integrate information about the processes used to generate the dataset [46] or use machine learning and related techniques to holistically estimate missing values [2].

While an exhaustive survey of imputation methods is beyond the scope of this paper, understanding the relationship between different

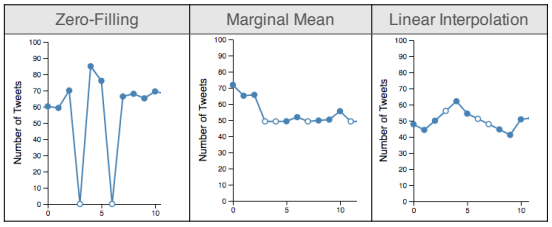


Fig. 3: We measured effects of three different imputation methods on data interpretation: zero-filling (substituting missing values with zeros), marginal means (substituting with the mean of available data), and linear interpolation of adjacent datapoints.

imputation choices and perceived data quality is critical for visualizing missing data. As Babad & Hoffer note, even if data values can be inferred with reasonable accuracy, it is important for analysts to understand when and where missing data occurs [7]. Missing data can have a significant impact on inference and decision making and can lend context to analyses. Most significantly for our work, missing data is a key component of *data quality*, a measure of the trust and suitability of data for addressing a given problem [44].

Time series data has specific considerations for data quality (see Gschwandtner et al. for a survey [30]). For example, non-uniform sampling may force interpolations. Joining data across two temporal sources with different granularities can create misalignment [6]. Measures taken at the same time may conflict. Since data is typically continuous, violations to trends may be especially salient. Due to these factors, we elect to use time series data for our study as it is common in both real-world analysis and empirical studies for visualization and these factors make it an important special case for understanding the implications of missing data for temporal analysis.

2.2 Visualizing Incomplete Data

Wong & Varga refer to missing data as *black holes* in a visualization–"a dark area of the cognitive workspace that by the absence of data indicates that one should take care [54, p.5]." They argue that it is unclear when and how visualizations should replace missing data to support sensemaking, yet it is clear that people should be able to detect and reason about missing data. Many visualization systems support data quality analysis, including quality change over time [11], data preprocessing [8], and highlighting missing, incorrect, or imputed values [9,10,23]. For example, Visplause [5] supports quality inspection for temporal data to assist analysts in inferring potential causes of missing data. Wrangler [36] uses statistical methods to help analysts impute missing values. xGobi [49], MANET [53], and VIM [50] offer visualization suites that allow analysts to understand the amount of missing data and compare different imputation methods.

Many visualization systems oriented towards specific domains or datatypes automatically process missing data. Some visualizations provide little to no visual indication of imputed data. For example, Turkay et al. [51] substitute missing values with feature means. Systems in meteorology [19] and psychology [31] interpolate missing data based on domain heuristics. Other systems leverage visual saliency to manipulate whether analyst attention is drawn to imputed values. For example, TimeSearcher uses brightly colored marks to indicate missing values [14]. Restorer uses grayscale to reduce the salience of missing spatial data and luminance to interpolate imputed values [52]. However, the influence of imputation and the corresponding visualization methods used in these systems is not well understood. We ground our exploration of imputation in current practices for missing data visualization.

2.3 Graphical Perception

Prior studies in graphical perception show how the methods used to visualize data change our interpretation of that data. For example, studies show that visualization design changes our abilities to estimate and compare statistical values [3,15,24,32] and shift our confidence in those estimations [1]. As imputed values represent uncertain information, we

draw on prior findings in uncertainty visualization to inform our study. Specific visual attributes, such as luminance, blurriness, and sketchiness, can indicate uncertainty in data and shift people's confidence in their conclusions [13, 16, 29, 37, 41]. Presenting data as "sketchy" additionally increases engagement with and willingness to critique data [55], which may have interesting ramifications for perceived data quality. Individual values can shift statistical perceptions of data [17], indicating imputed points introducing variation may potentially bias analyses. As many imputation methods provide no quantifiable measure of uncertainty, we evaluate encodings that both present either the level or the existence of uncertain information.

A handful of prior studies have explicitly evaluated the influence of visualization methods on perceptions of data quality. Xie et al. [56] measure how to communicate data quality in high dimensional data using size, brightness, and hue, and found hue and size to be strong channels for encoding quality information. Eaton et al. [21] compared how different methods for visualizing missing data in line graphs influenced accuracy and confidence in response for point-comparison and trend estimation recall tasks. They substituted missing values with zero, rendered no marks for missing data (*data absent*), and used gapped circles to indicate missing data. They found that people interpreted confidently even when critical data was missing, but found no significant differences between methods. Participants expressed an overall preference for visualizations that explicitly indicated missing data. Andreasson and Riveiro [4] conducted a similar study comparing the effects of absent data, fuzziness, and annotated absent data on analyst confidence in a decision making task. Their results showed that people had a strong preference for conditions with annotated absent data and a strong dislike for fuzziness.

Our work extends these findings by separating effects of imputation methods such as the zero-filling in Eaton et al. [21] from visualization methods, considering variable numbers of missing values, and leveraging a wider variety of visualization methods. We also evaluate bar charts in addition to line graphs as removing missing data from bar charts is indistinguishable from zero values. Prior studies indicate that the kinds of information people synthesize across bars and lines can vary [57], and these differences may significantly impact perceptions of missing data.

3 Motivation & Hypotheses

Data quality concerns how suitable a given dataset is to solve a problem or make a decision. Dimensions of data quality include several factors related to a data source (e.g., accessibility, volume, and relevance) and others relating to perceptions of the dataset (e.g., completeness, credibility, and reliability) [44]. While analysts must consider factors of a data source when choosing a dataset, the visualizations used to analyze data directly influence perceptions of that data. In this study, we measure how imputation and visualization choices impact response bias and perceptions when data is incomplete. We measure quality as a combination of confidence (how confidently can they complete a task given the data), completeness (how much data is available), credibility (how true is the data), and reliability (how correct is the data). Following best practices, we use the results from these metrics to construct a data quality scale ([18, 22, 48], c.f., §4.4).

Our inspiration for this study comes from collaborations with public health analysts. These analysts fuse data from sources of both low (e.g., social media) and high (e.g., CDC and WHO reports) quality data to develop holistic insights. Data collection errors and temporal misalignments after fusing these sources frequently lead to incomplete data. While our collaborators care about large scale patterns in this data, their imputation methods and whether or not they need to include imputed data in assessing these patterns is less well defined: analysts want to analyze patterns in light of missing data, but can often generate reasonable approximations about that data or want to know when and where data is missing to temper their decision making processes. As a result, we opt to evaluate missing data using similar methods to Jansen & Hornbæk where participants naturally integrate imputed values without explicit instructions as to how to consider those values in their estimates [34]. We measure performance using two common tasks employed by our collaborators: average and trend analysis.

We tested four categories of visualization type for communicating missing data that we encountered in the systems discussed in §2. The first category **highlights** missing data by leveraging bright colors to attract attention to missing data points (e.g., [10, 14]). The second category **downplayed** missing values by reducing the salience of imputed values relative to the rest of the data (e.g., [4, 9]). The third category used the encodings to **annotate** missing values with additional statistical information such as error bars drawing confidence from the imputation estimate (e.g., variance statistics for cold- or hot-deck methods) [9]. The fourth category used **information removal**, physically removing some element of missing values from the visualization (e.g., [4, 21]). As these semantically related to incompleteness, we anticipate that these encodings will also degrade data quality perceptions. Some tested manipulations were hybrids of these categories that examine dependencies across conditions. To mirror prior studies, we included a condition where missing data was entirely absent.

We draw our tested imputation methods from three methods we observed in existing visualization systems. **Zero-filling** substituted a single value (0) for all missing data points, as in many commercial systems. **Linear interpolation** linearly interpolated between adjacent available items (e.g., [31, 52]). **Marginal means** replaced each missing data value with the mean of all available signals (e.g., [23, 51]). For our data, zero-filling introduced the highest deviation from the original dataset, marginal means the second, and linear interpolation the lowest. While we experimented with more complex interpolation methods, we found no significant differences in our stimuli between those methods and the three selected. Figure 3 provides examples of the tested imputation and visualization categories.

Based on these conditions, we hypothesized that:

H1–Perceived data quality and response accuracy will both degrade as the amount of missing data increases.

H2–Highlighting methods will generate higher perceived data quality than downplaying and information removal methods.

H3–Linear interpolation will lead to higher perceived confidence and data quality than marginal means or zero-filling as it takes into account local trends in dataset.

H4–Imputed values will lead to higher perceived data quality than removed values.

H1 stems from the idea that completeness is a key indicator of data quality and provides a quality check for our experiment. In our experiment, data quality is measured as a combination of perceived confidence, credibility, reliability, and completeness. We anticipated people could effectively reason about missing values; therefore, no change in accuracy beyond that introduced by the amount of missing data. **H2** arises from certainty and completeness as aspects of data quality. As highlighting visualizations provide no visual indications associated with either completeness (as with information removal) or with reduced visual weight (as in downplaying), we anticipate it will lead to higher perceived quality. This corresponds with observations from Andreassen & Riveiro [4] who found evidence that "fuzzy" visualizations, correlated with downplaying, were not well-liked for decision-making with missing data [13]. We predict **H3** on the basis of potential biases introduced by zero-filled and mean values and that linear interpolation will create plausible variation in imputed values. This aligns with Correll & Heer's findings that values outside of a distribution can bias statistical perceptions in data [17]. **H4** stems directly from Eaton et al. [21], who showed a preference for visualizations using explicit visual indications of missing data.

4 Methods

We ran two 7 (visualization type) $\times$ 3 (imputation method) $\times$ 4 (percentage of missing data) full factorial within-participants studies to measure how visualization and imputation influence time series analysis, focusing on two conventional visualizations: line and bar graphs. Each study followed the same general procedure. Specific differences between the two studies are discussed in their respective sections. For

each study, we had three independent variables—visualization type, imputation method, and percentage of missing data—and five dependent variables—accuracy, confidence in response, data credibility, data reliability, and data completeness—combined to measure quality using scale construction [18].

4.1 Stimuli & Tasks

We generated each graph as a 1000×400 pixel graph using D3 [12] and Plot.ly [33] (see Fig. 1 for examples). Each graph visualizes 60 values representing the frequency of Tweets collected per minute over an hour to provide a concrete problem scenario where we often find missing data in the real-world. We simulated missing data completely at random (MCAR) by randomly removing a subset of values in each graph ($0\%, 10\%, 20\%$, or 30%). We replaced these values with imputed values computed using one of the three imputation methods described in §3 (zero-filling, linear interpolation, or marginal means). The 0% condition provided a baseline for measuring changes to our dependent variables due to data removal. The imputed values were then rendered using one of the seven candidate visualization methods per graph type (Figs. 4 and 6).

Above each graph, we provided a brief sentence contextualizing the data, a statement encouraging participants to complete the questions as quickly and accurately as possible, and a counter indicating current progress through the study. Below each graph, we enumerated five questions, answered using radio buttons. We evaluated two tasks each for line graphs and bar charts: average and trend comparison. Each task required participants to answer five questions for each stimuli with task language determined in piloting.

1. Were there more Tweets on average in the first or second half-hour? (*Averaging*)
   Is the overall rate of change larger in the first or second half-hour? (*Trend Detection*)
2. How confident are you in your response?
   1–Extremely Unconfident, 7–Extremely Confident
3. How credible is this data?
   1–Extremely Uncredible, 7–Extremely Credible
4. How complete is this data?
   1–Extremely Incomplete, 7–Extremely Complete
5. How reliable is this data?
   1–Extremely Unreliable, 7–Extremely Reliable

We chose to use averaging and trend comparison tasks in our evaluation as they forced participants to consider information from all points in the dataset and mitigated changes to the correct response and task difficulty introduced by randomly removing values. Prior studies in missing data visualization have relied on trend detection tasks (e.g., [21]), while our public health collaborators noted the importance of averaging for comparing relative frequencies across datasets.

4.1.1 Data Generation

Both noise and task difficulty may influence data perceptions and performance: noisier signals may change the effects of different imputation methods and confidence may correlate with difficulty. To control for these concerns, we used synthetic datasets to provide control over noise and difficulty. Each graph contained 60 y-values ranging from $y = 0$ to $y = 100$ uniformly spaced in time. We computed the y-values by first generating a signal from structured random noise and then adjusted each signal based on task constraints [43]. To assist with reproducing and extending our results and analyses, data and experimental infrastructures are available at `http://cmci.colorado.edu/visualab/MissingData/`.

Average Data: We generated signals using five different noise levels and considered noise as a random effect in our analyses. We then used a constraint-based optimization to adjust the mean difference between the first and last thirty points while minimizing deviation from the original random signal to control difficulty. We separated the means of the first and last half hour by 2.0, 4.0, and 6.0, randomly selecting which half hour was highest. We used this difference threshold as it achieved desirable response accuracy in prior studies [3]. For the average task, each graph visualized a randomly selected dataset from 110 total datasets generated using this method.

Trend Data: We generated signals using four different noise levels and considered noise as a random effect in our analyses. We separated the difference in the slopes of the first and last half hour by 0.5 or 0.7, randomly selecting which half hour larger overall rate of change. For the trend task, each graph visualized a randomly selected dataset from 96 total datasets generated using this method.

4.2 Procedure

Our study consisted of five phases: (1) consent, (2) screening, (3) instructional tutorial, (4) formal study, and (5) demographic questionnaire. Each participant first provided informed consent to participate in the study in accordance with our IRB protocol. We then screened participants for color vision deficiencies using a set of four Ishihara plates. Participants then received instructions about the study and were serially shown examples of each of the seven visualization conditions with one missing value. Each stimuli in the tutorial explained that some data was missing and that we had "guessed" at the values and described how we visualized "guessed" values. Participants were not informed of specific imputation methods or subjective tasks. Participants correctly identified the half-hour with the highest average or trend for each condition before beginning the formal study.

The formal study consisted of 87 trials presented serially (84 from our factorial design and 3 engagement checks). To mitigate effects from changing the visualization paradigm, we blocked stimuli by visualization method and randomized the order of blocks. Within each block, participants saw all twelve combinations of missing data (0%, 10%, 20%, and 30%) and imputation method (zero-filling, linear interpolation, and marginal mean) presented in random order. Each stimuli visualized a random dataset, with each dataset occurring at most once per participant. For averaging, engagement checks had 0% missing data where the average between halves of the dataset differed by 20.0. For trends, engagement checks differed in slopes by 1.0. These engagement checks were added to blocks 2, 4, and 6.

After completing the formal study, participants completed a demographic questionnaire, which included an opportunity for open-ended comments, and were compensated $1.00 for their participation.

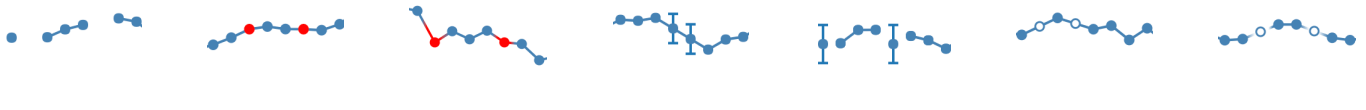
4.3 Participants

We collected data from 303 U.S. participants on Amazon's Mechanical Turk ($\mu_{age} = 36.3, \sigma_{age} = 12.7$, 150 female, 153 male). All participants reported normal or corrected-to-normal vision. To ensure honest participation and task understanding, we excluded any participants who answered two or more engagement checks incorrectly. Individual demographics and exclusions are reported in each Results section.

4.4 Measures & Analysis

We used three primary measures to analyze participant responses: perceived confidence in their answer (Question 2), credibility (Question 3), and a two-item scale describing perceived quality (Questions 4-5). We constructed the two-item scale by identifying correlation between the four quality questions at $\alpha = .7$. We combined correlated dimensions to construct our data quality scale per best practices [18, 22, 48]. We used this scale in place of the component questions to increase measure validity and use only descriptive analyses with single-item scales to mitigate effects of participant interpretation on our results. Accuracy both with and without imputed values formed a secondary metric to detect performance biases.

Unless otherwise specified, our main analysis used a repeated measures analysis of covariance (ANCOVA) to test for main and interaction effects with question order and noise treated as random effects and the actual (difference in when data points are removed) and imputed difference between means as covariates to mitigate effects of task difficulty. In both experiments, our response data was normally distributed. To control for Type I errors in planned comparisons between independently distributed settings of visualization method and imputation, we used Tukey's Honest Significant Difference test with $\alpha = .05$ for post-hoc analyses. We elected not to use response time as a measure. While



(a) Data Absent (b) Color Points (c) Color Points & Line Gradients (d) Connected Error Bars (e) Disconnected Error Bars (f) Unfilled Points (g) Unfilled Points & Line Gradients

Fig. 4: We tested seven different methods for visualizing missing values in line graphs manipulating both point and line appearance: two highlighting missing values, two downplaying missing values, two annotating missing values, and one removing missing values. .

understanding the effects of missing data on analysis speed is an interesting question, the inclusion of our subjective measures and use of crowdsourcing make it less reliable for our experiment.

5 Experiment One: Line Graphs

Prior studies in missing data visualization focused on line graphs, one of the most common and ubiquitous methods for visualizing data. We tested three factors we hypothesized that may effect missing data interpretation in line graphs: visualization type, percentage of missing data, and imputation method. We tested seven different visualization approaches that manipulated some combination of the point marks and lines themselves. We anticipated that since the connection between values in a line chart is salient, manipulating the lines may have different effects from manipulating the points alone.

Figure 4 shows the seven tested visualization designs: (a) Data Absent [21], (b) Color Points [14], (c) Color Points with Line Gradients (where the imputed value and its connections are both colored red) [14], (d) Connected Error Bars, with error corresponding to the standard deviation of the present values [4], (e) Disconnected Error Bars (where the line does not pass through imputed points) [4], (f) Unfilled Points [21], and (g) Unfilled Points with Line Gradients (where the line passing through imputed data is alpha-blended) [52]. Drawing on our four target visualization categories, the color conditions highlighted imputed values, unfilled points downplayed imputed values, error bars annotate data with additional statistical information, and data absent falls into information removal (Fig. 4). We conducted a 7 (visualization type) $\times$ 3 (imputation method) $\times$ 4 (percentage of missing data) within participants study to evaluate the effects of visualization and imputation on analyzing incomplete datasets using the procedure outlined in §4.

5.1 Line Graph Results

We found a strong correlation between perceived completeness and reliability (Cronbach's $\alpha > 0.70$) for both average and trend results. We constructed a two-item scale using these factors to describe data quality. We analyzed this scale as well as task performance using a three-factor full factorial rmANCOVA with question order and difficulty as random covariates. Figure 5 and Tables 1 & 2 summarize our results for perceived data quality and accuracy.

5.1.1 Averaging Results

Participants: We collected data from 66 U.S. participants on Amazon's Mechanical Turk ($\mu_{age} = 37.2, \sigma_{age} = 13.22$, 36 female, 30 male). Two participants were excluded for failing two or more of the large difference engagement check stimuli, resulting in 5,568 total trials. Table 1 summarizes our primary results.

Subjective Results

Data Quality: We found main effects of missing amount, visualization, and imputation on perceptions of data quality (Fig. 5). As the amount of missing data increased, perceived data quality decreased. Connected error bars and color points with line gradients led to higher perceived data quality, whereas the data absent conditions and disconnected error bars led to lower perceived data quality. Linear interpolation had higher perceived quality than marginal means, and both methods had higher perceived quality than zero-filling. We found three significant interactions effects on data quality: 1) visualization and missing amount, 2) visualization and imputation, and 3) imputation and missing amount. Color points with linear interpolation led to higher perceived quality, while disconnected error bars with zero-filling led to the lower. Connected error bars and zero-filling had significantly higher perceived quality than other zero-filling conditions.

Data Credibility & Analyst Confidence: Perceived data credibility and confidence mirrored our data quality results. Connected error bars and color points led to higher perceived credibility ($\mu_{connected} = 4.59 \pm .12$, $\mu_{colorpoint} = 4.61 \pm .11$) and confidence ($\mu_{connected} = 4.63 \pm .12$, $\mu_{colorpoint} = 4.66 \pm .11$), whereas the data absent condition led to lower credibility ($\mu_{absent} = 4.08 \pm .12$) and confidence ($\mu_{absent} = 4.29 \pm .12$). We found the same preference ranking for different imputation methods across both confidence and credibility: linear interpolation being highest ($\mu_{cred} = 4.88 \pm .07$, $\mu_{conf} = 4.67 \pm .07$), followed by marginal means ($\mu_{cred} = 4.76 \pm .07$, $\mu_{conf} = 4.48 \pm .08$), and zero-filling ($\mu_{cred} = 4.45 \pm .08$, $\mu_{conf} = 4.35 \pm .08$).

Objective Results

We computed accuracy both considering (with, Table 1) imputed values and excluding (without, statistics inline) imputed values to include both potential analysis strategies. Overall, participants correctly identified the half-hour with the higher mean in 78.8% of trials with and 78.3% without considering imputed values. We found main effects of missing amount ($F_{without}(1,62) = 23.81, p < .0001$) and imputation method ($F_{without}(2,61) = 8.76, p < .0002$) on accuracy. Linear interpolation and marginal means ($\mu_{linear} = 80.65\% \pm 1.15$ to $\mu_{means} = 77.10\% \pm 1.15$) both led to higher accuracy than zero-filling method ($\mu_{zero} = 72.87\% \pm 1.15$) in both cases. We found significant interaction effects of imputation and missing amount for both metrics ($F_{without}(2,61) = 6.14, p < .002$).

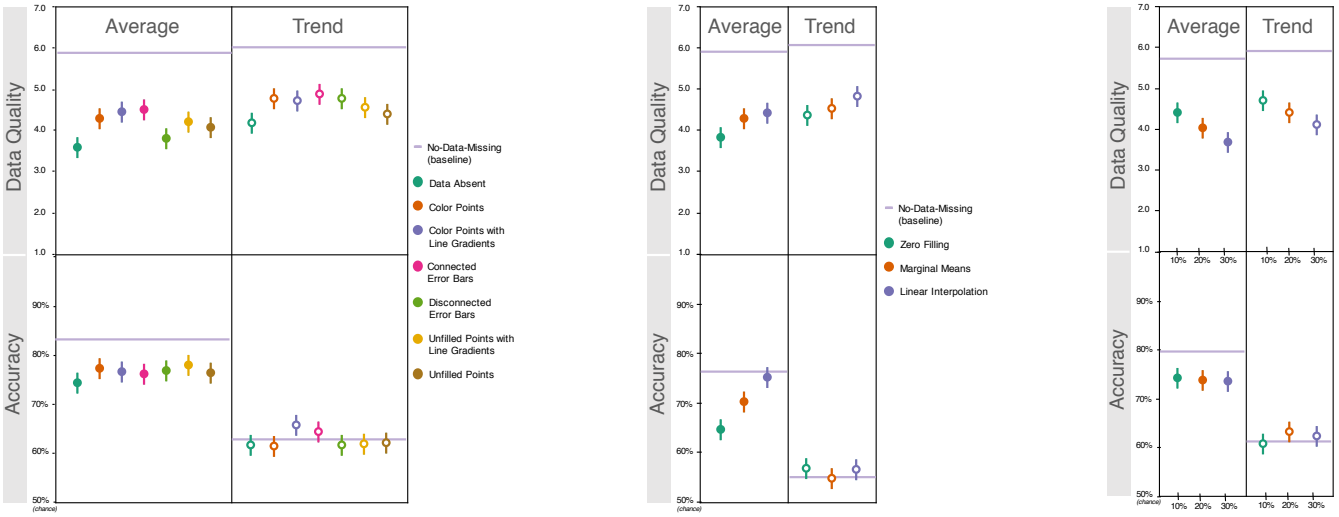
5.1.2 Trend Detection Results

Participants: We collected data from 80 U.S. participants on Amazon's Mechanical Turk ($\mu_{age} = 36.05, \sigma_{age} = 11.75$, 38 female, 42 male). No participants were excluded as all participants correctly answered two or more of the engagement checks, resulting in 6,960 trials. Table 2 summarizes our primary results.

Subjective Results

Data Quality: We found significant effects of the percentage of missing data, visualization, and imputation method on perceived data quality (Fig. 5). Connected error bars and color points led to higher perceived data quality while disconnected error bars and data absent led to the lower. Both linear interpolation and marginal means had higher perceived quality than zero-filling. We found a significant interaction between imputation method and missing amount, with linear interpolation being more robust as the amount increased.

Data Credibility & Analyst Confidence: Our confidence and credibility effects again mirrored effects of data quality. Connected error bars ($\mu_{cred} = 4.90 \pm .12$, $\mu_{conf} = 5.42 \pm .12$) and color points ($\mu_{cred} = 4.86 \pm .11$, $\mu_{conf} = 5.48 \pm .11$) led to higher perceived data credibility and confidence in trend detection, whereas disconnected error bars ($\mu_{cred} = 4.27 \pm .12$, $\mu_{conf} = 5.07 \pm .12$) and information removal ($\mu_{cred} = 4.51 \pm .10$, $\mu_{conf} = 5.27 \pm .11$) led to lower overall perceptions of confidence and credibility. Of note, connected error bars and zero-filling also had significantly higher perceived credibility than all other zero-filling conditions ($\mu_{error,zero} = 4.83 \pm .20$). We again found a consistent preference ranking across these factors for impu-



(a) Visualization Methods (b) Imputation Methods (c) Amount Missing

Fig. 5: Our results show that (a) visualization type, (b) imputation method, and (c) amount of missing data shift perceived data quality and can bias data interpretation for average and trend detection in line graphs. Error bars are bootstrapped 95% confidence intervals.

Table 1: Summary of significant results for line graph averaging (grey indicates non-significant results)

Factors	Data Quality	Accuracy (%)
Amt. Missing	$F(1,62) = 1519.38$, $p < .0001$	$F(1,62) = .32$, $p < .5694$
Vis.	$F(6,57) = 22.92$, $p < .0001$	$F(6,57) = .36$, $p < .9024$
(Connected)	$(\mu = 4.53 \pm .12)$	$(\mu = 76.50 \pm 1.77)$
(Data Absent)	$(\mu = 3.60 \pm .11)$	$(\mu = 74.00 \pm 1.78)$
(Color gradients)	$(\mu = 4.3 \pm .12)$	$(\mu = 77.00 \pm 1.78)$
(Disconnected)	$(\mu = 3.85 \pm .12)$	$(\mu = 76.03 \pm 1.78)$
Imp.	$F(2,61) = 61.77$, $p < .0001$	$F(2,61) = 20.78$, $p < .0001$
(Linear)	$(\mu = 4.43 \pm .07)$	$(\mu = 81.20 \pm 1.16)$
(Marginal)	$(\mu = 4.24 \pm .07)$	$(\mu = 76.54 \pm 1.16)$
(Zero-Filling)	$(\mu = 3.77 \pm .08)$	$(\mu = 70.59 \pm 1.17)$
Vis*Amt. Missing	$F(6,57) = 6.86$, $p < .0001$	$F(6,57) = 1.20$, $p < .3024$
Vis*Imp.	$F(12,51) = 3.047$, $p < .0002$	$F(12,51) = .48$, $p < .9267$
(Color Points*Linear)	$(\mu = 4.75 \pm .19)$	$(\mu = 83.50 \pm 3.08)$
(Disconnected*Zero)	$(\mu = 3.50 \pm .23)$	$(\mu = 69.78 \pm 3.09)$
(Connected*Zero)	$(\mu = 4.09 \pm .20)$	$(\mu = 69.35 \pm 3.09)$
Imp.*Amt. Missing	$F(12,51) = 17.04$, $p < .0001$	$F(12,51) = .77$, $p < .6727$

Table 2: Summary of significant results for line graph trends

Factors	Data Quality	Accuracy (%)
Amt. Missing	$F(1,78) = 1185.901$, $p < .0001$	$F(1,78) = .06$, $p < .8040$
Vis	$F(6,73) = 17.56$, $p < .0001$	$F(6,73) = .76$, $p < .5975$
(Connected)	$(\mu = 4.85 \pm .12)$	$(\mu = 64.88 \pm 1.65)$
(Data Absent)	$(\mu = 4.19 \pm .11)$	$(\mu = 62.62 \pm 1.65)$
(Color Gradients)	$(\mu = 4.76 \pm .17)$	$(\mu = 61.57 \pm 1.66)$
(Disconnected)	$(\mu = 4.06 \pm .13)$	$(\mu = 63.56 \pm 1.66)$
Imp.	$F(2,77) = 24.63$, $p < .0001$	$F(2,77) = .52$, $p < .5891$
(Linear)	$(\mu = 4.72 \pm .08)$	$(\mu = 62.29 \pm 1.09)$
(Marginal)	$(\mu = 4.46 \pm .09)$	$(\mu = 63.50 \pm 1.08)$
(Zero-Filling)	$(\mu = 4.30 \pm .08)$	$(\mu = 63.77 \pm 1.08)$
Imp.*Amt. Missing	$F(2,77) = 5.81$, $p < .003$	$F(2,77) = 1.55$, $p < .2114$

tation method: linear interpolation being highest ($\mu_{cred} = 4.90 \pm .07$, $\mu_{conf} = 5.55 \pm .06$), followed by marginal means ($\mu_{cred} = 4.62 \pm .07$, $\mu_{conf} = 5.40 \pm .07$), and zero-filling ($\mu_{cred} = 4.49 \pm .06$, $\mu_{conf} = 5.42 \pm .07$).

Objective Results

We again computed accuracy both considering (with, Table 2) imputed values and excluding (without, statistics inline) imputed values to include both potential analysis strategies. Overall, participants correctly identified the half-hour with the larger overall rate of change or slope with imputed values in 62.8% of trials and 63.2% without imputed values. We did not find any significant effects of our independent variables on trend detection accuracy.

5.2 Line Graphs–Synthesis of Results

Our results provide preliminary support for all of our hypotheses:

H1: As the percentage of missing data increased, perceived quality, credibility, accuracy, and confidence in analysis decreased.

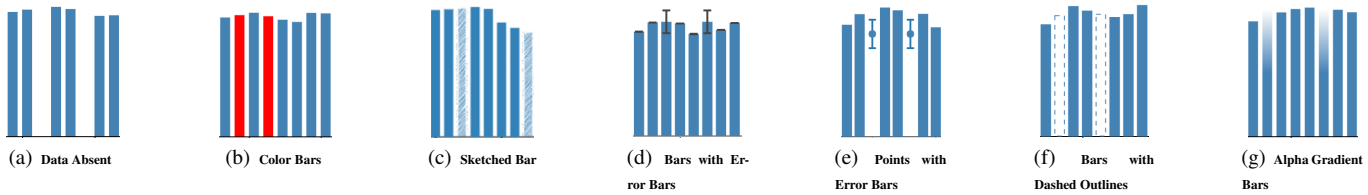
H2: *(partial)* We found that encodings highlighting missing values (e.g., our color points and color points and lines conditions) led to significantly higher perceived quality than visualizations that removed data. We failed to find evidence supporting the same comparison with downplaying techniques.

H3: Linear interpolation led to the highest perceived data quality

H4: The data absent condition led to significantly lower perceived data quality, credibility, and confidence than all other visualization conditions.

We found limited evidence of accuracy bias from imputation methods, consistent with Eaton et al. [21].

Our results showed that the participants regarded color points with line gradients as of higher quality overall while removing missing datapoints from the graph caused analysts to see the data as having lower quality and led to lower reported confidence in their task performance. While most methods within a given design category performed comparably, we also saw an interesting contrast between our two annotation conditions (connected versus disconnected error bars). Participants perceived connected error bars as having higher data quality, being more credible, and leading to higher confidence in interpretation. On the other hand, when the error bars were disconnected they reported significantly lower scores on these subjective measures. This conclusion offers an interesting consideration when juxtaposed with results from Andreasson and Riveiro [4] who found that fuzziness, another correlate for uncertainty, led to the lowest overall preference for decision making tasks. This discrepancy also suggests that adding additional information is insufficient to increase perceived data quality: the underlying structure of the visualization and integration of imputed values into the visual structure of a dataset may also play a role. Additionally,



(a) Data Absent (b) Color Bars (c) Sketched Bar (d) Bars with Error Bars (e) Points with Error Bars (f) Bars with Dashed Outlines (g) Alpha Gradient Bars

Fig. 6: We tested seven different methods for visualizing missing values in bar charts: one highlighting missing values, three downplaying missing values, two annotating missing values, and one removing missing values.

Table 3: Summary of significant results for bar graph averaging

Factors	Data Quality	Accuracy (%)
Amt. Missing	$F(1,78) = 2927.7$, $p < .0001$	$F(1,78) = 2.6$, $p < .1067$
Vis	$F(6,73) = 24.15$, $p < .0001$	$F(6,73) = 3.91$, $p < .0007$
(Bar Error)	$(\mu = 4.82 \pm .11)$	$(\mu = 81.48 \pm 1.52)$
(Color Bars)	$(\mu = 4.71 \pm .08)$	$(\mu = 83.23 \pm 1.52)$
(Gradient Bars)	$(\mu = 4.71 \pm .08)$	$(\mu = 79.62 \pm 1.5)$
(Sketched Bars)	$(\mu = 4.41 \pm .07)$	$(\mu = 78.69 \pm 1.54)$
(Dashed Outline)	$(\mu = 4.32 \pm .08)$	$(\mu = 75.35 \pm 1.52)$
(Point Error)	$(\mu = 4.26 \pm .09)$	$(\mu = 75.70 \pm 1.52)$
(Data Absent)	$(\mu = 3.60 \pm .11)$	$(\mu = 76.37 \pm 1.52)$
Imp.	$F(2,77) = 52.76$, $p < .0001$	$F(2,77) = 9.14$, $p < .0001$
(Linear)	$(\mu = 4.67 \pm .07)$	$(\mu = 80.51 \pm 1)$
(Marginal)	$(\mu = 4.60 \pm .07)$	$(\mu = 80.25 \pm 1)$
(Zero-Filling)	$(\mu = 4.26 \pm .07)$	$(\mu = 75.14 \pm 1)$
Vis*Amt. Missing	$F(6,73) = 10.73$, $p < .0001$	$F(6,73) = .89$, $p < .4988$
Vis*Imp.	$F(12,67) = 7.97$, $p < .0001$	$F(12,67) = 1.21$, $p < .2636$
(Bar Error*Linear)	$(\mu = 5.14 \pm .19)$	$(\mu = 88.52 \pm 2.64)$
(Point Error*Marginal)	$(\mu = 4.19 \pm .23)$	$(\mu = 79.37 \pm 2.64)$
(Gardient Bars*Linear)	$(\mu = 5.01 \pm .21)$	$(\mu = 85.41 \pm 2.61)$
(Color Bars*Linear)	$(\mu = 4.99 \pm .22)$	$(\mu = 83.92 \pm 2.64)$
Imp.*Amt. Missing	$F(2,77) = 21.35$, $p < .0001$	$F(2,77) = 2.52$, $p < .0808$
Vis*Imp.*Amt. Missing	$F(12,67) = 3.11$, $p < .0001$	$F(12,67) = 1.75$, $p < .0514$

the robustness demonstrated by connected error bars as the amount of missing data increased suggests that connected error bars may preserve perceived quality even as actual quality decreases, which could bias decision making. As this assumption is grounded in descriptive statistics and a lack of an effect, further testing is needed to determine the validity of this observation.

6 Experiment Two: Bar Charts

Bar charts provide an interesting case for visualizing missing data as they use bar height to encode data rather than position and connection. This change in encoding may shift the data patterns people observe [57]. Many techniques for bar charts also visualize absent data, zero-filling, and $y = 0$ the same way, which may change biases and quality perceptions compared to line graphs.

Several of our designs mirrored those used in the line graph conditions; however, we extended our techniques to include sketchy rendering [55] and dashed outlines [13] to prioritize encodings correlated to downplaying imputed values. Figure 6 shows the seven tested visualization designs: (a) Data Absent [21], (b) Color Bars [14], (c) Sketched Bars [55], (d) Bars with Error Bars [4], (e) Points with Error Bars [4], (f) Unfilled Bars with Dashed Outlines [13], and (g) Alpha-blended Gradient Bars [52]. Error bars again approximated the standard deviation of the data used in the imputation, and gradients used this amount to define the blend radius. Color bars highlighted imputed values; error bars and points with error bars annotated imputed values; sketchiness, gradient bars, and dashed outlines downplayed imputed values; and not showing the values exemplified information removal.

6.1 Bar Graph Results

We again found a strong correlation between perceived completeness and reliability ($\alpha > 0.70$) for both averaging and trend detection. We constructed a two-item scale using these factors to describe data quality. We analyzed this scale as well as task performance using a three-factor (visualization technique, imputation method, and amount of missing data) full factorial rmANCOVA with question order and difficulty as random covariates. Figure 7 and Tables 3 & 4 summarize our results for perceived data quality and accuracy.

6.1.1 Averaging Results

Participants: We collected data from 80 U.S. participants on Mechanical Turk ($\mu_{age} = 37.1, \sigma_{age} = 11.4$, 38 female, 42 male). All participants answered at least two of the engagement checks correctly, resulting in 6,960 trials. Figure 7 and Table 3 summarize our results.

Subjective Results

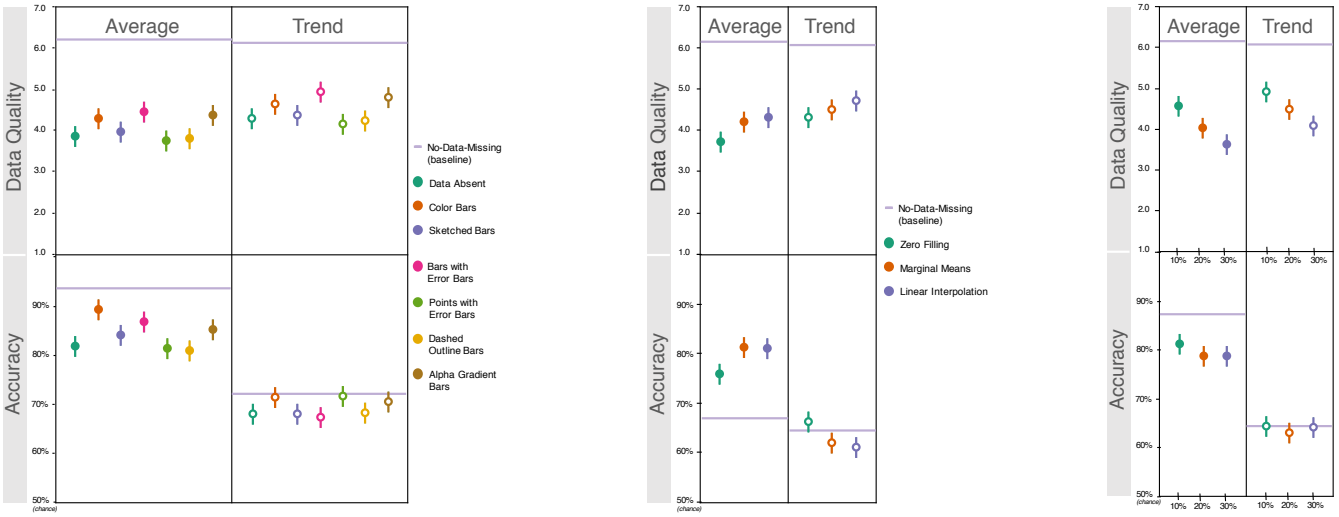
Data Quality: We found significant main effects of the percentage of missing data, visualization type, and imputation method on perceived data quality (Fig. 7). Increasing amount of missing data led to lower perceived data quality. Bars with error bars, color bars, and gradient bars led to higher perceived data quality, whereas points with error bars, and downplaying techniques—sketched bars and dashed outlines—and information removal led to lower perceived data quality. Both linear interpolation and marginal means had higher perceived quality than zero-filling. We also found four significant interactions: 1) visualization type and missing data, 2) visualization type and imputation, 3) amount of missing data and imputation method, and 4) visualization type, amount of missing data, and imputation method. Bars with error bars, gradient bars, and color bars using linear interpolation led to higher perceived data quality, while points with error bars and zero-filling led to lower data quality.

Data Credibility & Analyst Confidence: Self-reported perceptions of data credibility and analyst confidence generally mirrored data quality measures. Color bars led to the highest overall perceived confidence ($\mu_{color} = 4.75 \pm .10$). Color bars, bars with error bars, and gradient bars generated higher perceived credibility ($\mu_{color} = 4.85 \pm .07$, $\mu_{barerror} = 4.90 \pm .11$, and $\mu_{gradient} = 4.86 \pm .08$), while points with error bars led to consistently lower credibility ($\mu_{pointerror} = 4.52 \pm .09$) and confidence ($\mu_{pointerror} = 4.64 \pm .10$). The downplay techniques (sketching, outlines) and information removal also led to low reported confidence ($\mu_{sketch} = 4.57 \pm .07$, $\mu_{outline} = 4.56 \pm .08$, $\mu_{absent} = 4.60 \pm .07$).

Both linear interpolation and marginal means led to higher perceived credibility ($\mu_{local} = 4.81 \pm .07$, $\mu_{means} = 4.75 \pm .06$, and $\mu_{zero} = 4.52 \pm .07$) and confidence ($\mu_{local} = 4.71 \pm .07$, $\mu_{means} = 4.64 \pm .06$, and $\mu_{zero} = 4.47 \pm .07$). Unlike with lines, we found no evidence of significant differences between linear interpolation and marginal means for bars.

Objective Results

We again computed accuracy both considering (with, Table 3) imputed values and excluding (without, statistics inline) imputed values to include both potential analysis strategies. Overall, participants correctly identified the half-hour with the higher average in 80.76% of trials without imputed values and 80.73% with them. We found a main effects of visualization ($F_{without}(6,73) = 4.84, p < .0001$) and imputation ($F_{without}(2,77) = 3.38, p < .0338$) on accuracy in both cases. Bars with error bars and color bars led to higher accuracy, whereas the dashed outline bars led to lower accuracy ($\mu_{barerror} = 83.98\% \pm .01$



(a) Visualization Methods (b) Imputation Methods (c) Amount Missing

Fig. 7: As in line charts, our results show that (a) visualization type, (b) imputation method, and (c) amount of missing data shift perceived data quality and can even bias data interpretation for average and trend detection in bar charts.

Table 4: Summary of significant results for bar graph trends

Factors	Data Quality	Accuracy (%)
Amt. Missing	$F(1,72) = 1803.289$, $p < .0001$	$F(1,72) = 2.38$, $p < .1224$
Vis	$F(6,67) = 22.80$, $p < .0001$	$F(6,67) = 0.32$, $p < .9268$
(Bar Error)	$(\mu = 4.81 \pm .10)$	$(\mu = 61.08 \pm 1.60)$
(Dashed Outline)	$(\mu = 4.10 \pm .08)$	$(\mu = 60.97 \pm 1.59)$
(Data Absent)	$(\mu = 4.18 \pm .11)$	$(\mu = 60.28 \pm 1.61)$
(Points Error)	$(\mu = 4.03 \pm .07)$	$(\mu = 62.39 \pm 1.61)$
Imp.	$F(2,71) = 127.61$, $p < .0001$	$F(2,71) = 5.66$, $p < .0035$
(Linear)	$(\mu = 4.72 \pm .08)$	$(\mu = 59.57 \pm 1.06)$
(Marginal)	$(\mu = 4.39 \pm .04)$	$(\mu = 60.57 \pm 1.06)$
(Zero-Filling)	$(\mu = 4.14 \pm .04)$	$(\mu = 64.33 \pm 1.05)$
Vis * Amt. Missing	$F(6,67) = 7.64$, $p < .0001$	$F(6,67) = 2.01$, $p < .0607$
Vis * Imp.	$F(12,61) = 5.73$, $p < .0001$	$F(12,61) = 0.79$, $p < .6546$
Imp. * Amt. Missing	$F(2,71) = 9.34$, $p < .0001$	$F(2,71) = 1.42$, $p < .2401$

and $\mu_{color} = 83.4\% \pm .02$ vs. $\mu_{dashoutline} = 77.95\% \pm .02$). Linear interpolation and marginal means ($\mu_{linear} = 80.40\% \pm .991$ to $\mu_{means} = 80.19.\% \pm .991$) both led to higher accuracy than zero-filling method ($\mu_{zero} = 77.13\% \pm .992$).

For strategies inclusive of imputed values, we found additional main effects of imputation: linear interpolation and marginal means both led to higher accuracy than zero-filling. We found two interaction effects on accuracy: 1) imputation and missing amount and 2) visualization, missing amount, and imputation method. Bars with error bars using linear interpolation led to higher accuracy than linearly interpolated points with error bars and were more robust to changes in the amount of missing data.

6.1.2 Trend Detection Results

Participants: We collected data from 77 U.S. participants on Mechanical Turk ($\mu_{age} = 34.9, \sigma_{age} = 11.9$, 38 female, 39 male). We excluded 3 participants who failed to answer at least two of the engagement check stimuli correctly, resulting in 6,438 trials. Table 4 summarizes our primary results.

Subjective Results

Data Quality: We found main effects of the amount of missing data, visualization type, and imputation method on perceived data quality (Fig. 7). Bars with error bars led to higher perceived data quality while data absent, dashed outline bars, and points with error bars were seen as lower quality. Linear interpolation resulted in the highest perceived data quality, followed by marginal means, and data absent. We found three main interaction effects: 1) visualization and amount of missing data, 2) visualization method and imputation, and 3) amount of missing data and imputation. Bars with error bars and gradient bars with linear interpolation led to higher perceived data quality than all imputation methods using points with error bars and than dashed outlines with both zero-filling and marginal means.

Data Credibility & Analyst Confidence: We found limited effects of our independent variables on credibility and confidence that did not necessarily align with our quality scale. Specifically, gradient bars led to higher perceived credibility and confidence ($\mu_{cred} = 5.18 \pm .05$, $\mu_{conf} = 5.45 \pm .11$) than dashed outline bars ($\mu_{cred} = 4.85 \pm .10$, $\mu_{conf} = 5.17 \pm .11$) and points with error bars ($\mu_{cred} = 4.80 \pm .11$, $\mu_{conf} = 5.13 \pm .12$). Linear interpolation ($\mu_{cred} = 4.84 \pm .11$, $\mu_{conf} = 5.26 \pm .04$) led to higher perceived confidence than marginal means ($\mu_{cred} = 4.65 \pm .12$, $\mu_{conf} = 5.02 \pm .04$) and zero-filling ($\mu_{cred} = 4.55 \pm .12$, $\mu_{conf} = 5.10 \pm .05$).

Objective Results

Overall, participants correctly identified the half-hour with the larger overall rate of change in 60.6% of trials when discounting imputed values and 61.4% of trials when considering imputed values. We found no significant main effects for strategies discounting imputed values; however, we did find a surprising countereffect to averaging without imputed values in that zero-filling led to higher accuracy than linear interpolation and marginal means inclusive of imputed values.

6.2 Bar Graphs–Synthesis of Results

Our results provide support for each of our four hypotheses:

H1: (partial) As the amount of missing data increased, confidence in result and perceived data quality decreased. We found no corresponding evidence for accuracy.

H2: *(partial)* Our data absent condition and two downplaying techniques (sketching and dashed outlines) led to consistently low perceived data quality, whereas color bars led to high quality perceptions. However, gradient bars, which also downplay imputed values, led to consistently high perceived data quality.

H3: *(partial)* Linear interpolation outperformed zero-filling and led to higher quality perceptions, confidence, and credibility for the tested visualization types. However, linear interpolation only led to higher subjective impressions for our trend detection task.

H4: *(partial)* The data absent condition again led to low overall perceived quality, confidence, and credibility; however, other techniques performed comparably poorly on this metric.

Across both tasks, we found that bars with error bars, gradient bars, and color bars led to consistently high perceived confidence and overall accuracy. However, our other annotation condition—points with error bars—and downplay conditions—sketched bars and dashed outline bars—led to lower perceived confidence, credibility, data quality, and task accuracy. Combined with our results from Experiment One and findings from Andreasson & Riveiro [4], these findings suggest that visual encodings that break the continuous visual structure of a graph reduce perceptions of data quality and may actually inhibit analysis. We anticipate that these effects may be more critical to consider for bar charts as we found significant discrepancies between accuracy effects inclusive and exclusive of imputed values.

We hypothesize the observed subjective and objective bias for encodings that break continuity may be due to attentional selection. The process used to visually average information may be impacted by the point condition's use of height and position encodings. Our lowest performing conditions—data absent, points with error bars, dashed outlines, sketching, and zero filling—reduce or remove the weight of imputed bars. Based on known mechanisms of visual attention (see Gleicher et al. [28] for a discussion), these reductions may cause participants to group consecutive sets of imputed and real data values prior to estimating aggregate statistics, complicating visual aggregation tasks. Gradient bars would not exhibit these effects as the majority of the bar appears consistent with the rest of the bars in the sequence. Future testing is needed to verify this hypothesis.

7 Discussion

We measured the effect of missing data on interpretation accuracy and perceived data quality in time series data across 14 visualization methods and three imputation types. Our results show:

- Perceived data quality and confidence generally degrade as the amount of missing data increases.
- Data visualized by highlighting missing values tends to be seen as higher quality than downplay or information removal.
- Information removal can significantly degrade perceptions of data quality, and confidence. These methods even lead to incorrect responses if missing values break the visual continuity of a visualization.
- Linear interpolation leads to higher perceptions of quality and confidence in analysis.

While avoiding bias is a critical element of effective visualization, we find that the ways systems impute and visualize missing data can also manipulate perceived data quality and confidence in results. Whether ideal perceptions of quality are high or low is likely dependent on parameters of the data, problem, and domain.

We found preliminary evidence that high confidence, credibility, and perceived quality in interpreting incomplete datasets depend on multiple factors. Our studies show that visualizing imputed datapoints using highlighting and annotation while preserving the continuity of available data lead to the highest perceived data quality and confidence in result. We see this in color bars, gradient bars, and bars with error bars in bar graphs and with connected error bars and color points with line gradient in line graphs. This conclusions runs somewhat contrary to prior work on preference in decision support [4]; however, it is inline with research on visual selection [28] and uncertainty and trust (see Sacha et al [47] for a discussion). Specifically, visual explanations of errors, such as those indicated by error bars, may increase trust in a system [20]. This trust may manifest in increased confidence and perceived data quality. Examining further design components associated with uncertainty visualization could offer further insight into this phenomena (e.g., alternative designs [37] and cognitive biases [16]). However, we found that error bars that do not preserve continuity—disconnected error bars in line graphs and point with errors bars in bar charts—lead to low perceived confidence, credibility, and data quality, which indicate critical factors beyond the integration of uncertainty.

Linear interpolation consistently produced the highest confidence and perceived data quality, while zero-filling led to the lowest. The increased quality perceptions associated with linear interpolation indicate that imputation methods should favor methods drawing from the data distribution in scenarios requiring high analyst confidence and perceived quality. Sophisticated imputation methods may be worth added complexity to avoid response bias in these scenarios. We offer first steps towards a systematic understanding of the role of imputation in visualization; however, our studies focused on a simple domain and relatively smooth signals. Further evaluation is needed to more deeply understand the effects of imputation in visualization.

While our results enumerate how design choices for visualizing incomplete datasets might modulate perceived quality, a number of factors may inform the desirable level of perceived quality, including:

Decision Risk: Accuracy is paramount in high risk situations and missing values may jeopardize that accuracy. Visualization systems can encourage caution in interpreting flawed datasets by using representations that avoid bias and appropriately decrease perceived quality.

Data Fusion: Combining data from multiple sources may mean that the overall anticipated data quality varies between those sources. In these scenarios, systems can leverage domain knowledge to guide analysts to visualizations that modulate confidence and apparent quality appropriately for each source.

Confidence in Imputation: Individual imputation methods may differ in how faithfully they represent the original data. In scenarios like cold-deck imputation, analysts may know how well the imputed data mirrors the original data. Visualizations can leverage this knowledge to choose methods that adapt perceived quality proportionally to the imputation quality.

7.1 Limitations & Future Work

We made several simplifying assumptions in our experiment. Our narrative scenario used familiar but simple and low-risk tasks (i.e., the cost of getting the wrong answer is minimal). While these choices allowed us strong control over our tested conditions to encourage general understanding, future testing should extend our work to real-world datasets and scenarios to better understand the impact of these choices and how different analytic workflows and data characteristics might change these perceptions.

Further, we tested a small set of possible imputation and visualization methods, drawing inspiration from visualization tools that actively manage missing data. However, we found few tools explicitly discuss missing data management. Future work should extend to a broader set of visualization and imputation methods, such as multiple imputation and machine learning-based approaches to understand their broader utility for data in different domains. Future studies should additionally test more subtle amounts of missing data and consider more formal modeling of salience, uncertainty, and other perceptions with perceived quality and accuracy.

8 Conclusion

We used time series data with missing values to measure how missing data influences factors of perceived data quality and found that the design choices and interpolation methods used to represent data significantly influence analysts' perceptions of data. Highlighting imputed values and using linear interpolation led to higher perceived confidence, credibility, and data quality. Downplaying visual encodings, zero-filling, and electing not to draw data can lead to lower subjective perceived measurements.

Our results enumerate design trade-offs for designers to consider when crafting behaviors for handling missing data in visualization in order to tailor subjective and objective responses to the demands and domains of stakeholders.

Acknowledgments

The authors wish to thank reviewers and other members of VisuaLab at University of Colorado Boulder. This research was funded by NSF CRII: CHS # 1657599.

REFERENCES

[1] M. Adnan, M. Just, and L. Baillie. Investigating time series visualisations to improve the user experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5444–5455. ACM, 2016.

[2] S. Ahuja, M. Roth, R. Gangadharaiah, P. Schwarz, and R. Bastidas. Using machine learning to accelerate data wrangling. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pp. 343–349. IEEE, 2016.

[3] D. Albers, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 551–560. ACM, 2014.

[4] R. Andreasson and M. Riveiro. Effects of visualizing missing data: an empirical evaluation. In *Information Visualisation (IV), 2014 18th International Conference on*, pp. 132–138. IEEE, 2014.

[5] C. Arbesser, F. Spechtenhauser, T. Mühlbacher, and H. Piringer. Visplause: Visual data quality assessment of many time series using plausibility checks. *IEEE transactions on visualization and computer graphics*, 23(1):641–650, 2017.

[6] A. Aris, B. Shneiderman, C. Plaisant, G. Shmueli, and W. Jank. Representing unevenly-spaced time series data for visualization and interactive exploration. *Lecture notes in computer science*, 3585:835, 2005.

[7] Y. M. Babad and J. A. Hoffer. Even no data has a value. *Communications of the ACM*, 27(8):748–756, 1984.

[8] J. Bernard, T. Ruppert, O. Goroll, T. May, and J. Kohlhammer. Visual-interactive preprocessing of time series data. In *Proceedings of SIGRAD 2012; Interactive Visual Analysis of Data; November 29-30; 2012; Växjö; Sweden*, number 081, pp. 39–48. Linköping University Electronic Press, 2012.

[9] M. Bögl, P. Filzmoser, T. Gschwandtner, S. Miksch, W. Aigner, A. Rind, and T. Lammarsch. Visually and statistically guided imputation of missing values in univariate seasonal time series. In *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*, pp. 189–190. IEEE, 2015.

[10] C. Bors, M. Bögl, T. Gschwandtner, and S. Miksch. Visual support for rastering of unequally spaced time series. In *Proceedings of the 10th International Symposium on Visual Information Communication and Interaction*, pp. 53–57. ACM, 2017.

[11] C. Bors, T. Gschwandtner, and S. Miksch. Qualityflow: Provenance generation from data quality. In *Poster Proceedings of the EuroGraphics Conference on Visualization*, 2014.

[12] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011.

[13] N. Boukhelifa, A. Bezerianos, T. Isenberg, and J.-D. Fekete. Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2769–2778, 2012.

[14] P. Buono, A. Aris, C. Plaisant, A. Khella, B. Shneiderman, H. Hochheiser, and B. Schneiderman. Interactive pattern search in time series. In *Proc. SPIE*, vol. 5669, pp. 175–186, 2005.

[15] M. Correll, D. Albers, S. Franconeri, and M. Gleicher. Comparing averages in time series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1095–1104. ACM, 2012.

[16] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics*, 20(12):2142–2151, 2014.

[17] M. Correll and J. Heer. Regression by eye: Estimating trends in bivariate visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1387–1396. ACM, 2017.

[18] R. V. Dawis. Scale construction. *Journal of Counseling Psychology*, 34(4):481, 1987.

[19] S. Djurcilov and A. Pang. Visualizing gridded datasets with large number of missing values (case study). In *Proceedings of the conference on Visualization'99: celebrating ten years*, pp. 405–408. IEEE Computer Society Press, 1999.

[20] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6):697–718, 2003.

[21] C. Eaton, C. Plaisant, and T. Drizd. Visualizing missing data: Graph interpretation user study. *Human-Computer Interaction-INTERACT 2005*, pp. 861–872, 2005.

[22] A. L. Edwards. *Techniques of attitude scale construction*. Ardent Media, 1983.

[23] S. J. Fernstad and R. C. Glen. Visual analysis of missing dataâĂŤto see what isn't there. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pp. 249–250. IEEE, 2014.

[24] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3237–3246. ACM, 2013.

[25] T. Furche, G. Gottlob, L. Libkin, G. Orsi, and N. W. Paton. Data wrangling for big data: Challenges and opportunities. In *EDBT*, pp. 473–478, 2016.

[26] J. Gao. Adaptive interpolation algorithms for temporal-oriented datasets. In *Temporal Representation and Reasoning, 2006. TIME 2006. Thirteenth International Symposium on*, pp. 145–151. IEEE, 2006.

[27] W. Githungo, S. Otengi, J. Wakhungu, and E. Masibayi. Infilling monthly rain gauge data gaps with satellite estimates for asal of kenya. *Hydrology*, 3(4):40, 2016.

[28] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE transactions on visualization and computer graphics*, 19(12):2316–2325, 2013.

[29] T. Gschwandtnei, M. Bögl, P. Federico, and S. Miksch. Visual encodings of temporal uncertainty: A comparative user study. *IEEE transactions on visualization and computer graphics*, 22(1):539–548, 2016.

[30] T. Gschwandtner, J. Gärtner, W. Aigner, and S. Miksch. A taxonomy of dirty time-oriented data. *Multidisciplinary Research and Practice for Information Systems*, pp. 58–72, 2012.

[31] K. Gülensoy, C. Gawrilow, and T. von Landesberger. Visual exploration of dirty activity sensor and emotional state data from psychological experiments. In *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, p. 19. ACM, 2014.

[32] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1303–1312. ACM, 2009.

[33] P. T. Inc. Collaborative data science, 2015.

[34] Y. Jansen and K. Hornbæk. A psychophysical investigation of size as a physical variable. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):479–488, 2016.

[35] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[36] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3363–3372. ACM, 2011.

[37] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5092–5103. ACM, 2016.

[38] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee. A taxonomy of dirty data. *Data mining and knowledge discovery*, 7(1):81–99, 2003.

[39] M. J. Lajeunesse, J. Koricheva, J. Gurevitch, and K. Mengersen. Recovering missing or partial data from studies: a survey of conversions and imputations for meta-analysis. *Handbook of Meta-analysis in Ecology and Evolution*, pp. 195–206, 2013.

[40] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[41] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505, 2012.

[42] H. Müller and J.-C. Freytag. *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik, 2005.

[43] K. Perlin. Improving noise. In *ACM Transactions on Graphics (TOG)*, vol. 21, pp. 681–682. ACM, 2002.

[44] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[45] L. Pouchard. Revisiting the data lifecycle with big data curation. *International Journal of Digital Curation*, 10(2):176–192, 2016.

[46] S. Rässler. Data fusion: identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, 33(1&2):153–171, 2016.

[47] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, 22(1):240–249, 2016.

[48] P. E. Spector. *Summated rating scale construction: An introduction*.

Number 82. Sage, 1992.

[49] D. F. Swayne and A. Buja. Missing data in interactive high-dimensional data visualization. *Computational Statistics*, 13(1):15–26, 1998.

[50] M. Templ, A. Alfons, and P. Filzmoser. Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification*, 6(1):29–47, 2012.

[51] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser. Representative factor generation for the interactive visual analysis of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2621–2630, 2012.

[52] R. Twiddy, J. Cavallo, and S. M. Shiri. Restorer: A visualization technique for handling missing data. In *Proceedings of the conference on Visualization'94*, pp. 212–216. IEEE Computer Society Press, 1994.

[53] A. Unwin, G. Hawkins, H. Hofmann, and B. Siegl. Interactive graphics for data sets with missing values:MANET. *Journal of Computational and Graphical Statistics*, 5(2):113–122, 1996.

[54] B. W. Wong and M. Varga. Black holes, keyholes and brown worms: Challenges in sense making. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, pp. 287–291. SAGE Publications Sage CA: Los Angeles, CA, 2012.

[55] J. Wood, P. Isenberg, T. Isenberg, J. Dykes, N. Boukhelifa, and A. Slingsby. Sketchy rendering for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2749–2758, 2012.

[56] Z. Xie, S. Huang, M. O. Ward, and E. A. Rundensteiner. Exploratory visualization of multivariate data with variable quality. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pp. 183–190. IEEE, 2006.

[57] J. Zacks and B. Tversky. Bars and lines: A study of graphic communication. *Memory and Cognition*, 27:1073–1079, 1999.