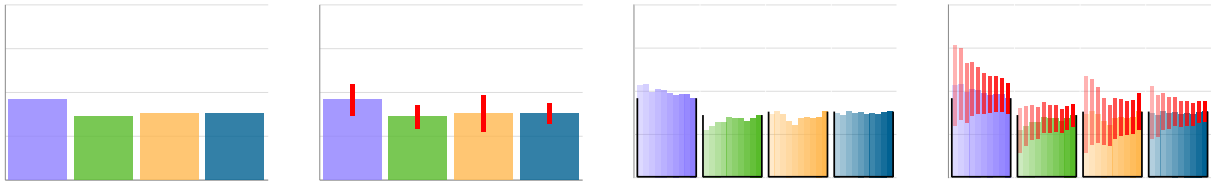
Studying Early Decision Making with Progressive Bar Charts

Ameya Patil, Gaëlle Richer, Christopher Jermaine, Dominik Moritz, Jean-Daniel Fekete

Studying Early Decision Making with Progressive Bar Charts

Ameya Patil, Gaëlle Richer, Christopher Jermaine, Dominik Moritz, and Jean-Daniel Fekete, *Senior Member, IEEE*



(a) ***baseline*** (b) ***ci*** (c) ***history*** (d) ***history_ci***

Fig. 1: Four progressive bar chart designs showing the means for four columns of a data table loaded progressively, updated every second: (a) baseline bar chart—***baseline***, (b) bar chart with confidence intervals (95%)—***ci***, (c) bar chart with near-history—***history***, and (d) bar chart with near-history and confidence intervals—***history_ci***. All the bar charts represent the same data at one step of the progression. For our proposed new designs (c) and (d), opacity and position along the X-axis encode the recency of the updates; opaque bars on the right represent more recent updates, and transparent bars on the left represent older updates.

Abstract—We conduct a user study to quantify and compare user performance for a value comparison task using four bar chart designs, where the bars show the mean values of data loaded progressively and updated every second (progressive bar charts). Progressive visualization divides different stages of the visualization pipeline—data loading, processing, and visualization—into iterative animated steps to limit the latency when loading large amounts of data. An animated visualization appearing quickly, unfolding, and getting more accurate with time, enables users to make early decisions. However, intermediate mean estimates are computed only on partial data and may not have time to converge to the true means, potentially misleading users and resulting in incorrect decisions. To address this issue, we propose two new designs visualizing the history of values in progressive bar charts, in addition to the use of confidence intervals. We comparatively study four progressive bar chart designs: with/without confidence intervals, and using near-history representation with/without confidence intervals, on three realistic data distributions. We evaluate user performance based on the percentage of correct answers (accuracy), response time, and user confidence. Our results show that, overall, users can make early and accurate decisions with 92% accuracy using only 18% of the data, regardless of the design. We find that our proposed bar chart design with only near-history is comparable to bar charts with only confidence intervals in performance, and the qualitative feedback we received indicates a preference for designs with history.

Index Terms—Progressive visualization, Uncertainty, Bar charts, Confidence intervals.

1 Introduction

Interactive visual data analysis on big data can suffer from poor user experience due to long wait times, which affect the user's analysis process [33]. Progressive visualization systems address this interactivity issue by incrementally loading, processing, and visualizing data. As more data is processed over time, the visualization is refined from a crude to an increasingly accurate estimate. Although the intermediate visualizations enable users to decide early, their limited accuracy can make them unreliable, possibly leading to incorrect decisions. In this article, we study how fast and accurate people are when making decisions using progressive bar charts, and more specifically, the benefits of confidence intervals, and that of *near-history* representation which visualize the recent history of each bar, as shown in Fig. 1.

To understand the benefits and challenges of progressive analysis, consider the example of tracking the vote count for a US presidential election. Typically, the results are tallied and sometimes even visualized as votes are received, to determine as early as possible which party will win. To compare estimates accurately, one needs to know if the intermediate visualization can be relied on to make inferences or if it is still insufficient. In the case of tracking vote counts, the processing order of states and the various types of ballots—paper, digital, mail-in, etc., create known biases in the early results. In a progressive system, the process is often simpler and is not biased since the data is loaded or processed in reasonably random order. Yet temporal artifacts can still happen, resulting in possibly misleading false patterns. Thus, the main challenge of progressive visualization is to accurately convey the uncertainty of intermediate results to help people avoid mistakes while supporting them in making decisions as soon as possible. Uncertainty also has a dynamic aspect related to the convergence of results: the instability of an estimate over time signals that it may not reflect the final result. Intuitively, users will monitor its stability for some time before trusting it. To help gauge stability, some systems display the evolution of the quality of intermediate results over time (e.g., [7]).

In this article, we focus on progressive bar charts and investigate both uncertainty and stability cues for early decision-making. Progressive bar charts are standard bar charts where the heights of the bars change at every update (e.g., every second) as estimates are refined until they settle to their final value. Many progressive visualization systems use bar charts [13, 19, 21, 38, 45, 60], usually with confidence intervals shown as error bars to indicate estimate uncertainty. Prior work on uncertainty visualization has shown that uncertainty in data is difficult to understand in static visualizations [25, 27], and in particular confidence intervals visualized as error bars are inaccurately interpreted for static bar charts [11]. We thus study the usability of confidence intervals for bar charts in the progressive setting. Further, to better show

- *Ameya Patil is with University of Washington, Seattle. E-mail:* `ameyap2@cs.washington.edu`.
- *Gaëlle Richer and Jean-Daniel Fekete are with Inria & Université Paris-Saclay. E-mail:* `firstname.lastname@inria.fr`.
- *Christopher Jermaine is with Rice University. E-mail:* `cmj4@rice.edu`.
- *Dominik Moritz is with Carnegie Mellon University. E-mail:* `domoritz@cmu.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

the uncertainty due to the instability of the progression, we introduce a *near-history* representation that displays the history of estimates as traces (Fig. 1c). The goal of this representation is to ease assessing the stability of the estimate by explicitly visualizing its trend over the progression. We also pair near-history with confidence intervals (Fig. 1d), thus also visualizing the trend of the uncertainty of the estimates.

We present an empirical study on human decision-making performance using progressive bar charts and investigate whether near-history representations and confidence intervals improve early decision-making, i.e., lead to earlier decisions or fewer mistakes. We compared four designs for progressive bar charts: standard, with confidence intervals as error bars, with near-history, and with both confidence intervals and near-history. In a crowdsourced study, we ask participants to compare pairs of bars in a progressive setting. We used three realistic datasets: data generated from commonly found distributions (*Normal* and *Power law*) [24], and data generated progressively using the *Wander Join* algorithm for join operations in databases [32] on a real-world dataset. Note that we generated randomly ordered data while not all real-world datasets come in random order. We evaluated the four designs quantitatively based on participants' response time and their error rate computed against the ground truth (the comparison of the final bar values) and qualitatively based on participants' confidence ratings. To estimate how early participants answer compared to the optimal time for a given accuracy, we also compared their performance to those of automated decision procedures. The results of our study suggest that people can make early decisions reliably using progressive bar charts, with decisions taken in 22 s on average, on our test having a maximum duration of 120 s (18% of the data), with 8% error. We measured a 14 s difference on average between the response times of participants and automated decision procedures on the same tasks for the same error rate. We report that confidence intervals and the near-history display significantly improve user performance over the standard bar charts, and bar charts combining both confidence intervals and near-history, but effect sizes remain small. User confidence ratings align with this trend and show the potential of near-history display for informing users on the uncertainty and stability of intermediate results. In summary, we contribute the following:

- an empirical study on the usability of progressive bar charts for a value comparison task;
- new visualization designs for progressive bar charts showing *near-history* information;
- a comparative evaluation of standard bar charts with/without confidence intervals, and with/without near-history.

We also introduce the use of sequential tests in the context of progressive visualization to assess human performance.

2 Background

Our research questions and the motivation behind our proposed bar chart designs are informed by the existing progressive visualization systems and practices in decision-making. We briefly present these points in this section.

2.1 Progressive Visualization

An important aspect of visual data analysis is the interactive latency [33, 50]. Progressive visualization techniques reduce interactive latency, thus improving user experience, by breaking down different stages of the visualization pipeline into iterative steps (e.g., [19, 20, 37, 45, 52]). These techniques have also been shown empirically to be effective at mitigating the latency in visual explorations [7, 19, 58]. However, this iterative procedure means that the initial versions of the visualization show approximations which are refined over time. This gives rise to certain challenges with using progressive visualizations.

Analysis in progressive visualizations can be characterized temporally in three phases: *early partial results* that are not yet reliable, *mature partial results* that are trustworthy but susceptible to small changes, and *definitive results* that are stable [3]. The length of these three phases is affected by the speed of access to the data, the data distribution, the order in which it is processed, and the underlying computation processing it. The user essentially tries to balance between speed and accuracy along these three phases [32, 44]. The challenging aspect is the uncertainty in the data or computations and the resulting visualizations. Uncertainty in the visualizations tends to make users less confident about the conclusions drawn from them [7, 38]. Thus, in our study, we evaluate the efficacy of confidence intervals in the progressive setting and propose new designs for progressive bar charts.

2.2 Decision-making in the Progressive Setting

Two factors can affect decision-making in the progressive setting: errors due to repeated testing, and biases due to uncertainty.

Inferential methods inform on a property about a population, by studying a sample from the population and generalizing the results to the population. When significance tests are applied to multiple independent-drawn samples, it increases the probability of false positives. This undesirable effect of repeated testing is known as inflation of the family-wise error rate [1]. Significance tests performed on progressively increasing samples are not independent since data points from previous samples are also accounted for in subsequent significance tests. Repeated testing causes problems of inflated error rate even in this case, although not as severe as in the case of multiple testing on independent samples [5, 55]. In both cases, the issue occurs due to repeated application of the test until the p-value drops below the chosen significance level, biasing the outcome of the experiment. This is a well-studied problem in the field of *sequential analysis* that develops methods for performing repeated tests on progressive samples without the assumption of a fixed sample size [55].

The problem of inflated error rates has been addressed before by correcting the significance level for each test [4, p. 27] depending on the number of times the test is planned to be applied and using confidence sequences [23]. However, these corrections may be too strong in the context of progressive visualizations where the user does not necessarily decide at every time step, potentially switching focus between different tasks. More relevant to our work is the work of Zgraggen et al. [59], which highlights the problem of repeated testing specifically during visual analysis, but for static visualizations. We investigate whether humans are susceptible to this issue in the progressive setting, where there may not be enough time to decide at every step.

Decision-making may also be affected by how progressive estimates change over time and the biases that this may introduce. Users predict the ground truth based on the intermediate results they have seen so far. This reliance on trends in past data to predict the future could lead to *inference uncertainty* [51]. Yet another possible source of error is misjudgments of the uncertainty due to the dynamic nature of visualizations in a progressive setting, also known as *uncertainty bias* [36, 44]. This bias includes biases due to human tendencies to distrust fluctuating values—*ambiguity bias* [6], and neglect the visual glyphs for uncertainty visualization—*neglect of probability bias* [53]. Finally, false patterns that occur in intermediate results have also been shown to affect decision-making, known as *illusion bias* [36, 44].

2.3 Visualizing Uncertainty for Progressive Analytics

Skeels et al. [51] categorize uncertainty as: *measurement* uncertainty—due to imprecise data collection, *completeness* uncertainty—due to missing data, and *inference* uncertainty—due to making inferences based on models created from past data. Progressive visualizations systems potentially suffer from 2 levels of uncertainty—*completeness* uncertainty, and *inference* uncertainty. Progressive visualizations should ideally depict any uncertainty for users to make informed decisions. However, the very nature of uncertainty and the challenge of defining the correct behavior when using uncertainty visualizations, makes it difficult to account for uncertainty when analyzing data [25, 27].

Many techniques have been proposed to visualize the *measurement* and *completeness* uncertainty [43]. Traditionally, uncertainty is represented with error bars and box plots [8]; more recently, violin plots [22] and gradient plots [28] have been used as alternatives that allow for a continuous encoding of uncertainty [11]. Gradient plots and violin plots were found to be better than error bars for understanding uncertainty [9, 11]. Animation has also been shown to be effective for judging uncertainty in static data, even for non-trained users [30]. In

progressive visualizations, using animation to convey the uncertainty of intermediate results may conflict with the animation due to progressive updates. Despite the new designs developed for uncertainty visualizations, bar charts with error bars for confidence intervals remain the standard in many progressive visualization systems today [38, 45, 60]. We thus narrow the scope of our study to bar charts with error bars and their variations since they have not been evaluated yet in the progressive context and the proposed near-history representation also integrates more easily with the standard encoding than with e.g., violin plots.

On the other hand, there is very little prior work on visualizing the stability or convergence, which is one of the sources of *inference* uncertainty in progressive visualizations. Fisher [18] called for research on the visualization of convergence trends in progressive visualization systems. Visualizing stability in progressive visualization is connected to visualizing changes in data. Robertson et al. [48] studied the effectiveness of visualizing changes in data using traces, animation, and small multiples, and found traces and small multiples to be better for visual analytics. This motivates our proposed near-history representation that follows the principle of traces, and we briefly discuss an alternative based on small multiples in Sect. 3.2.

3 Study Rationale

Comparing values is one of the fundamental low-level visual analysis tasks, used in many of the high-level tasks like filtering, sorting, finding trends, finding extrema, etc. [2]. These tasks are the basis for user intents like understanding who won the elections or understanding crop yield trends over a year. Since value comparison tasks are performed efficiently on bar charts compared to other visualization types [39, 49, p. 152], we focused the study on value comparison using bar charts.

Comparing values with a progressive bar chart differs from a traditional static bar chart. Typically, the decision is that the value depicted by the height of bar A is higher than, lower than, or is not significantly different from the other bar B. With a static bar chart, the user can answer with any of the three options. However, with a progressive visualization, at any time of the progression, the user has also the option to *wait before deciding*. The progressive setting has an additional uncertainty due to the way data is sampled until it has been entirely processed. Even after all the data is processed, it can have an intrinsic uncertainty, which is not considered in our experiment. We want to determine if humans can make accurate decisions with progressive bar charts and if they can make these decisions early i.e., before the end of the progression, ideally as soon as the level of uncertainty allows for it.

Specifically, we focus on the task of comparing the mean estimates of two distributions A and B, represented as two progressive bars. We generate data distributions to have a non-zero significant difference between their true means so that either A or B is truly larger. We do so to avoid making the user wait unnecessarily until the end of the progression due to the possibility of there being no significant difference. We investigate whether users make mistakes, deciding that A is larger when actually it is B, or the opposite, by deciding too early.

3.1 Research Questions

The goal of our study is to understand how useful progressive bar charts are for decision-making and find ways to improve them. In particular, we identify the following research questions:

- **(R1)** How fast and accurate are humans when making decisions using progressive bar charts?
- **(R2)** How do different designs for progressive bar charts affect human performance when making decisions?
- **(R3)** How do different designs for progressive bar charts affect human confidence?

To answer these research questions, we performed a crowdsourced study, described in Sect. 4. We evaluated four progressive bar chart designs, shown in Fig. 1. Two are commonly used in progressive systems today: the standard bar chart—***baseline*** (Fig. 1a), and the bar chart with confidence intervals—***ci*** (Fig. 1b). In the following subsections, we motivate our proposed near-history designs for progressive bar charts, describe design considerations for near-history visualizations, and present our hypotheses.

3.2 Why Near-History Visualizations

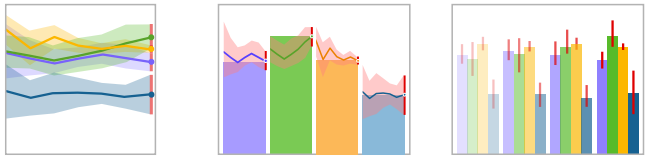
In progressive analytics, one goal is to help users to decide when to decide. A decision can be made when (a) the estimates have low uncertainty or variance, and (b) the estimates have stabilized over time. $x\%$ confidence intervals (CI) for an estimate give a range of values that is $x\%$ likely to contain the ground-truth value [14], taking into account the estimate's variance and number of contributing data points. While CIs convey the uncertainty or variance of the estimates, users still have to rely on their memory of the past estimates to assess the stability of estimates. To avoid this additional cognitive load, we propose to extend progressive bar charts with *near-history* (traces). Instead of visualizing only the latest progressive estimate (and CI), we visualize the last n progressive estimates (and CIs), refer Fig. 1c and Fig. 1d.

In a progressive setting, since the convergence of estimates to the true values cannot be assessed in practice, patterns of stability in the progressive estimates are a useful indicator of their convergence. More precisely, an instability pattern acts as a proxy for the non-convergence of the estimate. Eventually, after a substantial amount of data is processed, it is more likely that both the CI and near-history information will suggest that a decision can be made i.e., narrow CI with a stability pattern. Even in this scenario, if the progression happens to process an outlier next, a decision made before that point may be incorrect. Both CIs and near-history only provide precursory information about how the progression might evolve in different ways. They provide sufficient conditions to **not decide** at any point, but not to **decide** at any point. The value of near-history is in providing the user a glimpse of how the estimate and the CIs have evolved so far so that any inaccuracy in decision-making can be better anticipated. The benefit of the near-history representation is to offload the task of remembering the trend of the estimates and variance. We thus hypothesize, based on the findings from Hullman et al. [26], that more visual information should lead to more correct decisions with more confidence, which we formally present in Sect. 3.4.

3.3 Design of Near-History Visualizations

When iterating on near-history designs, we compared different encodings, refer Fig. 2. Our main design requirements were: (D1) visually scale to many bars, (D2) compare the latest estimates, (D3) compare the error bars of the latest estimates, and (D4) assess the trend of estimates.

We first considered encoding sequences of estimates with overlaid colored polylines with error bands for CIs, Fig. 2a. Although this encoding is a natural choice, it can become cluttered by many categories close in range of values (D1). Closer to the traditional bar charts, we considered a mixed encoding where the polyline is nested in a bar representing the last estimate value, Fig. 2b. However, to keep the encoding of near-history estimates and CIs as close as possible to their encoding in their counterpart standard bar charts, we discarded these designs. Focusing on encodings with both the history of estimates and CIs as bars, we considered two arrangements. The first arrangement spatially groups historical mini-bars by recency similar to juxtaposing multiple bar charts (small multiples), Fig. 2c. The second arrangement is the design used in our experiment which we refer to as ***history*** (Fig. 1c). It spatially groups historical mini-bars by category where the recency of updates is encoded by the position along X-axis, and opacity. Thus, mini-bars for old estimates are on the left and are more transparent than those for the newer estimates. These two designs have different trade-offs regarding our requirements. The design chosen in our study



(a) Overlaid bands. (b) Bar-nested bands. (c) Juxtaposed bars.

Fig. 2: Alternative near-history designs, with confidence intervals.

can make requirements (D2) and (D3) difficult due to an increased gap between the visual marks to compare [54]. On the other hand, the alternative design can make requirement (D4) difficult. To mitigate the issue due to increased separation in our chosen design, we frame each group of mini-bars with a black outline whose height encodes the latest estimate, thus bringing the marks to be compared, closer. We discarded the alternative design considering requirements (D1) and (D4)—with more bars, trend assessment would be almost impossible. We touch upon the use of interactions to achieve requirement (D1) in Sect. 6.4, but do not consider it in our study.

We also visualize the history of the CIs in ***history_ci*** (Fig. 1d). The idea is to combine the benefit of CIs with that of near-history information. Effectively, this design offloads the task of remembering the trend of both the estimate and its uncertainty or variance over time, by explicitly visualizing it. Although the increased number of visual elements in ***history***, and even more in ***history_ci***, could increase the time required to process or engage with it, it can also improve understanding [26].

Other techniques like gradient plots have been found to be better than error bars [11]. However, prior work by Procopio et al. [44] did not find significant difference between the two in the progressive setting. Also, since bar charts are ubiquitous in progressive visualization systems today, we first evaluate near-history visualization with bar charts, and leave their evaluation with gradient or violin plots as future work.

The number of mini-bars in the near-history encodings can affect user performance. Fewer mini-bars would give lesser cues about the stability pattern, thus requiring users to memorize the trend more, especially if the progressive updates include frequent outliers and are thus chaotic. On the other hand, more mini-bars may overwhelm the user with increased movement in the visualization and take more screen real estate. Although it would be important to evaluate the effect of the number of mini-bars, we fixed it in our study, to evaluate the feasibility of near-history visualizations as the first step, and reduce the parameter space of the experiment. We fixed the number of mini-bars to 10 as a balance between providing sufficient stability cues and keeping the chart area small and leaving the determination of an optimum number of mini-bars for future work. The chart area is kept the same across all four designs (with/without near-history) in our study to avoid any confounding bias due to the chart area.

3.4 Measures and Hypotheses

For the value comparison task in the progressive setting in our user study, we use the following measures to express our hypotheses:

Accuracy: Accuracy is the ratio of correct decisions to the total number of decisions. A decision is a choice between "A is larger than B" and "B is larger than A" at an intermediate stage of the progression. The correctness of a decision is determined by the ground-truth i.e., the relation between A and B once all the data is processed.

Response Time: Response time is the time taken to decide from the beginning of the progression in seconds. Equivalently, it can be expressed in the number of updates or the amount of data processed since the same amount of data is added each second in the experiment.

User Confidence: User confidence is a subjective measure of usability collected per visualization on a 7-point Likert scale and done retrospectively so that participants can reflect on all visualizations. Note that we asked participants to rate the visualizations and not their decisions on individual tasks, since we want to evaluate the visualization designs themselves and not the decisions.

Our hypotheses for each of our research questions are:

- **(H1)** *Humans can achieve high accuracy using progressive bar charts.* We hypothesize that participants can achieve an accuracy of at least 80% for the value comparison task using progressive bar charts. We chose the threshold for accuracy partly based on results from prior work in decision making for value comparison tasks on **static bar charts** [49] (92%), progressive bar charts [44] (86%), and partly based on our subjective view of how accurate we think is good enough in the **progressive setup**, similar to the choice of .05 as the standard level of statistical significance [12].
- **(H2)** *In terms of user performance (accuracy and response time), bar charts with history and confidence intervals are better (higher accuracy and earlier response) than only with history or only with confidence intervals, both of which are better than plain bar charts i.e.,* ***history_ci*** $>$ (***history*** *and* ***ci***) $>$ ***baseline***.
- **(H3)** *In terms of user confidence* ***history_ci*** $>$ (***history*** *and* ***ci***) $>$ ***baseline***.

We base the last two hypotheses on the motivation behind near-history visualizations discussed in Sect. 3.2.

To assess whether the response times are *early* or rather *late*, we would ideally compare them with a reference response time. To compute these references, we investigated optimal sequential tests adapted to our comparison task, which would also be free of parameters and distribution assumptions to be fair to the participant's knowledge of the data. Here, optimal means minimizing the expected response time (also called sample size) for a fixed upper bound on the probability of error δ (usually $\delta = .05$). However, having an optimal decision test for unknown parameters even under the normality assumption is challenging: most available solutions are only approximate or asymptotically optimal [15, 42]. Without an ideal test, we compared human response times to approximations of optimal times computed with automated decision procedures, which we describe in Sect. 4.4. Note that in this work we only consider automated decision procedures as a means to obtain reference response times, to quantify how efficient humans are. We *do not* study these automated procedures to integrate them into progressive analytics systems or to completely replace humans in the decision-making process.

4 Study Design

To address our research questions, we conducted a crowdsourced study on Prolific. We asked participants to compare the mean values of pairs of distributions using the four progressive bar chart designs, as **accurately** and **quickly** as possible. Existing literature on crowdsourced studies shows that participants tend to complete the tasks as quickly as possible [29, 34, 35, 46]. Therefore, we awarded a £0.31 bonus (10% of the base rate of £3.13) if their overall accuracy was more than 50%, to discourage participants from botching tasks and encourage them to balance their accuracy and speed. Participants could also track their overall accuracy and thus, expected bonus through the experiment.

The study was performed in two rounds—a pilot study and the final study. Based on the findings and participant feedback from the pilot study, we made changes to the study design for the final study on two broad aspects: participant screening, and task complexity. The results did not change significantly between the two rounds, so the subsequent discussion will refer to the final study unless specified otherwise.

4.1 Participants

We restricted our participant pool to people with at least a college-level degree and formal education in science, mathematics, economics, and finance. Although we provided the required background about the use of progressive visualization, confidence intervals, and near-history information, our pilot study revealed possible difficulties for participants without some basic statistics background in understanding the experiment. Thus, to ensure a low rejection rate of responses, we added the aforementioned restrictions.

4.2 Tasks and Procedure

We performed a full-factorial within-subjects study for two factors: bar chart design (4) and dataset (3), and with four serial repetitions per pair of factors. We used a Latin square design for the design and dataset combinations. Each participant was randomly allocated a set of 48 comparison tasks (owing to 16 per dataset and 12 per bar chart design). Each comparison task corresponds to a pair of precomputed data sequences represented by two progressive bars, A and B, that are to be compared by participants. Each dataset is made of 100 pairs of precomputed data sequences, with their associated ground truth, i.e., which of A and B is truly larger. The pilot study had only one comparison task per trial page, i.e., asked participants to compare only two bars at a time, with which we saw a ceiling effect for accuracy. To mitigate this ceiling effect and also to make our setting more realistic,

we switched to 4-bar bar charts in the final study and asked participants to compare two separate pairs of bars simultaneously on each. Fig. 3 shows a sample trial page for our revised study. The bars to be compared were randomly interleaved (compare A with B, C with D **or** compare A with C, B with D), with an equal number of each configuration for each combination of design and dataset. Different colors were used for each of the four bars to ensure a clear separation .

Each user session started with a consent page, followed by a tutorial on progressive visualization, the usage of confidence intervals, and near-history, followed by an explanation of the task—to compare the mean values of pairs of distributions using the progressive bar chart, as accurately and quickly as possible. We asked participants to self-report their level of skill in information visualization and statistics and provided one training task per design before starting the recorded tasks. Each trial page included two questions, four answer buttons, and a blank visualization. When ready, participants could click the 'Start' button to start the progressive visualization updates; this also started a timer. To answer the questions, participants could click on one of the two answer buttons for each question (Fig. 3). The progression stopped when both questions were answered or when all data were processed. Above the chart, a progress bar showed the progression as a percentage of processed data. Updates were made each second, for a total time of 120 s; however, participants were not explicitly informed that all progressions would last for 120 s, so as to nudge them towards thinking in terms of data processed and not time spent. As illustrated in Fig. 3, clicking an answer button shows textual feedback about the correctness of the answer and updates the bottom bar with the new percentage of correctly answered questions. This was done to nudge participants to focus on accuracy rather than time. Finally, after answering all tasks, participants rated each design based on how confident they were using it to answer the tasks and explained their ratings.

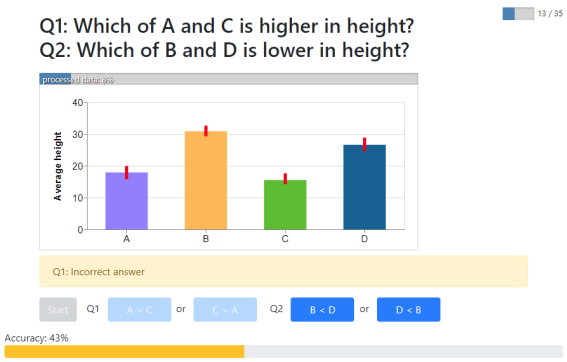


Fig. 3: Example trial page with two simultaneous tasks (Q1, Q2). A progress bar (blue) shows the percentage of data processed. Besides immediate textual feedback upon answering, participants can track their overall accuracy throughout the study on the yellow progress bar.

4.3 Datasets

The bar charts shown to the study participants visualized the mean of progressive samples of data drawn from a distribution. We used three different realistic data distributions to generate the datasets for our experiment. Two datasets are synthetically generated using two common distributions in real-world datasets: *Normal* and *Power law* [24]. The third is made from the output of a progressive algorithm that computes the mean of the result of a join operation over real-world data. The four bars on each trial page corresponded to four different data sequences from the same dataset, i.e., generated from the same process but with different parameters. Each data sequence is an ordered set of 10k data points drawn from a distribution with a known true mean. Each dataset is a set of 100 pairs of data sequences to be compared, generated such that the difference between their means is always the same.

Synthetic Data: *Normal* and *Power law* Distributions

For the first dataset, ***normal***, we used *Normal* distribution with mean and standard deviation chosen to represent real-life distributions of heights of dogs ($\mu \in [12,30], \sigma^2 \in [25,35]$). For the second dataset, ***power law***, we used *Power law* (or Pareto) distributions to represent salaries in a population [47]. A *Power law* distribution has a density function of the form $p(x) = (\alpha-1)x_{\min}^{\alpha-1}x^{-\alpha}$ and a finite mean for $\alpha > 2$. Most naturally occurring *Power law* distributions, including salary distributions, have $\alpha \in [2,3]$ [40]. Thus, to keep the dataset realistic while ensuring meaningful mean estimates, we generate parameter $\alpha \in [2.75, 3.5]$ for our experiment (and parameter $x_{\min} \in [3,10]$). To generate a data sequence for the ***normal*** or ***power law*** datasets, we first pick its distribution parameters at random within the ranges specified above. Next, we draw a sample of 10k data points from the distribution thus defined. The sequence of *progressive samples* is formed by accumulating chunks of 84 data points from this sample. For each progressive sample, the empirical mean and associated 95% confidence intervals are computed using bootstrapping [16].

Real-world Data: *Wander Join* Algorithm

Our third dataset was generated using the *Wander Join* algorithm [32] on the flights dataset [57] to get the flight delay data. *Wander Join* algorithm computes approximate results for join queries and has been the foundation of subsequent work on progressive aggregation over joins [45] owing to its suitability for OLAP use cases. The algorithm works by performing *random walks*. In each random walk, the algorithm randomly selects a tuple across the tables that satisfies the join clause and computes the required aggregation over multiple such walks, along with the associated uncertainty. In our study, we asked participants to compare the average flight delay for two categories of flights. To sample approximately 10k data points for the two categories, we performed 2×10k random walks with appropriate filter clauses in the query. The mean estimates and confidence intervals were computed as part of the algorithm i.e., no bootstrapping was performed. We refer to the dataset created using this algorithm as ***wander join***.

The stationarity of a process generating time series data means that the distribution of the generated data is constant over time, and consequently the statistical properties, such as the mean and variance. In a strong sense, this means that the distribution of different chunks of the generated data is the same. This kind of statistical consistency is desirable in the progressive setting because it allows for reliable predictions about the true estimate based on the intermediate progressive estimates. The generation process for ***normal*** and ***power law*** aims for stationarity: data sequences are random samples from a single distribution. *Wander Join* is an example of an algorithm that tries to generate a stationary sequence of progressive data but cannot avoid abrupt changes due to possible outlier values.

We generated distributions in pairs, such that the true mean of either one is necessarily greater than the other and that the empirical means computed at the end of the progression reflect the order of the true means, i.e., shows the correct answers. We also ensured that the mean estimate at the end of the progression was within 5% of the true mean of the distribution. Further, to avoid any confounding bias due to the difference between the true means of the two distributions, we ensured the same absolute difference of 2 units between the true means. This was done by adding an offset value to one of the two distributions in a pair. In some cases, this offset caused some data points to have impossible values, e.g., negative salary. However, the true means remained possible values so the data distribution was not changed in any way. We also used the same Y-scale for the bar charts in all trials, which ensures that the absolute difference of 2 units between true means is translated to an absolute visual difference as well.

To summarize, the participants were shown a progression of a total of 10k data points per bar, with updates once per second corresponding to 84 new data points. Similar to Procopio et al. [44], the data generation, computation of the means, and confidence intervals of each progressive sample were done offline. This allowed for assigning the same datasets to multiple participants and ensured that all saw the same progressions. During the experiment, the visualization was updated every second with the corresponding pre-computed mean estimates and confidence intervals for all four bars.

4.4 Calculation of Measures

We used accuracy and response time as evaluation measures for human performance and compared them with the accuracy and response time of decision procedures to estimate how far people are from the optimal response time. Since finding a sequential and distribution-agnostic decision procedure is an open problem, we use three decision procedures, chosen for their specific benefits: the *confidence interval procedure*, the t-test, and the *General Likelihood Ratio Test* (GLRT). The *confidence interval procedure* is a valuable baseline since it uses the same information as the one visually shown to participants with the ***ci*** and ***history_ci*** designs. The *t-test procedure* is based on the well-known t-test, and, as such is easy to implement. Finally, the GLRT is a sequential test tailored to our comparison task. We run the procedures at the same data rate as in the user study i.e., progressive samples grow by the same amount of data points as between visualization updates and until a decision is reached (that either A or B is larger) or until the progression completes. For a fair comparison with humans, the procedures are entirely non-parametric; they decide solely based on empirical statistics (mean and variance) of the current progressive samples and the specified error threshold that we denote by δ.

Confidence Intervals Procedure This procedure uses $(1-\delta)\%$ confidence intervals (CI) for the mean for every pair of progressive samples, as computed by the *Wander Join* algorithm or by the bootstrapping method without any correction for repeated testing. To compare the mean estimates of A and B, it compares the overlap between the CIs for A and B. If at any instant there is no overlap, the procedure decides that the highest mean is the largest. Otherwise, it follows *Rule 4* described by Cumming [14, p. 7] which computes the *proportion overlap* of the CIs and allows to decide with 95% confidence when proportion overlap is 50% or less. With a larger proportion overlap, the procedure does not decide yet. Essentially, this inference rule performs a t-test using CIs, which we apply to the bootstrap and *Wander Join* CIs rather than regular CIs. This inference rule was included in the user study tutorial.

t-test Procedure This procedure runs a Welch's two-sample t-test [56] for every pair of progressive samples. If the p-value is larger than δ, the procedure continues. Otherwise, the procedure decides which of A or B is larger by comparing their empirical means. Said graphically, the procedure computes the function $f(t)$ giving the p-value at every step t of the progression and stops once it crosses the boundary $y=\delta$.

Generalized Likelihood Ratio Test Wald's Sequential Probability Ratio Test (SPRT) [55]), also called sequential likelihood ratio test, is a sequential test to decide between two simple hypotheses (of the form $\mathcal{H}_0 : \mu = \mu_0$ against $\mathcal{H}_1 : \mu = \mu_1$) for fixed error probabilities. The test is formulated as a boundary-crossing procedure: at each step, it compares the likelihood ratio, which measures how likely the observed values support one hypothesis against the other, with a boundary function defined to guarantee certain levels of error. While SPRT's boundary functions are proven to be optimal, the test does not support the decision task of our study. Generalized likelihood ratio tests (GLRT) [41] extends the likelihood ratio test approach to support composite testing problems such as the one modeling the task of our study: $\mathcal{H}_0 : \mu_A > \mu_B$ against $\mathcal{H}_1 : \mu_B > \mu_A$. The GLRT statistic is the ratio between the likelihoods of the best-fitting parameter value from each hypothesis. The decision rule of the GLRT compares this statistic with a threshold function. If the statistic is smaller, the procedure continues. Otherwise, the procedure decides which of A or B is larger by comparing their empirical means. To achieve a reliable testing procedure, one has to devise a threshold function matching the given levels of error, in our case δ for both hypotheses. Our chosen threshold function follows the results developed for A/B testing by Kaufmann et al. [31].

The three procedures remain heuristics since their underlying decision rules assume normality, which is broken for two out of the three datasets. Therefore, while all procedures take a δ parameter, they might be too optimistic, i.e., fail more often than specified over many examples, or pessimistic i.e., decide late and fail less often than the specified error rate δ, by a large margin. This is even more likely for the CI and the t-test procedures that also ignore the sequential context and could suffer from repeated testing bias.

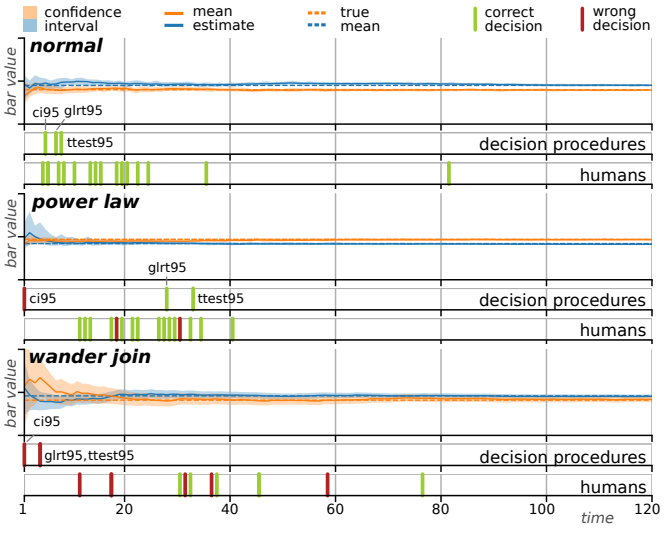


Fig. 4: Examples of task data from the three datasets. For each dataset, the top plot shows the mean estimate and confidence interval at each time step for the two values to be compared (in blue and orange), and the bottom plots show the corresponding response times and correctness of the decision procedures and humans (12 to 21 participants) with stripes. A decision stripe is green if the decision is correct, red otherwise.

5 Results & Analysis

Based on the results of our study, we present an analysis of human performance in the progressive setting, evaluate the progressive bar chart designs, and compare human performance with that of automated procedures. For the final study, we rejected 12 participants out of 90 owing to quality issues, to have a total of 78 participants. 80% of participants were at least competent in information visualization and around 60% were at least competent in statistics. We used a t-test to compute the significance of differences between techniques for task accuracy, response time, and confidence rating (shown with a dashed line on figures), and Cohen's d coefficient to compute their effect sizes.

5.1 Task Complexity Level

Due to the nature of the distributions—*Normal* and *Power law*, and the *Wander Join* algorithm, the respective data sequences have different convergence behaviors. Thus, each dataset presents a different level of task complexity. Fig. 4 presents an example of pairs of data sequences (colored blue and orange) for each dataset, showing the mean estimates and confidence intervals at each step of the progression. We also show the decisions taken by the statistical procedures (described in Sect. 4.4) and for the sake of comparison, by all our study participants who were allotted that example pair of data sequences in their comparison tasks.

In general, the ***normal*** dataset is more well-behaved than the others, as can be seen in Fig. 4. Very rarely during the progression, an incorrect decision is made by both the procedures and humans. With the ***power law*** dataset, the two mean estimates tend to fluctuate early in the progression, which produces misleading intermediate results. Both the procedures and humans tend to decide later than for ***normal***. The ***wander join*** dataset shows the most chaotic progression of the three, with more fluctuations between the two mean estimates and their confidence intervals. Both the procedures and humans tend to decide later or more incorrectly compared to the former two datasets. Thus, the general trend is that tasks from the ***wander join*** dataset are the most complex, followed by ***power law***, and then ***normal***, as illustrated by Fig. 4. However, the random choice of distribution/algorithm parameters for generating the individual examples may result in certain examples being easier to compare in ***wander join*** than in ***power law*** or ***normal***. We do not explicitly control the task complexity but instead rely on the random process to generate a variety of examples per dataset. We present results at the dataset level during our analysis.

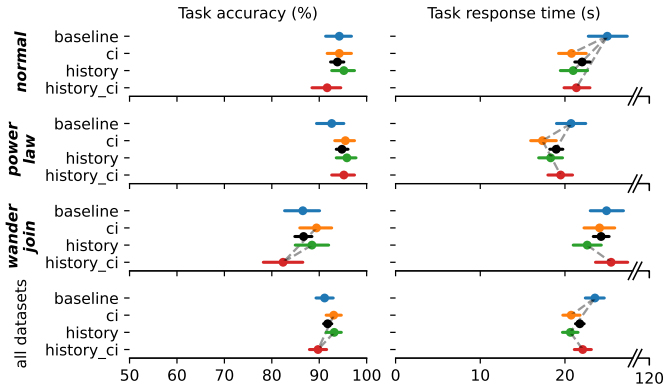


Fig. 5: Participant accuracy and response time for each bar chart design (colored marks), dataset (black marks), and overall (last row), with 95% confidence intervals for mean values and dashed lines for significant differences ($p < .05$).

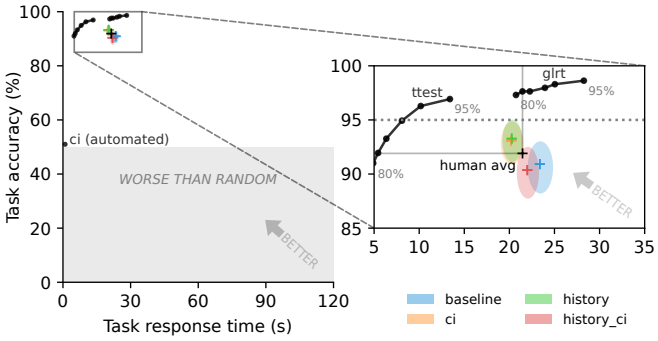


Fig. 6: Task accuracy vs. response time for each bar chart design across datasets (colored ellipses) with 95% confidence intervals for the mean visualized by ellipse height and width. Black lines show the reference performance of decision procedures.

5.2 Human Performance Evaluation

In this section, we report on the results of our study for task accuracy, response time, and user confidence ratings.

5.2.1 Task Accuracy and Response Time

Fig. 5 shows the percentage accuracy (left) and response time (right) averaged across all participants, for each design, for each dataset, and across all datasets. Both the ***normal*** and ***power law*** datasets show an accuracy of around 95% across all designs. Since the accuracy is high, these two datasets do not provide much insight into the efficacy of different designs; there is no significant difference in accuracy and the effect sizes are small ($d < .2$). For the more complex ***wander join*** dataset, the average accuracy is around 85% which highlights more differences. Specifically, ***history_ci*** design was found to be less accurate compared to both ***ci*** and ***history*** with a small effect size ($d \approx .2$), and the same is observed across all datasets ($d \approx .1$). For all datasets combined, participants were 92% accurate on average, with no clear winner/loser in terms of task accuracy. However, the comparatively lower accuracy for ***history_ci*** design in the complex ***wander join*** dataset and across all data sets suggests possible usability issues for ***history_ci*** in real-life scenarios.

The task response time is roughly the same at around 22 s, corresponding to less than 20% processed data (less than 2k data points out of 10k per bar) for each dataset and each design. There are significant differences between the response times for ***baseline*** and the other designs, in both the ***normal*** and ***power law*** datasets. However, the effect sizes are not large enough to be considered important with $d \approx .3$ (absolute difference of 2– 3 s). The ***wander join*** dataset has higher average response time compared to the other two owing to its higher complexity. This could indicate that users wait longer before deciding on chaotic progressions regardless of the visual representation. However, it gives a significant advantage to ***history*** compared to ***history_ci***. Overall, participants decided significantly earlier with the ***history*** and ***ci*** designs than with the ***baseline*** and ***history_ci*** designs. However, none of the differences have a considerable effect size ($d \approx .1$), with absolute differences of about 3 s.

Fig. 6 summarizes the results of human performance in terms of both task accuracy and response time. The bar charts which have high accuracy and low response time—the top-left corner of the plot, are better. The ***ci*** and ***history*** designs are comparable and marginally better compared to the ***baseline*** and ***history_ci*** designs.

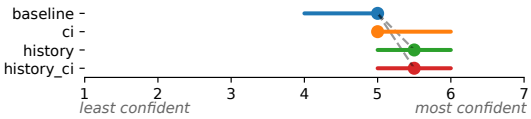


Fig. 7: Confidence rating for each bar chart design, with 95% confidence intervals for the **median** and dashed lines for significant differences ($p < .05$, Wilcoxon t-test).

5.2.2 Confidence Rating

To evaluate how confident users felt while using each design, we compared the confidence rating assigned to each, at the end of the study (1 for least, 7 for most confident). Although there are no decisive results about the efficacy of different designs in terms of accuracy and response time, Fig. 7 clearly shows that participants were the least confident with the ***baseline*** design, followed by the ***ci*** design with the ***history*** and ***history_ci*** designs enjoying the highest confidence. Only the differences between the confidence rating for ***baseline*** and ***history***, and for ***baseline*** and ***history_ci*** are significant.

5.3 Comparison to Decision Procedures

To put the performance of the participants in perspective, in particular how early their responses were, we compared them with the decision procedures described in Sect. 4.4 on the same data.

First, we checked how reliable the decision procedures are by comparing their accuracy with their pre-specified error threshold δ, here 5%. Fig. 8 shows the resulting accuracy and response time for the decision procedures run on the experiment data, for each dataset, and across all datasets, with a dotted line at 95% accuracy (the expected minimum accuracy). The t-test and GLRT were reliable as they exhibit more accuracy (97% and 99% respectively) than the pre-specified 95% threshold, and are even relatively robust on the non-normal datasets; only the t-test fails to match the threshold on ***wander join***. The CI procedure exhibits an average accuracy close to random overall, making it unusable. The fact that it also decides almost immediately shows that it is overly optimistic and probably not calibrated for the sequential setting. Thus, we exclude it from the subsequent comparison.

Next, we look at the average response time of the t-test and GLRT. They respond relatively early, with decisions at 14 and 30 time steps on average i.e., with less than 25% of the data or 2500 data points. Since the procedures are not theoretically optimal, we can only assume that optimal decision times are earlier than their actual response times.

Finally, to compare the human response time with decision procedure response times at equal accuracy, we run the t-test and GLRT while varying the accuracy threshold from 95% to 80% (i.e., δ from 5% to 20%). We adopt this sampling approach because both procedures are conservative and achieve higher accuracy than their pre-specified threshold. Fig. 6 shows the corresponding variation of accuracy relative to response time for t-test and GLRT as a black curve. This highlights how conservative GLRT is, with an achieved accuracy above 97% even at an 80% threshold. The reticle in Fig. 6's inset is centered on the human average and highlights the results at the same accuracy/response time. Compared to the average human performance, GLRT achieves 97% accuracy in the same response time (21 time steps), and the t-test

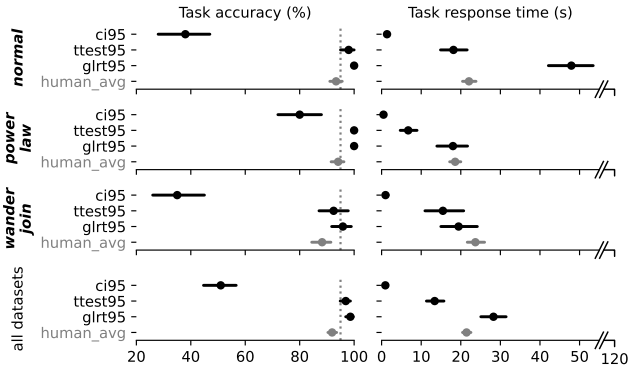


Fig. 8: Task accuracy and response time of automated decision procedures with 95% accuracy threshold (dotted line), for each dataset and overall, with 95% confidence intervals for mean values. In gray, the average human performance for reference.

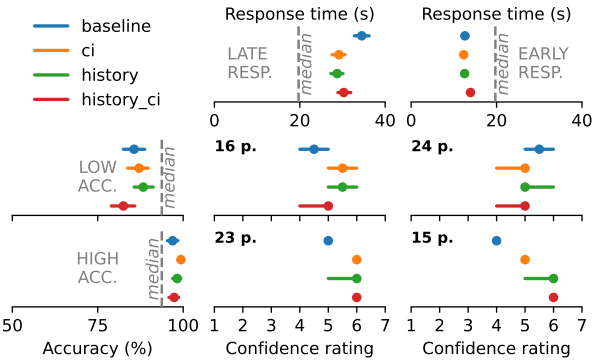


Fig. 9: Confidence ratings are broken down as per the participants' average compared to the median, for accuracy (rows) and response time (columns), with 95% confidence intervals for the **median** of confidence rating, and **mean** of accuracy and response time.

achieves the same accuracy (92%) in only 6 time steps. In summary, we find that human performance—both accuracy and response time, is close to the reference performance of decision procedures (GLRT and t-test). Although we find a difference of 15 time steps on average between the response of humans and the t-test for the same accuracy, the difference cannot solely be explained by a difference in the sample size required to decide, since measured response time for humans also includes a reaction time. We discuss this aspect in Sect. 6.5.

5.4 Hypotheses Evaluation

We now evaluate our hypotheses based on the results and analysis, thereby answering the corresponding research questions:

- ***(H1)*** *Humans can achieve high accuracy using progressive bar charts.* Our study participants were around 92% accurate across all conditions, with an average response time of around 22 s (18% data processed). We thus **accept** it.
- ***(H2)*** *In terms of user performance (accuracy and response time), bar charts with history and confidence intervals are better (higher accuracy and earlier response) than only with history or only with confidence intervals, both of which are better than plain bar charts i.e.,* ***history_ci*** $>$ (***history*** *and* ***ci***) $>$ ***baseline***. The quantitative evaluation failed to bring out a clear winner in both accuracy and response time. Both ***ci*** and ***history*** were found to be comparable and slightly better in accuracy or response time compared to ***baseline*** and ***history_ci***. We thus only **partially accept** it.
- ***(H3)*** *In terms of user confidence* ***history_ci*** $>$ (***history*** *and* ***ci***) $>$ ***baseline***. Participants only rated designs with near-history (***history*** and ***history_ci***) significantly higher than the ***baseline*** (by a small margin). We thus only **partially accept** it.

6 Discussion

The results of this experiment demonstrate that participants could perform comparisons accurately and decide quickly, even with the ***baseline*** design for progressive bar charts, surprisingly. Progressive bar charts with confidence intervals (***ci***) and with near-history (***history***) facilitated earlier responses and participants felt more confident using the near-history designs (***history*** and ***history_ci***), although the differences between the four designs were not always significant and rather small. We now discuss some interpretations of our findings, generalizability of our work, and possible future work.

6.1 Understanding Errors

This user study of progressive bar charts was aimed at assessing how well users performed on realistic data but not at explaining the causes of their decision errors. On average, participants responded after 20 s. Would their accuracy have been different if they only saw the snapshot of the progressive bar chart at this point i.e., without seeing the progression? Without comparison to a static or a one-shot approximate bar chart, we cannot evaluate how much the error rates are explained by factors related to the progression such as the *illusion bias* or *uncertainty bias*, or if they are due to other factors that are not related to the progression, such as misreading confidence intervals.

Inspecting the verbal explanations for confidence ratings, it seems some participants were not used to confidence intervals e.g., reporting "I am not 100% comfortable with confidence intervals, [***ci***] required more time to analyze and being confident with my answer" or faced difficulty in decision making—"When is the confidence interval small enough for me to trust the size of the bars?" Overall, 37% of participants declared limited statistical literacy and about a third scored ***baseline*** higher than ***ci*** in the confidence rating questionnaire. Since we did not collect a confidence rating per answer, we cannot analyze the relationship between perceived and displayed uncertainty (e.g., in the case of ***ci*** and ***history_ci***). In the comments on their understanding of confidence intervals, many participants reported difficulty *using* it, stating that the error bars "look complicated just by looking at it", "[are] kind of misleading", "got [them] very distracted, [...] especially because they changed sizes way too often" and that "it is like one more data to take into account and it deconcentrates [them]". These results could relate to known inference issues in error bars (binary interpretation [11]) or loss of trust due to fluctuations (*ambiguity bias* [6]). It would thus be worth investigating the differences between progressive bar charts and one-shot approximate bar charts, and also how user confidence is affected by the displayed uncertainty.

6.2 Breakdown by Human Performance

Participants rated ***ci***, ***history***, and ***history_ci*** higher than the ***baseline*** on the confidence by the end of the study, although they did not always perform largely better with them during the study. This difference hints toward the higher learning curve of near-history designs or unfamiliarity with confidence intervals. Still, some participants reported the number of visual elements in the near-history designs, and even in ***ci*** design in a few cases, to be overwhelming, even by the end of the study. To explore the relationship between confidence and performance, we look at a breakdown of confidence ratings according to performance, shown in Fig. 9. We split participants into four groups according to their average accuracy and response time compared to the median accuracy and response time of the experiment. We present this analysis as exploratory due to the small number of participants (≈ 20) and possible variations in the trial data of each group.

We can assume that the participants with the earliest responses (rightmost column) are relying on the cues provided in the visualization more than the others, who generally wait more to confirm their decision. Accordingly, the performance of these participants is more likely to highlight the effectiveness of each design. Further, participants with the earliest responses and the highest accuracy can be said to have a relatively better understanding of the study and the features of the different designs. Thus, participants in this category (bottom-right

cell in Fig. 9) might give a measure of how good users can get with more practice and exposure. From the low to high accuracy categories (middle to bottom row in Fig. 9), there is a clear drop in confidence in the ***baseline*** design, and a slight increase in confidence in the ***history_ci*** and ***history*** designs relative to the other designs. This hints toward the potential of near-history for more experienced users.

Finally, we look at the discrepancy between the confidence and performance results for ***history_ci***: it is rated among the highest in confidence rating while leading to the lowest accuracy overall. Participants with the lowest accuracy (middle row in Fig. 9) make the most decision errors, i.e., have the lowest accuracy, using ***history_ci*** while the rest of the participants achieve comparable accuracy with all designs. It may imply that ***history_ci*** is a comparatively worse option than ***history*** or ***ci*** for users with less experience.

6.3 Generalizability

Our study represents a simplified version of a progressive visual analysis scenario, with two tasks to perform on a 4-bar bar chart without distracting elements. Accuracy and response time could suffer in the real-life scenario having a busier interface with multiple components that could distract users from focusing on only one or two progressions and more bars on the screen; the near-history representation could become more effective because it requires less recall. We designed the experiment with two simultaneous questions concerning two pairs of bars, interleaved for half the trials, so the task is less likely to be done by memorizing previous values of one bar and is more realistic. As the number of bars increases, progressive bar charts would require interaction to ease comparison between bars. New experiments could be done to verify if our results still hold when progressive systems are more widespread.

Progressive visualization might take hours to complete while our experiment only runs for 2 min. Our experiment setting provides only a small time benefit to participants as they could afford to wait till the end for securing their decision. Making decisions early is even more important for longer progressions to save more time; shorter progressions would not need progressiveness at all. Consequently, when users necessarily have to decide early, the scope for error, and thus the efficacy of near-history could increase. Our experiment uses relatively small sequences (10k data points) over a short but realistic time (120 s). In a real setting, a progressive system could easily download 100k table rows per second from a remote server, leading to faster convergence. Only data produced through complex computations or very slow data channels could lead to the scenario in our experiment.

The three datasets in our experiment represent commonly found and realistic data distributions and three levels of complexity for the decision task. We expect these datasets to cover both the easiest and most difficult, yet realistic, convergence behaviors. The Pareto nature of *Power law* distributions (80/20 rule) makes it more likely than for *Normal* distributions to observe fluctuations in the early estimates. The ***wander join*** dataset uses real-world data which we found contains more outliers on average than the synthetic datasets, with outliers defined as values falling 1.5×IQR below the first quartile or 1.5×IQR above the third quartile, where IQR is the interquartile range. Our experiment relies on randomly ordered data whereas real-life datasets can come in a particular order e.g., chronological. Progressive computations are sensitive to data order and progressive systems do their best to operate on data as shuffled as possible [10]. The *Wander Join* algorithm [32] simulates random order and stationarity.

Although value comparison is a low-level task, it is very common and the basis for more complex compound tasks [17]. Future work is needed to validate other visualization tasks in a progressive setting. With respect to generalizability to other chart types like scatter plots, we first need to study how and when, near-history traces can be employed meaningfully, which is another avenue of future work. Overall, we believe our experiment is ecologically valid for dataset complexity, visual task, and chart type. For realistic progressive system interfaces and the progression duration, the differences between the efficacy of different bar charts studied in our work will become clearer.

6.4 Improvements with Interactions

Part of the difficulty of decision-making with progressive visualization is in making precise judgments under (dynamic) uncertainty. On bar charts, confidence intervals facilitate bar-to-value and bar-to-bar comparison. Pairwise comparison could be eased by highlighting significant differences upon hovering on a bar (or mini-bar). Interactions could be used to improve scalability to more bars and support other tasks. Showing the mini-bars of the near-history design upon interaction only can improve its scalability while retaining the ability to assess the convergence of an estimate. The four visual annotations of Ferreira et al. [17] designed for bar charts with confidence intervals could be integrated into the four proposed progressive bar chart designs to support more comparison tasks (bar-to-bars, bar-to-value, bar-to-range), and extrema identification. Visually, these techniques could integrate easily as they use color scales to encode the probability relevant to the task, e.g., the probability of being the maximum estimate, or of being over/under another estimate. We see their computational adaption to the progressive setting as a great avenue for future work.

6.5 Assessing Performance with Decision Procedures

In this work we introduce the use of decision procedures, inspired by sequential tests, to assess how efficient participants are for our chosen comparison task. Decision procedures are merely used as a means to estimate the optimal decision times for our tasks at a given accuracy threshold, i.e., to obtain a reference for participants' response times. We see two main methodology limitations to this comparison that we believe are also promising directions for future work. The first is to identify a valuable comparison procedure. While searching for a baseline for our experiment results, we could not find an optimal non-parametric statistical test designed for sequential testing and suited to our comparison task. All existing tests make assumptions about the data distribution and their parameters whereas our humans were shown data with no prior information about their distribution. The initial version of the GLRT we devised with an expert turned out to be too conservative in deciding on our ***power law*** examples. We eventually opted for two well-performing decision procedures and make all experiment data available to facilitate future analyses with more optimal procedures. The second limitation is the measure of response time. Without any existing model of human decision-making in the progressive context, we cannot determine how much of the measured response time is taken by cognition efforts or the action to click on the answer buttons. Indeed, in our experiment design, the view is updated every second with new data, even without an explicit decision from the participants to continue. As a result, the measured response time of our participants includes response delays that we cannot measure with our experiment. For this reason, we refrain from presenting the found time differences as ratios.

7 Conclusion

In summary, we conducted an empirical study to evaluate how efficient progressive bar charts are for early-decision making and compare four alternative designs: standard bar chart with/without error bars for confidence intervals, and near-history with/without error bars. Our results provide evidence that humans can do accurate comparisons early in the progression using progressive bar charts, even for data with difficult convergence behavior. More experiments are needed to understand how this finding generalizes to other tasks and chart types.

In terms of performance and confidence, our results suggest that confidence intervals and near-history individually support users in deciding earlier and with higher confidence, more than the standard bar chart design and even more than the design with both near-history and confidence intervals. Although not all differences were significant in our controlled experiment, we believe that this trend would be accentuated for real-life scenarios and for trained users.

Acknowledgments

We thank the participants of our pilot study for their time, Emilie Kaufmann for her guidance with statistical tests, Pierre Dragicevic, Catherine Plaisant, and Emanuel Zgraggen for their feedback. This article is a result of the Dagstuhl seminar 18411.

REFERENCES

[1] C. Albers. The Problem with Unadjusted Multiple and Sequential Statistical Testing. *Nature Communications*, 10(1):1921, Dec. 2019. doi: 10.1038/s41467-019-09941-0

[2] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 111–117, 2005. doi: 10.1109/INFVIS.2005.1532136

[3] M. Angelini, G. Santucci, H. Schumann, and H.-J. Schulz. A Review and Characterization of Progressive Visual Analytics. *Informatics*, 5(3):31, 2018. doi: 10.3390/informatics5030031

[4] P. Armitage. *Sequential medical trials*. Wiley New York, 2d ed. ed., 1975.

[5] P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235–244, 1969.

[6] D. A. Asch, J. Baron, J. C. Hershey, H. Kunreuther, J. Meszaros, I. Ritov, and M. Spranca. Omission Bias and Pertussis Vaccination. *Medical Decision Making*, 14(2):118–123, 1994. doi: 10.1177/0272989X9401400204

[7] S. K. Badam, N. Elmqvist, and J.-D. Fekete. Steering the Craft: UI Elements and Visualizations for Supporting Progressive Visual Analytics. *Computer Graphics Forum*, 36(3):491–502, 2017. doi: 10.1111/cgf.13205

[8] G. Bonneau, H. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. Overview and State-of-the-Art of Uncertainty Visualization. In C. D. Hansen, M. Chen, C. R. Johnson, A. E. Kaufman, and H. Hagen, eds., *Scientific Visualization*, Mathematics and Visualization, pp. 3–27. Springer, 2014. doi: 10.1007/978-1-4471-6497-5_1

[9] S. C. Castro, P. S. Quinan, H. Hosseinpour, and L. Padilla. Examining effort in 1d uncertainty communication using individual differences in working memory and nasa-tlx. *IEEE Trans. Vis. Comput. Graphics*, 28(1):411–421, 2021.

[10] Y. Cheng, W. Zhao, and F. Rusu. Bi-Level Online Aggregation on Raw Data. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17. ACM, New York, NY, USA, 2017. doi: 10.1145/3085504.3085514

[11] M. Correll and M. Gleicher. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Trans. Vis. Comput. Graphics*, 20(12):2142–2151, 2014. doi: 10.1109/TVCG.2014.2346298

[12] M. Cowles and C. Davis. On the Origins of the .05 Level of Statistical Significance. *American Psychologist*, 37(5):553–558, May 1982.

[13] A. Crotty, A. Galakatos, E. Zgraggen, C. Binnig, and T. Kraska. Vizdom: interactive analytics through pen and touch. *Proc. VLDB Endow.*, 8(12):2024–2027, 2015. doi: 10.14778/2824032.2824127

[14] G. Cumming and S. Finch. Inference by Eye: Confidence Intervals and How to Read Pictures of Data. *American Psychologist*, pp. 170–180, 2005. doi: 10.1037/0003-066X.60.2.170

[15] H. Dyrssen and E. Ekström. Sequential testing of a wiener process with costly observations. *Sequential Analysis*, 37(1):47–58, 2018. doi: 10.1080/07474946.2018.1427973

[16] B. Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

[17] N. Ferreira, D. Fisher, and A. C. König. Sample-oriented task-driven visualizations: allowing users to make better, more confident decisions. In M. Jones, P. A. Palanque, A. Schmidt, and T. Grossman, eds., *Proc. SIGCHI Conf. Human Factors Comp. Sys.*, pp. 571–580. ACM, 2014. doi: 10.1145/2556288.2557131

[18] D. Fisher. Incremental, approximate database queries and uncertainty for exploratory visualization. In *2011 IEEE Symposium on Large Data Analysis and Visualization*, pp. 73–80. IEEE, Providence, RI, USA, Oct. 2011. doi: 10.1109/LDAV.2011.6092320

[19] D. Fisher, I. Popov, S. Drucker, and M. Schraefel. Trust me, I'm partially right: Incremental visualization lets analysts explore large datasets faster. In *Proc. SIGCHI Conf. Human Factors Comp. Sys.*, pp. 1673–1682. ACM, New York, NY, USA, 2012.

[20] M. Glueck, A. Khan, and D. J. Wigdor. Dive in! Enabling Progressive Loading for Real-Time Navigation of Data Visualizations. In *Proc. SIGCHI Conf. Human Factors Comp. Sys.*, CHI '14, pp. 561–570. ACM, New York, NY, USA, 2014. doi: 10.1145/2556288.2557195

[21] J. M. Hellerstein, P. J. Haas, and H. J. Wang. Online Aggregation. In *Proc. SIGMOD*, SIGMOD '97, pp. 171–182. ACM, New York, NY, USA, 1997. doi: 10.1145/253260.253291

[22] J. L. Hintze and R. D. Nelson. Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, 52(2):181–184, May 1998.

[23] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[24] K. Hu, N. Gaikwad, M. Hulsebos, M. Bakker, E. Zgraggen, C. Hidalgo, T. Kraska, G. Li, A. Satyanarayan, and Çağatay Demiralp. VizNet: Towards A Large-Scale Visualization Learning and Benchmarking Repository. In *ACM Human Factors in Computing Systems (CHI)*, p. 1–12. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300892

[25] J. Hullman. Why Authors Don't Visualize Uncertainty. *IEEE Trans. Vis. Comput. Graphics*, 26(01):130–139, Jan. 2020. doi: 10.1109/TVCG.2019.2934287

[26] J. Hullman, E. Adar, and P. Shah. Benefitting InfoVis with Visual Difficulties. *IEEE Trans. Vis. Comput. Graphics*, 17(12):2213–2222, 2011. doi: 10.1109/TVCG.2011.175

[27] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Trans. Vis. Comput. Graphics*, 25(1):903–913, 2019. doi: 10.1109/TVCG.2018.2864889

[28] C. Jackson. Displaying Uncertainty With Shading. *The American Statistician*, 62:340–347, 2008.

[29] D. R. Johnson and L. A. Borden. Participants at your fingertips: Using amazon's mechanical turk to increase student–faculty collaborative research. *Teaching of Psychology*, 39(4):245–251, 2012. doi: 10.1177/0098628312456615

[30] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Trans. Vis. Comput. Graphics*, 25(1):892–902, 2018.

[31] E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of A/B testing. In *Conference on Learning Theory*, pp. 461–481. PMLR, 2014.

[32] F. Li, B. Wu, K. Yi, and Z. Zhao. Wander Join: Online Aggregation via Random Walks. In *Proc. SIGMOD*, SIGMOD '16, pp. 615–629. ACM, New York, NY, USA, 2016. doi: 10.1145/2882903.2915235

[33] Z. Liu and J. Heer. The Effects of Interactive Latency on Exploratory Visual Analysis. *IEEE Trans. Vis. Comput. Graphics*, 20(12):2122–2131, 2014. doi: 10.1109/TVCG.2014.2346452

[34] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *Proc. SIGKDD Workshop on Human Computation*, HCOMP '09, pp. 77–85. ACM, New York, NY, USA, 2009. doi: 10.1145/1600150.1600175

[35] B. McInnis, D. Cosley, C. Nam, and G. Leshed. *Taking a HIT: Designing around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk*, pp. 2271–2282. ACM, New York, NY, USA, 2016. doi: 10.1145/2858036.2858539

[36] L. Micallef, H.-J. Schulz, M. Angelini, M. Aupetit, R. Chang, J. Kohlhammer, A. Perer, and G. Santucci. The Human User in Progressive Visual Analytics. In J. Johansson, F. Sadlo, and G. E. Marai, eds., *EuroVis 2019 - Short Papers*. The Eurographics Association, 2019. doi: 10.2312/evs.20191164

[37] H. Mohammed, Z. Wei, E. Wu, and R. Netravali. Continuous Prefetch for Interactive Data Applications. *Proc. VLDB Endow.*, 13(12):2297–2311, July 2020. doi: 10.14778/3407790.3407826

[38] D. Moritz, D. Fisher, B. Ding, and C. Wang. *Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data*, pp. 2904–2915. ACM, New York, NY, USA, 2017. doi: 10.1145/3025453.3025456

[39] T. Munzner. *Visualization Analysis and Design*. A K Peters Visualization Series. CRC Press, 2014.

[40] M. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, 2005. doi: 10.1080/00107510500052444

[41] J. Neyman and E. S. Pearson. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika*, 20A(1/2):175–240, 1928.

[42] L. Opperman and W. Ning. Sequential probability ratio test for skew normal distribution. *Communications in Statistics - Simulation and Computation*, 50(10):2823–2836, 2021. doi: 10.1080/03610918.2019.1614623

[43] L. Padilla, M. Kay, and J. Hullman. *Uncertainty Visualization*, pp. 1–18. John Wiley & Sons, Ltd, 2021. doi: 10.1002/9781118445112.stat08296

[44] M. Procopio, A. Mosca, C. E. Scheidegger, E. Wu, and R. Chang. Impact of Cognitive Biases on Progressive Visualization. *IEEE Trans. Vis. Comput. Graphics*, 2021. To appear. doi: 10.1109/TVCG.2021.3051013

[45] M. Procopio, C. Scheidegger, E. Wu, and R. Chang. Selective Wander Join: Fast Progressive Visualizations for Data Joins. *Informatics*, 6(1), 2019.

[46] A. Rapoport and A. Tversky. Choice behavior in an optional stopping task. *Organizational Behavior and Human Performance*, 5(2):105–120, 1970.

doi: 10.1016/0030-5073(70)90008-5

[47] W. J. Reed and B. D. Hughes. From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 66:067103, Dec. 2002. doi: 10.1103/PhysRevE.66.067103

[48] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. Effectiveness of Animation in Trend Visualization. *IEEE Trans. Vis. Comput. Graphics*, 14(6):1325–1332, Nov. 2008. doi: 10.1109/TVCG.2008.125

[49] B. Saket, A. Endert, and Ç. Demiralp. Task-Based Effectiveness of Basic Visualizations. *IEEE Trans. Vis. Comput. Graphics*, 25(7):2505–2512, 2019. doi: 10.1109/TVCG.2018.2829750

[50] B. Shneiderman. Response Time and Display Rate in Human Performance with Computers. *ACM Computing Surveys*, 16(3):265–285, 1984. doi: 10.1145/2514.2517

[51] M. Skeels, B. Lee, G. Smith, and G. Robertson. Revealing uncertainty for information visualization. In *Proc. AVI '08*, p. 376. ACM Press, Napoli, Italy, 2008. doi: 10.1145/1385569.1385637

[52] C. D. Stolper, A. Perer, and D. Gotz. Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics. *IEEE Trans. Vis. Comput. Graphics*, 20(12):1653–1662, 2014. doi: 10.1109/TVCG.2014.2346574

[53] C. Sunstein. Probability Neglect: Emotions, Worst Cases, and Law. *The Yale Law Journal*, 112, Nov. 2001. doi: 10.2139/ssrn.292149

[54] J. Talbot, V. Setlur, and A. Anand. Four Experiments on the Perception of Bar Charts. *IEEE Trans. Vis. Comput. Graphics*, 20(12):2152–2160, 2014. doi: 10.1109/TVCG.2014.2346320

[55] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945. doi: 10.1214/aoms/1177731118

[56] B. L. Welch. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, Jan. 1947. doi: 10.1093/biomet/34.1-2.28

[57] H. Wickham. ASA 2009 Data Expo. *Journal of Computational and Graphical Statistics*, 20(2):281–283, 2011.

[58] E. Zgraggen, A. Galakatos, A. Crotty, J.-D. Fekete, and T. Kraska. How Progressive Visualizations Affect Exploratory Analysis. *IEEE Trans. Vis. Comput. Graphics*, 23(8):1977–1987, 2017. doi: 10.1109/TVCG.2016.2607714

[59] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. In *Proc. SIGCHI Conf. Human Factors Comp. Sys.*, pp. 1–12. ACM, Apr. 2018. doi: 10.1145/3173574.3174053

[60] H. Zhao, H. Zhang, Y. Liu, Y. Zhang, and X. Zhang. Pattern discovery: A progressive visual analytic design to support categorical data analysis. *Journal of Visual Languages & Computing*, 43:42–49, 2017. doi: 10.1016/j.jvlc.2017.05.004