# Exploring the Effects of Aggregation Choices on Untrained Visualization Users' Generalizations from Data

F. Nguyen[1] X. Qiao[1] J. Heer[2] and J. Hullman[1]

[1]Northwestern University
[2]University of Washington

**Abstract**

*Visualization system designers must decide whether and how to aggregate data by default. Aggregating distributional information in a single summary mark like a mean or sum simplifies interpretation, but may lead untrained users to overlook distributional features. We ask, How are the conclusions drawn by untrained visualization users affected by aggregation strategy? We present two controlled experiments comparing generalizations of a population that untrained users made from visualizations that summarized either a 1000 record or 50 record sample with either single mean summary mark, a disaggregated view with one mark per observation, or a view overlaying a mean summary mark atop a disaggregated view. While we observe no reliable effect of aggregation strategy on generalization accuracy at either sample size, users of purely disaggregated views were slightly less confident in their generalizations on average than users whose views show a single mean summary mark, and less likely to engage in dichotomous thinking about effects as either present or absent. Comparing results from 1000 record to 50 record dataset, we see a considerably larger decrease in the number of generalizations produced and reported confidence in generalizations among viewers who saw disaggregated data relative to those who saw only mean summary marks.*

**CCS Concepts**

• ***Human-centered computing*** → *Visual analytics; Visualization design and evaluation methods;* • ***Computing methodologies*** → *Uncertainty quantification;*

## 1. Introduction

Summary visualizations are indispensable to exploratory data analysis (EDA), the process by which a user visually assesses data to characterize distributions and trends, identify discrepancies, and formulate generalizations or predictions about the larger population from which data are drawn. Designing an effective summary visualization for EDA can involve trade-offs, with different design choices facilitating comprehension in ways that may conflict. For example, representing each data point with a mark in a *disaggregated* view affords reasoning about variance and distribution and depicts sample size (Figure 1 top). On the other hand, presenting the data using an aggregation (e.g., mean aggregation, Figure 1 center), simplifies the display, reduces error in perceiving expected value, and scales to multivariate displays and large samples. Presenting a mean annotation atop a disaggregated view may appear to be the best of both worlds, as it renders both central tendency and variance visible, assuming users attend to all of the information in the display (Figure 1 bottom).

Systems for exploratory data analysis currently vary in default aggregations. For example, Tableau Software [Tab18] defaults to sum aggregating data, often resulting in visualizations that display only a single mark as in Figure 1 center. Visualizing the same data in Voyager [WMA*16] shows data disaggregated by default. We ask, What are the implications of such differences? While some research examines how well viewers can estimate central tendency from disaggregated data [CH17; SHGF16], little empirical work poses the more applied question of how different aggregation strategies impact the types of conclusions people draw from summary visualizations in order to inform practice.

Designers of exploratory visual analysis systems may assume that users will customize views to support their target inferences regardless of the default aggregation strategy used by the system. However, default settings may have a significant influence on novice users of software, because they may not be aware of the need to change from a default or how to do so [SK06]. As data analysis and visualization tools reach a broader audience of users, from journalists [Tea18a] to students across a variety of disciplines [Sch13], understanding the impacts of aggregation choices is increasingly important for making sure inferences from visualized data are sound.

We contribute the results of two controlled experiments comparing untrained visualization users' generalizations—conclusions about a population drawn from a sample—made from three visual aggregation strategies: disaggregating the data, mean aggregation of the data, and disaggregating the data but superimposing the mean aggregation (Figure 1). Participants in each of our pre-registered within-subjects experiments viewed 15 visualizations representing a mix of aggregations strategies, with each visualization depicting a data sample drawn from a known ground truth population. For each visualization they saw, participants had the option of recording any "reliable generalizations", or statements they believed held for the the population from which the data was sampled, for example "As the age of visitor increases, the purchase amount also increases". For each generalization they made, participants also provided a numeric confidence rating.

In our first pre-registered experiment, participants examined a relatively large data sample consisting of 1000 records drawn from a population describing website visitation. While we observe no reliable effect of aggregation strategy on the average accuracy of generalizations, the number of generalizations reported and the confidence that participants express in their generalizations varied by aggregation condition. In particular, relative to viewing disaggregated data, viewing mean aggregated data led participants to record fewer generalizations, about which they were more confident. Additionally, we find in an exploratory analysis that participants who viewed disaggregated data, with or without a mean summary, were only 0.43x to 0.57x as likely to exhibit dichotomous thinking about effects (i.e., "there is a difference" or "there is not a difference" without any mention of effect size) on average.

In a second pre-registered experiment, we assess the robustness of our results to changing the sample size of the dataset. We again see no difference in accuracy based on aggregation strategy. However, in comparing the number of generalizations reported for the small sample to that reported for the large sample, we see a much bigger drop in the number of generalizations drawn from the small sample from disaggregated views than mean aggregated views, as we would hope to see if participants are being sensitive to how a smaller sample size makes conclusions drawn from the data less reliable. We also observe slightly larger effects on confidence and dichotomous thinking in the same direction as those observed in Experiment 1, where viewing disaggregated data reduced both relative to mean aggregated data.

Our results confirm that aggregation strategies impact generalizations drawn by untrained users, and provide insight into how sensitive they are to the informativeness (i.e., sample size) of data given different strategies. When aggregation is used as a default without also plotting disaggregated data, untrained users may engage in more superficial analyses of differences in data. We describe implications of these results for visualization system designers and how future work might employ alternative study designs to evaluate possible implications of aggregation strategies for exploratory data analysis.

## 2. Related Work

### 2.1. Visual Summary Strategies

While aggregation has well-studied implications for statistical analyses (e.g., [CA76]), the impacts of aggregation strategies on people's conclusions from visual analysis of data have not been thoroughly explored. Gschwandtner et al. compare 6 different visual summaries of aggregated univariate data [GBFM16] across tasks. Their results suggest that success of visual encoding correlates to task, although they fail to compare disaggregated views of data. Sarikaya et al. describe visual summaries in visual analytics tools through a quantitative content analysis of 180 publications describing visual analysis systems [SGS18]. Their results suggest a strong correlation between summarization methods (such as aggregation) and intended tasks. Aggregation appeared in 74% of the summaries they coded, including all presentation-oriented summaries and 9 out of 10 univariate summaries, and typically corresponded to intended tasks of characterizing the entire data set through specific measures. They suggest critically examining the task specificity versus flexibility trade-off between aggregation and other summarization approaches like showing all the data. However, it is difficult to define any single common task for the aggregation defaults used in visualization systems. We contribute two controlled experiments aimed at assessing how the quantity, nature, and accuracy of generalizations that system users make about the data can vary as a result of aggregation strategy as a step toward better understanding how naturalistic inferences may be affected by such defaults.

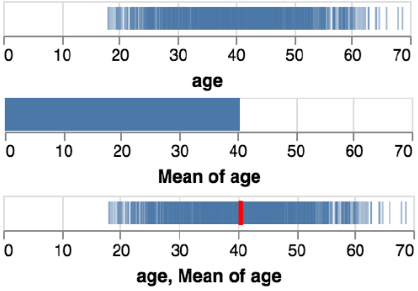


**Figure 1:** *E1 stimuli depicting a sample size of n=1000. Shown are univariate visualizations demonstrating disaggregated data (top), mean aggregated data (center, encoded using a bar rather than line mark), and disaggregated data with mean (bottom). All stimuli are included in the supplementary material.*

Some researchers have suggested *ensemble perception* as an alternative to aggregation, where a person perceives distributional properties from groups of visual items spread over space or time [CG15; CH17; HRA15; KNKH18; SHGF16]. Accumulating evidence suggests that the visual system is capable of quickly, automatically, and accurately extracting mental representations of ensembles [AO07; Alv11; HW09; HZ84; HB15; LKW16; MPOW17]. Visualization researchers have examined how well people can infer statistics from encodings that vary in their level of aggregation for particular types of data, like time series [ARH12; ACG14], multiclass scatterplots [GCNF13], or hierarchical data sets [EF10].

Correll and Gleicher [CG15] propose that ensemble processing may have benefits beyond flexibility of inferences. They suggest that "implicit uncertainty visualization," in which designers forego the addition of summary marks like mean annotations in favor of presenting ensembles of data points, may result in viewers having greater trust in their internal representation of uncertainty because they constructed it themselves. If this is true, we would expect viewers to be more confident in those generalizations that they report from disaggregated views. We elicit confidence for each generalization that a person provides in our study. This allows us to assess whether confidence varies reliably based on whether the visualization employs ensemble processing or not.

### 2.2. Aggregation Strategies for Simple Datasets

We are interested in how aggregation strategies impact the generalizations that untrained users draw as they examine visualizations. Common strategies for univariate and multivariate data include:

1. Disaggregating data by default (Figure 1, top).
2. Presenting classed (i.e., binned) data, where the number of bins is less than the number of data points (e.g., univariate or multivariate histogram).
3. Using a single summary mark to represent a chosen aggregation (such as a mean or sum) (Figure 1 center).
4. Presenting multiple measures as summary marks, such as by conveying central tendency and variance in a bar chart with error bars or quartiles in a boxplot.
5. Visualizing a density function, such as a univariate density plot or multivariate visualization like a heatmap or continuous scatterplot [BW08].
6. Presenting some combination of disaggregated data and summary marks, such as a disaggregated view with an overlaid summary mark representing central tendency (Figure 1 bottom).

The above strategies can be distinguished based on whether they explicitly present central tendency (#3), variance (#1), or provide access to both (#4, #5, #6) (with the fidelity of the information in #3 depending on the number of bins chosen). Taking into account human ensemble processing capabilities, disaggregating data by default (#1) enables judgments of both central tendency and variance, though with the possibility of more error in estimates of central tendency relative to the other techniques.

Strategies #2, #3, #4, and #5 avoid presenting individual data points, and consequently better scale to very large data sets. Strategy #3 reduces distributions to a single point estimate, potentially simplifying visual judgements, albeit with error proportional to how well the chosen aggregation summarizes the distribution. This strategy may be chosen as a default in some existing visualization systems because it both scales well to large data and aligns with well-established preferences among people to avoid uncertainty information where possible, such as by using heuristics based on representativeness or central tendency [TK74]. However, when the sample size is not very large, omitting distributional information (such as sample size and variance) prevents users of a visualization from reasoning about effect size, or the difference between the means of two distributions taking into account the variance in each [Coe02]. Consequently, users cannot account for the reliability of the difference between two distributions, making inference error prone. Our study looks at the implications of presenting only an aggregated measure for untrained users' generalizations, as these users may not be aware of the risks of drawing conclusions without access to other distributional information.

Empirical evidence suggests that using summary marks to encode distributional information, as in #4 leads to misinterpretations of data, among both trained and untrained users [BFWC05; CG14; Neu12]. Strategies like #5 scale well to large data and implicitly convey central tendency while also providing a view of the distribution. However, existing visualization systems do not typically employ these methods by default, perhaps because they are considered too specialized for some users. We compare untrained users' inferences from examples of #1, #2, #3 and #6 as these currently appear in systems.

To understand the extent to which existing systems use aggregation defaults, we surveyed those used in Microsoft Power BI [Cor18] Tableau [Tab18], Tibco Spotfire [Inc18], and Voyager [WMA*16]. Power BI requires the user to choose a chart type, putting the choice of whether to aggregate in the user's hands. For univariate data, the user might see a bar, line, or dot chart. Tableau presents a bar showing sum aggregation, and Spotfire presents a histogram. Voyager defaults to a one dimensional scatterplot (i.e., strip plot, Figure 1) where each data point is encoded as an individual mark. For data consisting of a quantitative variable and a categorical variable, depending on the chart type chosen PowerBI presents a bar, line, or dot chart. Both Tableau and Spotfire present a bar chart with the category on the $x$-axis and the bar height representing a sum aggregation. Voyager presents a facetted strip plot (Figure 3). Given two quantitative variables, Excel and PowerBI again make suggestions depending on the chart type shown (e.g., a scatterplot, two series line chart, or grouped bar chart if the user selects scatterplot, line, and bar respectively). Tableau uses sum aggregation by default on both variables, resulting in a scatterplot with a single point. Voyager and Spotfire present scatterplots of disaggregated data. The disparity in aggregation defaults across these systems motivates our controlled comparison of how aggregation strategies impact novice users' inferences.

## 3. Experiment 1: Comparing Aggregation Strategies

To better understand how different aggregation strategies may impact what novice visualization users extrapolate from a sample to a population, we conducted a controlled online experiment on Amazon's Mechanical Turk [Ama18]. Our experiment asked participants to view and respond to visualizations that use different aggregation strategies to depict subsets of a data sample that we generated from a ground truth population. While an in-depth naturalistic study of exploratory data analysis [ZZZK18] optimizes for external validity, an online experiment provides the control over stimuli and statistical power needed to compare the effects of aggregation strategies on inferences.

### 3.1. Experiment Conditions & Research Questions

As described above, aggregate measures of central tendency or sum (#3 above) may be preferred by some visualization system developers because they simplify views and avoid overplotting. Our **aggregate strategy** condition uses mean aggregation, because we expect untrained users to be more familiar with reasoning about averages based on their common usage for summarizing data in the media, science, and other everyday applications [Gal95].

When it comes to presenting uncertainty or distributional information, the empirical evidence suggests that using disaggregated data (#1 above) rather than summary marks or plots (#3, #5 above) to convey distributional information is less likely to confuse users (particularly untrained users) [BFWC05; CG14; Neu12]. We therefore include a **disaggregated strategy** condition.

One way of enabling both estimates of central tendency and distribution for untrained users is to plot disaggregated data but to add a mark representing the mean (#6). Adding a mean mark may reduce error in mean estimates for some users, but may also affect the aspects of the view that the user focuses on, such that generalizations are more likely to describe central tendency. We include a **disaggregated with mean strategy** condition to examine this trade-off.

### 3.1.1. Unit of Analysis: Generalization

In contrast to prior work that examines how well users can do narrowly defined tasks like reading probabilities or judging central tendency with different visual representations of distributions (e.g., [CG14; FWM*18; HRA15; KKHM16; KNKH18]), we are interested in more naturalistic behaviors that might occur as a user views plots in an exploratory data analysis setting. Based on an understanding of visualization interpretation as statistical inference [Tuk77], we define a **generalization** as a statement about the sample that a viewer believes also describes the population from which the data is drawn.

As inferential statements, generalizations provide information about how visualization users think about and place confidence in data. We can assess generalizations from an accuracy perspective, as commonly used in visualization evaluations [LBI*12; IIS*14; HQC*19], asking whether different visual aggregation strategies lead to systematic differences in generalization accuracy. Recent work by Zgraggen et al. and Zhao et al. [ZZZK18; ZDZ*17] on strategies for addressing the multiple comparisons problem demonstrates a data generation process in which generalizations or "insights" produced during exploratory data analysis can be definitely labeled as correct or incorrect. When data shown to users are drawn from a known ground truth model, generalization accuracy can be determined parametrically where test assumptions are clearly met or non-parametrically through bootstrapping [EfIM82].

Assessing generalization accuracy requires that the user's subjective sense of the generalization's truth value is equivalent to a statistical notion of what it means for a statement to hold true (e.g., statistically significant at $\alpha$=0.05). In reality, even when prompted for "reliable" generalizations only, a user who is not accustomed to reasoning about subjective uncertainty may instead report observations about the sample alone, or hunches about the population for which their level of certainty is not easily comparable to a statistical definition of 95% certain [Hul16].

To allow for the fact that untrained users' estimates of how likely a statement is to describe the population may not perfectly align with a notion of reliability as Null Hypothesis Significance Testing (NHST) at $\alpha$=0.05, we elicit a participant's confidence in their generalization using a 100 point slider. Eliciting confidence allows us to evaluate whether different visual aggregation strategies produce systematically different levels of confidence, and how confidence changes with a smaller sample size (Experiment 2).

Another approach for assessing generalizations beyond simply labeling them true or false prioritizes *estimation* (e.g., [Cum14; HFKJ06]). In an estimation paradigm, the focus is on the magnitude and uncertainty of an effect, which are argued to be associated with more valid scientific inferences that use of significance testing. We can assess the extent to which a participant's generalizations exhibit estimation by coding when a generalization exhibits reasoning about effect magnitude (e.g., "Ad B very slightly increased mean purchases"). Though not about effects, we can also compare generalizations in how frequently they contain quantitative predictions about values of variables in the data (e.g., "It seems most people made 2-4 purchases on the site").

Finally, we expect that different visual aggregation strategies may lead users to emphasize different aspects of the data in their generalizations. As a descriptive analysis, we can look at the focus of the generalizations (i.e., whether they focus on central tendency, variance or distribution shape, etc.), taking into account the stimuli that the user viewed as necessary to help distinguish their intention.

## 3.2. Methods

We designed a within-subjects experiment consisting of 15 trials. In each trial, we presented participants with a visualization that used either mean aggregation, disaggregation or disaggregation plus mean aggregation. Participants had the option of recording one or more generalizations that could be made from the data.

### 3.2.1. Data Generation and Sample Size

To create the dataset from which each visualization stimuli was generated, we adapted the data generation method, domain and sample size used by Zgraggen et al. [ZZZK18]. We replicated their procedures with the exception of several small changes. We chose one of the two datasets they tested, (*online shopping*), which contains customer information from a fictional online website. The dataset consists of 12 attributes (4 quantitative, 3 nominal, 5 ordinal), including information like ages of customers, income, average number of purchases per month and average time spent on the site. We chose to use this dataset rather than Zgraggen et al.'s sleep dataset because results from a pilot study we conducted indicated that participants relied more on the data and less on their own intuitions and prior assumptions about the domain with the shopping dataset compared to the sleep dataset. We used the same synthetic dataset for each participant.

We embed ground truth labels in the dataset, which allows us to later code any generalization as correct or not. For an $n$-attribute dataset, we generate $n/2$ "true" relationships as correlated pairs. These $n/2$ pairs are chosen randomly from the set of variables and given a non-zero absolute correlation coefficient. While Zgraggen et al. counted any non-zero correlation as a ground truth correlation, we required that these correlations be greater than 0.4 to ensure that participants would not be "penalized" for failing to observe small correlations that might be difficult to notice in plots. We sampled data from bivariate normal random variables parameterized by these correlation coefficients. This process results in a dataset with 12 total attributes (4 quantitative, 3 nominal, 5 ordinal), and 6 correlated attribute pairs. We used the above process to produce a dataset containing 1000 records.

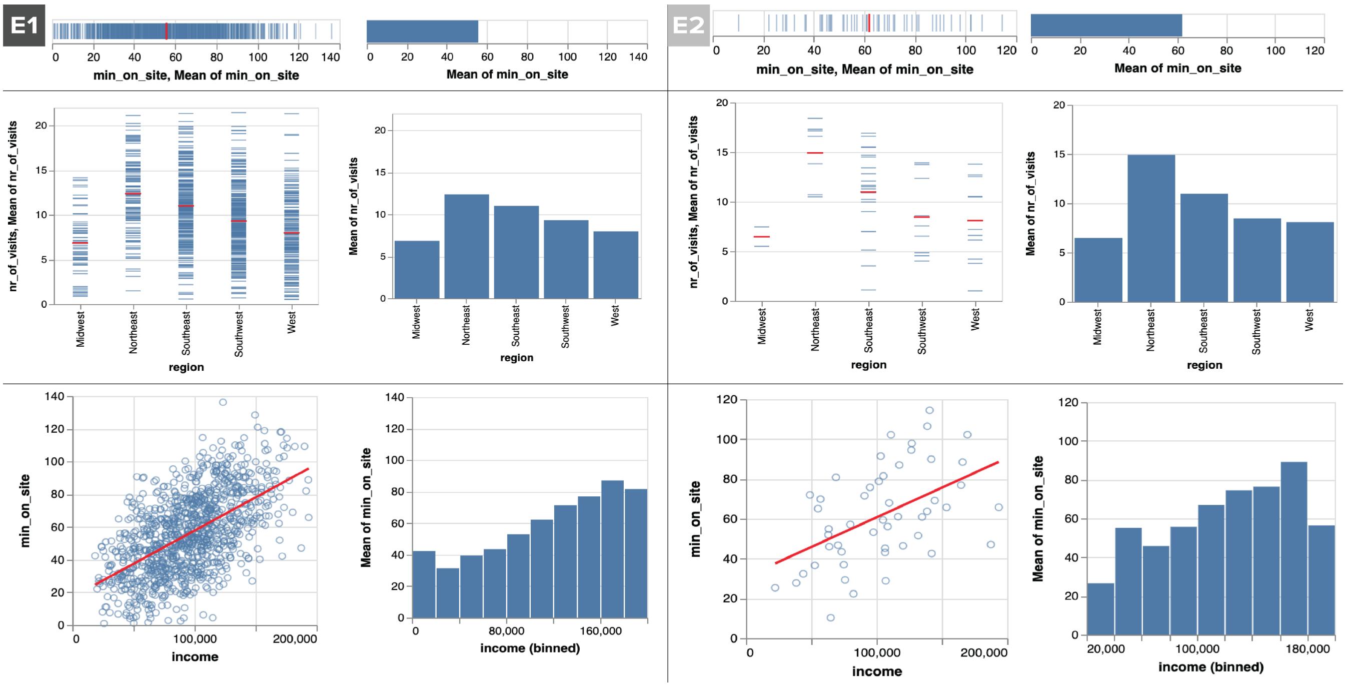


**Figure 2:** *Examples of stimuli used in Experiment 1 and Experiment 2. Shown are disaggregated data with mean (left) and mean aggregated data (right) for Experiment 1 (n=1000, left column) and Experiment 2 (n=50, right column). Disaggregated data stimuli simply have the mean mark in red removed for each chart. Each row depicts the data-type combination: univariate (top), one quantitative and one categorical attribute (middle), and two quantitative attributes (bottom). All stimuli are included in the supplementary material.*

### 3.2.2. Stimuli Generation

To include realistic data combinations in the visualizations we presented to participants, we formulated three data-type combinations. These data attribute combinations include *univariate*: where the chart displays a single quantitative data attribute (i.e. age); a chart that displays *one quantitative attribute and one nominal attribute* (i.e. purchase amount and education level); and *two quantitative attributes* (i.e. purchase amount and number of purchases).

To identify the particular subsets of variables that we would plot for each data-type combination we relied on a small formative study of exploratory data analysis with the online shopping dataset. Four participants from our university who were familiar with data analysis spent 70 minutes using the Voyager system [WMA*16] to generate and examine visualizations to support data analysis. Analyzing user-generated views, we identified attributes of the dataset that participants deemed worth exploring. In total we identified five data attribute subsets for each data-type combination (e.g. income, region and purchases, age and purchase amount), resulting in 15 total data-type combination stimuli. When by crossing these with the three aggregation strategies, we generated 45 total stimuli.

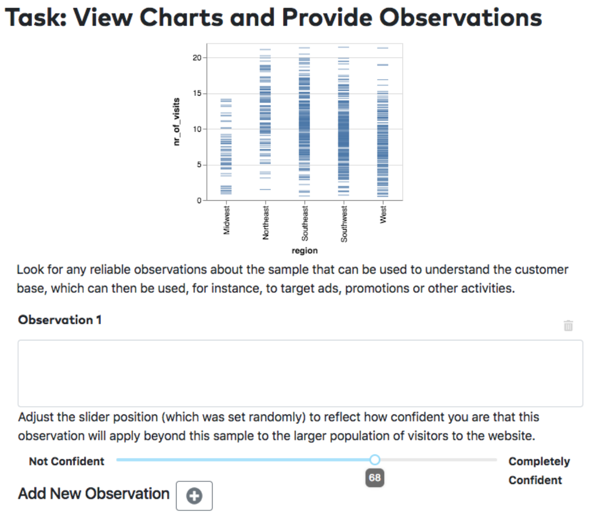
### 3.2.3. Procedure

**Preregistration:** We pre-registered our conditions, analysis and data collection criteria on the Open Science Framework (`https://osf.io/v87wd/`) before collecting data.

We conducted our experiment on Amazon Mechanical Turk (MTurk) as a single Human Intelligence Task (HIT) composed of 15 trials. Each participant saw at least 4 stimuli that applied each aggregation strategy. We ensured that the combined assignment of data-type combination (univariate, categorical and quantitative, quantitative and quantitative) and aggregation strategy (disaggregated, mean aggregated, and disaggregated with mean aggregation) were fully balanced across participants, and randomized the order of presentation of the trials for each participant.

Upon accepting the "Human Intelligence Task" (HIT), participants were re-directed to a web page containing instructions for the task. Participants were given a description of the dataset (including sample size) and asked to "note any observations you can make about the larger population of visitors to the website given the sampled data," and to "only include reliable observations that you would report to a superior like a manager if this was your job, or that you might make decisions based off of." For each trial, participants were asked to record all of their generalizations one at a time through text entry. They were free to record as many generalizations as they would like about each visualization stimuli presented to them, or no generalization at all (Figure 3).

Participants were asked to report "how confident they are that this observation will apply beyond this sample to the larger population of visitors to the website," using a slider ranging from 0 (labeled not confident) to 100 (labeled completely confident). To control for anchoring effects, the initial position of the slider handle was randomly assigned between participants. Upon completion



Task: View Charts and Provide Observations

Look for any reliable observations about the sample that can be used to understand the customer base, which can then be used, for instance, to target ads, promotions or other activities.

Observation 1

Adjust the slider position (which was set randomly) to reflect how confident you are that this observation will apply beyond this sample to the larger population of visitors to the website.

Not Confident 68 Completely Confident

Add New Observation ⊕

**Figure 3:** *A depiction of the task interface used in our studies (disaggregation by default stimuli shown). The participant uses buttons to add additional generalizations and is prompted to record their confidence with a slider after completing generalizations on the single page.*

of the study we ensured that there were no systematic differences in initial position by aggregation condition.

### 3.2.4. Participants

To determine the number of participants for the experiment, we conducted a pilot study with 9 participants. A simulation-based power analysis on accuracy differences observed in the pilot data suggested a target of 90 participants to achieve 80% power under $\alpha = 0.05$. We recruited workers from the United States with an approval rating of 97% or above. Participants were compensated with $5.00 for completing all the tasks. We ensured counterbalancing of the order of trials by running small batches, and after each batch applying our pre-registered exclusion criteria to the collected data. This was repeated until we had a total of 90 complete sets of generalizations from participants after exclusions. During the collection process, one participant was excluded due to the same confidence value on all trials, and six participants were excluded because the majority of their generalizations were not about the data they saw, all omissions per our preregistered exclusion criteria (see Preregistration). We used participant identifiers on MTurk to deny those who had previously completed the experiment to retake it.

We did not exclude participants based on statistical familiarity. In an exit survey, 80 participants self-reported their familiarity with statistics on a five point scale. Of these, four (5%) participants reported that they "used statistics often or were experts", 15 (19%) that they "took [statistics] in college and sometimes use it", 28 (35%) that they "took it in college but rarely use statistics" and 33 (41%) that they have "very little experience, and never took a course". These results suggest that the large majority of participants were not very familiar with statistics. Additionally, 80 participants reported their usage of "charts and graphs" on a five point scale. 6 (8%) participants reported that they "never" use charts, 33 (41%) that they "rarely (less than once a month)", 19 (24%) that they "Sometimes (1-5 times per month)", 17 (21%) "Often (1 - 5 times per week)", and 2 (3%) participants use charts "Very often (about everyday)". These results suggest that a majority of participants were not highly familiar with visualization.

## 3.3. Analysis Design

We analyze all generalizations reported by participants with a few exceptions. We apply pre-registered exclusion criteria and exclude generalizations that simply reiterated attributes of the visualization (e.g. "Chart shows the income and mean income"), generalizations based on prior or personal knowledge (e.g. "Graduate students are most likely budgeting money to pay off student loans"), or generalizations that misinterpreted the visualization (e.g. "There is a 3.5% dropout rate" when the participant viewed a visualization of number of purchases by education level).

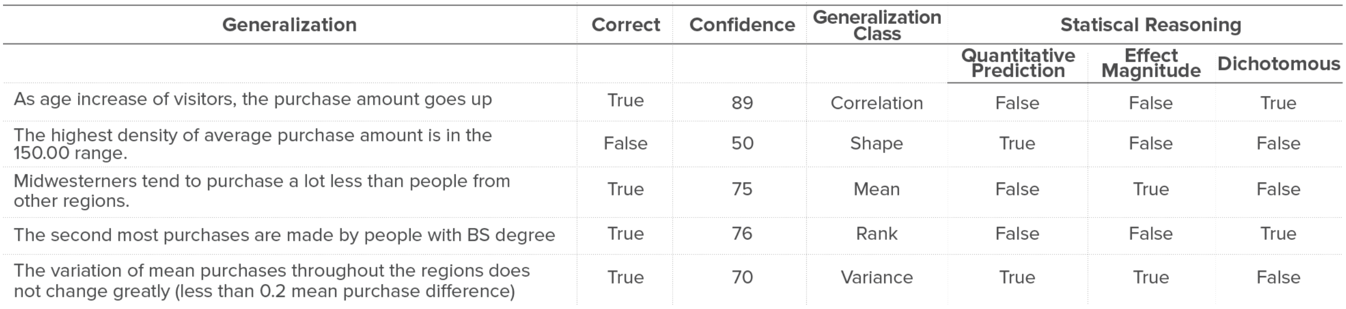
### 3.3.1. Coding of Generalizations

We coded participants' generalizations into one of five *generalization classes* modeled after Zgraggen et al.'s insight classes: *mean, variance, correlation, shape, ranking*. *Mean* refers to direct estimates or comparisons describing expectations for the population mean (e.g. "The average purchase is about $128."). *Variance* refers to direct estimates or comparisons of the distribution's variance (e.g. "The variation of purchases throughout the regions does not change greatly"). *Correlation* refers to statements describing a perceived relationship between two parameters (e.g. "Income doesn't significantly affect how many purchases a buyer makes."). *Shape* refers to statements about the shape or density of the distribution (e.g. "Most customers have a bachelors degree."). Finally, *rank* refers to statements where participants rank more than two levels of a categorical variable by a quantitative variable (e.g. "Blue is the most popular color.", "Red is the least popular color."). Three of the authors participated in coding, with the first and second author completing the majority. In cases where it was not obvious how to code a generalization, the three coders discussed the code until resolution was reached.

Per our preregistration we also coded when a generalization made a *quantitative prediction* (QP) or mentioned *effect magnitude* (EM). If a generalization contains a numeric statement about the value of a visualized variable (e.g. "Mean income is around 90,000", "Most buyers make between 3 and 4 purchases."), we mark it true for quantitative prediction. If the generalization mentions the magnitude of an observed effect of one parameter on the other (e.g. "Ad campaign had little impact on the 100-150k range of income customers."), we code it true for effect magnitude.

### 3.3.2. Evaluation of Generalization Accuracy

The accuracy of each generalization can be evaluated against the ground truth model [ZZZK18]. For example, "Age correlates with amount of spending" can be coded as true if the model that generated the dataset included an embedded correlation between these two variables. If two attributes are instead sampled from independent normal random variables, the generalization is false. However,



| Generalization | Correct | Confidence | Generalization Class | Statiscal Reasoning | | |
|---|---|---|---|---|---|---|
| | | | | Quantitative Prediction | Effect Magnitude | Dichotomous |
| As age increase of visitors, the purchase amount goes up | True | 89 | Correlation | False | False | True |
| The highest density of average purchase amount is in the 150.00 range. | False | 50 | Shape | True | False | False |
| Midwesterners tend to purchase a lot less than people from other regions. | True | 75 | Mean | False | True | False |
| The second most purchases are made by people with BS degree | True | 76 | Rank | False | False | True |
| The variation of mean purchases throughout the regions does not change greatly (less than 0.2 mean purchase difference) | True | 70 | Variance | True | True | False |

**Figure 4:** *Example generalizations and their respective correctness, reported confidence, and codes for generalization class, quantitative prediction, effect magnitude and dichotomous thinking.*

only some of the generalizations can be evaluated for ground truth by evaluating attribute relationships in the ground truth. We translate all other generalizations into testable queries and compare them against our 10M record sample dataset generated from the ground truth model (see Data Generation).

To convert an encoded generalization into a query, we first need to translate ambiguities in the generalization. For instance, a statement about an estimate of central tendency, "The amount per purchase seems to average about 150" can be interpreted as implying some uncertainty in the participant's estimate of the mean. Research in number sense suggests that people are sensitive to the ways in which rounded numbers imply uncertainty [BHG*11]. If the population mean is close, but not exactly 150, this would seem to align with what the participant stated. On the other hand, a generalization like "The average number minutes on the site is 55.76" indicates either a misunderstanding of the task (to report reliable generalizations that are expected to hold for the population) or an absurdly precise (i.e., confident) estimate of the population mean. When a generalization implies rounding, we follow Zgraggen et al.'s process to account for the potential uncertainty by adding an interval of 10% around the hypothesized central tendency. When a generalization is reported to decimal points or to values that don't imply rounding, then we do not account for the aforementioned 10% interval, as [BHG*11] shows that higher precision reflects less uncertainty about predictions. We expect that while noisy, this strategy is relatively conservative in concluding a generalization is wrong (e.g., a generalization that the mean of a variable is 150 is considered accurate so long as the interval [135, 165] contains the population mean, a generalization that the mean is 500 is accurate if the interval [450, 550] contains the population mean).

Other examples of resolving ambiguities around generalizations include interpreting textual descriptions of uncertainty as numerical thresholds, such as "Very few shoppers make purchases after 60 years old" (see Preregistration). We interpret phrases that imply a small number ("very few", "a small portion") conservatively as "20% or less of the sample." We interpret phrases that imply a strong majority ("most people make 3 purchases") as "80% or more of the sample." We follow the procedure of Zgraggen et al. [ZZZK18] and evaluated ambiguities conservatively, giving the participant the benefit of the doubt in assuming an implied level of precision. Once queries were specified, we ran them against the 10M record sample and recorded whether the result was true (correct) or false (incorrect).

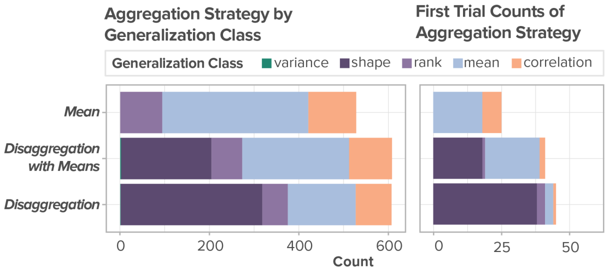


**Figure 5:** *The distribution across generalization types by aggregation strategy across all trials (left) and the first trial only (right).*

### 3.4. Results

90 full sets of generalizations that passed our pre-registered exclusion criteria were collected. We omitted from analysis seven sets of data from participants who completed the study but did not pass our pre-registered exclusion criteria. The average time to complete the task was 21.6 min (sd: 9.9 min), with an average trial completion time of 93.1 sec (sd: 107.5 sec).

#### 3.4.1. Generalization Frequency and Type

Participants submitted a total of 1,941 generalizations. Of these, we omitted 52 generalizations that did not appear to be about the data or that were gross misinterpretations of the visualization that was shown, as described in §3.3.1 (24 disaggregated, 19 disaggregated with means, 19 mean). We omitted 134 trials from analysis where the participant did not express any generalizations for that view (43 disaggregated, 43 disaggregated with means, 48 mean). This process left us with 1,743 trials (90%). Not counting trials for which no generalizations were produced, our omission rate is slightly lower than that of Zgraggen et al. [ZZZK18], who removed 6 out of 161 generalizations.

We observed the fewest generalizations in the mean aggregation condition with 528 (30.2%), with the two disaggregated con-

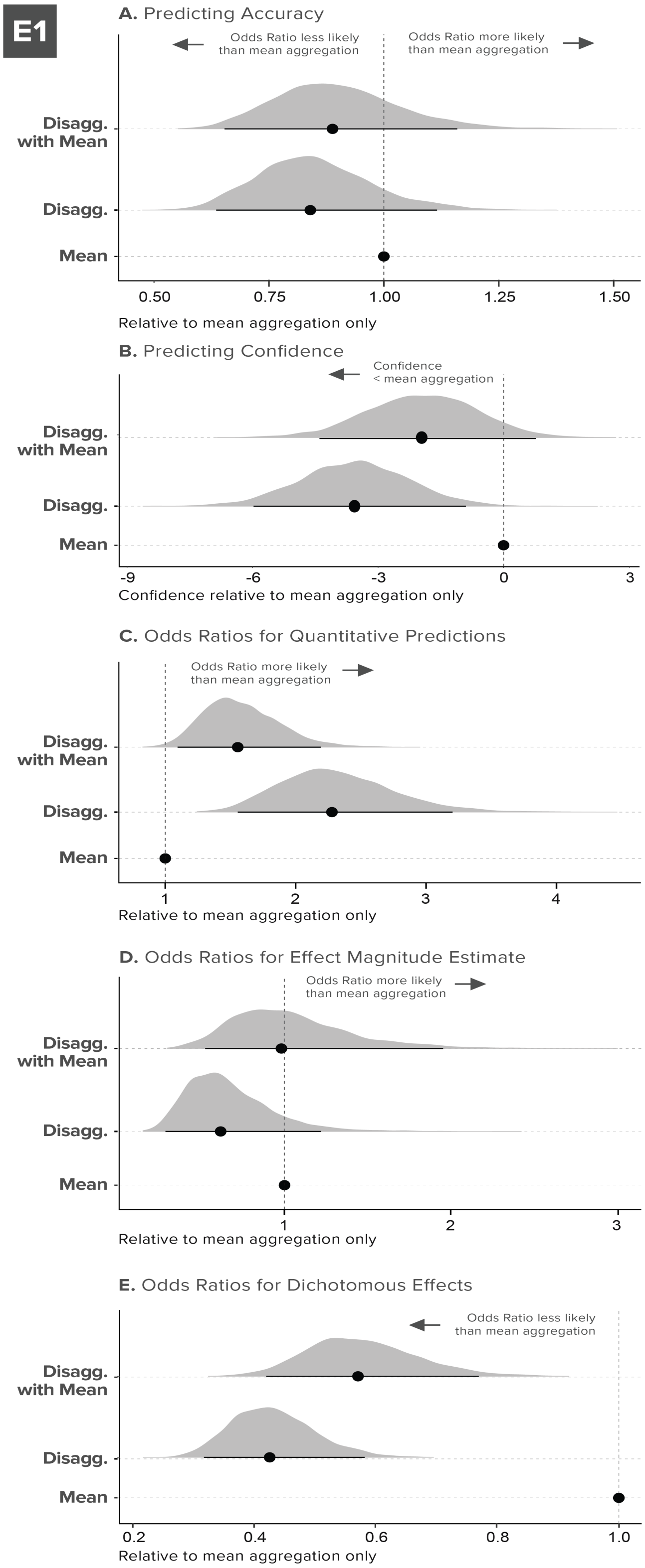


**Figure 6:** *Posterior estimates in E1 of effects of aggregation strategy on the probability that (a) a participant's generalization is correct, (b) a participant reports lower confidence, (c) a participant's generalization expresses a quantitative prediction, (d) a participant's generalization references effect magnitude and (e) a participant's generalization references a dichotomous effect. Effects in all models except confidence are expressed in terms of odds ratios relative to viewing a mean aggregation. Intercept and estimate effect of trial are available in Supplemental Material.*

ditions producing similar numbers (with mean: 608, 34.9%; without: 607, 34.8%). Across data-type combinations we observed the fewest generalizations for univariate data (482; 27.6%), followed by two quantitative variables (610; 35.0%) and one nominal, one quantitative data (651; 37.3%).

Of our total generalizations, 718 (41.2%) described a mean or difference in means, 519 (30.0%) described distribution shape, 283 (16.2%) described a correlation, 220 (12.6%) described a rank, and only 3 (0.01%) concerned variance (Figure 5). Breaking these percentages down by aggregation strategy, we see obvious expected patterns like more shape generalizations when data is disaggregated by default relative to mean aggregation alone (52.2% vs 17.8%; 33.1% for disaggregation with mean), less mean generalizations (25.0% vs 43.9%; 39.2% for disaggregation with mean), and less correlation generalizations (13.2% vs 20.3%; 15.7% for disaggregation with mean).

It is possible that our use of mean aggregation led workers to focus more on means even on trials where means were not plotted (disaggregated). To understand how much of an influence this may have had we analyzed the types of generalizations that participants generated on only the first trial. We observe the same ranking by frequency of generalization types, with roughly 16% fewer mean generalizations in the disaggregated condition (Figure 5 right), suggesting that viewing means in earlier trials may inflate the probability of mean generalizations in later trials. We therefore control for trial number in our models, described below.

### 3.4.2. Accuracy

Of the generalizations that we analyzed, 66.4% were accurate. The average participant was accurate for 67.4% of their generalizations, though this rate ranged considerably across participants (min: 31.6%, max: 100.0%, sd: 16.1%).

Overall accuracy rates were similar across aggregation strategies, with disaggregated views producing a rate of 65.5%, mean aggregated views producing a rate of 67.2%, and disaggregated views with mean annotation producing a rate of 66.7%.

To assess accuracy by aggregation strategy while controlling for the random effects of participant and dataset ID (which of the 15 datasets, with 5 of each data-type combination, the participant saw), we specified a pre-registered Bayesian hierarchical binomial model. Our model predicts the mean effect ($\beta$ coefficients) of disaggregation and disaggregation with mean annotation (represented by dummy variables) on the probability that a generalization is correct. Mean aggregation is the reference class to which coefficients refer. We also estimated the mean effect of trial number. We estimated varying intercepts for participant ID and dataset ID. We specified identical weakly regularizing Gaussian priors centered on 0 for each effect ($\beta$; standard deviation of 5). We specified half-Cauchy priors centered on 0 for the estimated variance for varying intercepts effects ($\sigma$; standard deviation of 1). Half-Cauchy distributions are Cauchy defined over positive real numbers only; Cauchy distributions are thick tailed and preferable to Gaussian distributions as a weakly regularizing prior [McE15]. We specified and ran the accuracy model and all other Bayesian models reported below using the *rethinking* package for R[McE16], which uses R Stan [Tea18b] for MCMC sampling.

We report results as the distribution of posterior odds ratio estimates for effects of disaggregation and disaggregation with mean (relative to mean aggregation, which is the reference group in the model). We report estimates for the model intercept, effect of trial number, and sigma estimates for this and all subsequent models in Supplemental Material. Rather than using p-values to judge which effects are reliable, we present 95% Credible Intervals on estimated coefficients (reported in text, henceforth CIs). To judge reliability of effects in a way that is analogous to looking for significance, a reader can look for whether these CIs include zero (indicating the possibility of no effect), or equivalently for the logistic models, the degree to which 95% CIs on the estimated odds ratios include one. For example, the CI shown in Figure 6(b), for the disaggregation condition, does not include zero though that for the disaggregation with mean condition does.

Figure 6(a) presents the results of our accuracy model. While both disaggregation and disaggregation with mean reduce accuracy slightly on average (i.e., generalizations are 0.84x and 0.89x as accurate) compared to mean aggregation, these effects are not reliable (95% CIs on odds ratios include 1: [0.64, 1.1] and 2: [0.68, 1.2] respectively).

To better understand whether accuracy rates are affected by the nature of the data, we calculated accuracy by data-type combination. The one nominal, one quantitative variable data resulted in the lowest accuracy rates for generalizations at 45.2% (95% CI: [43.2, 47.1]). The two quantitative variables and univariate datatypes produced much higher accuracy rates at 79.7% (95% CI: [77.9, 81.5]) and 82.9% (95% CI: [81.1, 84.6]) respectively. Examining generalizations for the more error prone nominal and quantitative combination showed that many participants tended to focus on pairwise differences between nominal values that were often not supported in the population dataset.

### 3.4.3. Confidence

Our prompt, following [ZZZK18], asked participants to only report "reliable generalizations that you would report to a superior like a manager if this was your job, or that you might make decisions based off of." While others have interpreted generalizations prompted in this way as being at least 95% confident [ZZZK18], we found that participants were at least 90% confident for only 429 (24.7%) generalizations. For a portion of generalizations (7.0%), participants provided confidence values of 0.

It is possible that participants were more liberal in reporting generalizations when they realized they could report confidence as well. Per our pregistration, we assessed confidence on only the first trial. The results were very similar to the confidence trends across all trials, with 27.9% reporting confidence $\geq 90$, and 5.4% reporting confidence of 0.

As we might hope, participants were less confident on average when their generalization was incorrect, albeit only slightly (between 1 and 5 points less confident out of 100). We observed the largest disparity in confidence across accurate versus inaccurate generalizations for views that only depicted a mean mark: 4.7 points less confident when they were incorrect vs 3.6 points less confident in the disaggregated view and 1.8 points less in the disaggregation view with means overlaid. Following our preregistration, we evaluated accuracy for generalizations filtered to only include confidence at some value $k$ or above, allowing us to examine whether confidence can be reliably thresholded across participants to identify more accurate generalizations. However, we see little change in accuracy levels when generalizations are thresholded by confidence, possibly because confidence values are subject to individual differences in how the participant assigns numeric values to different levels of subjective belief, and may reflect other contextual characteristics [Hul16].

We find some evidence that aggregation strategy has an effect, albeit small, on confidence. In the mean aggregation condition, participants reported an average confidence of 71.8% (95% CI: 70.2, 73.3). Participants reported an average confidence of 68.6% (95% CI: 67.0, 70.2) for the disaggregation with means strategy and 67.3% (95% CI: 65.0, 68.8) for the disaggregation strategy.

To investigate the effect of aggregation strategy on confidence while taking into account the study design, we use a pre-registered Bayesian hierarchical model while controlling for the random effects of participant ID and dataset ID. Our model estimate the mean effect ($\beta$ coefficients) of disaggregation and disaggregation with mean annotation (both represented by dummy codes) as well as trial number on a participant's reported confidence. We again apply weakly regularizing identical Gaussian priors centered on 0 for each effect ($\beta$; standard deviation of 5), and half Cauchy priors centered on 0 for the estimated variance for varying intercepts effects ($\sigma$; standard deviation of 50).

Figure 6(b) presents posterior mean estimates for effects of both disaggregation and disaggretation with mean relative to mean aggregation. On average viewing a disaggregated visualization reduced confidence by 3.6 points (95% CI [-6.2, -1.0 ]). Viewing a disaggregated with mean visualization reduced confidence on average by 1.9 points, though not reliably (95% CI [-4.4, 0.6]).

### 3.4.4. Reasoning about Quantities & Effect Magnitude

We coded 991 *quantitative prediction* generalizations mentioning numeric estimates of variable values, and 211 *effect magnitude* generalizations mentioning the size of a relationship or difference between variables. When comparing the rate of these codes by aggregation strategy, we see that in the mean aggregation condition the percentage of quantitative predictions is sightly lower: 50.0% of all mean aggregation generalizations were coded as quantitative prediction; 58.0% for disaggregation with means; 61.6% for disaggregation.

Overall, participants were much less likely to refer to effect magnitude than make quantitative predictions about single variables. We observe similar rates of effect magnitude generalizations when data is disaggregated by default: 3.0%, mean aggregated: 4.2%, and disaggregated with means: 4.4%.

To assess these differences while accounting for subject and dataset specific effects, we specified two (pre-registered) Bayesian hierarchical binomial models identical to the model we used for accuracy. However, this time we regressed the mean effects ($\beta$ coefficients) on the probability that a generalization mentions Effect Magnitude and makes a Quantitative Prediction, respectively. We used identical weakly regularizing priors for mean effects and standard deviations to those used in the accuracy model.

Figure 6(c) and Figure 6(d) present posterior odds ratio estimates for effects of disaggregation and disaggregation with mean relative to mean aggregation for effect magnitude and quantitative predictions.

We find that relative to using only mean aggregation, using a disaggregated view or a disaggregated view with mean reliably increases the rate of quantitative predictions by 2.3x and 1.6x, respectively (95% CIs: [1.6, 3.3], [1.1, 2.3]). More quantitative predictions in views that show disaggregated data is likely a result of the greater number of possible inferences available when raw data is plotted.

Using a disaggregated view made participants slightly less likely to reference effect magnitude on average, but not reliably (0.61x 95% CI: [0.30, 1.2]). Relative to using only mean aggregation, using disaggregated data with a mean annotations made no real difference (0.97x; 95% CI [0.48, 1.9]). This is not surprising given the low rates of effect magnitude generalizations across aggregation strategies.

In the process of coding generalizations for effect magnitude, we observed that many described effects as either present or absent without mentioning magnitude, such as “There is/is not a relationship between number of visits and education”. While statements that suggest there is no effect could be interpreted as effect magnitude (e.g., predictions that the effect is 0), discussions of the prevalence of *dichotomous thinking* among statistical reformers [HFKJ06] would suggest that these statements are examples of such black and white thinking about effects as simply present or absent.

To better understand if dichotomous thinking was more likely with one strategy than another, we conducted an exploratory analysis. We explicitly all generalizations that mentioned either an effect or the absence of an effect without any other description of magnitude (e.g., “There is/is not a difference in number of visits based on education.”). Rates of dichotomous generalizations were much higher than rates of effect magnitude references: 29.7% for disaggregation, 32.4% for disaggregation with mean, and 37.9% for mean aggregation. To assess whether these rates differed reliably by aggregation strategy while accounting for subject and dataset specific effects, we specify an hierarchical model identical to our model for effect magnitude where we instead predicted the proportion of dichotomous generalizations. We found that indeed, relative to mean aggregation both disaggregated and disaggregated with mean views reduced the probability of dichotomous statements, by 0.43x and 0.57x, respectively (95% CIs:[0.32, 0.58], [0.42, 0.77]). These results are exploratory, as we did not preregister the model.

### 3.5. Discussion of Results

Overall, we do not find evidence that accuracy rates differ by aggregation strategy. Viewing a disaggregated view without an overlaid mean did reduce confidence slightly, by an average of 3.6 points over mean aggregation alone. Disaggregating data, with or without an overlaid mean, increased how likely participants were to make quantitative statements about single variables, roughly doubling the rate on average relative to mean aggregation alone. Our exploratory analysis provides evidence suggesting that disaggregating data, with or without an overlaid mean, may reduce the probability that a participant engages in dichotomous thinking about effects by noting the presence or absence of a difference between variables without any mention of effect size. A preregistered followup is however necessary to confirm this, which we contribute in Experiment 2.

There are several reasons that may explain less confident generalizations with disaggregated views. In the absence of a mark summarizing the mean, the participant must work harder to estimate central tendency. This may result in a meta-cognitive realization of the difficulty associated with the task, engendering deeper processing as suggested by research in educational psychology [DOV11]. Similarly, participants who use disaggregated views may perceive generalizations about distribution or variance (which are not supported by mean aggregation alone) as more complex than generalizations about central tendency, and therefore feel less confident. We see few differences however in the confidence reported across different types of insights, suggesting the latter reason may be less explanatory.

## 4. Experiment 2: Small Sample Size

The relatively large size (n=1000) of the sample participants viewed in Experiment 1 may be behind the relatively consistent accuracy rates across aggregation strategies: even if one has access to only a mean to assess differences or make quantitative predictions, a large normally distributed random sample may support relatively accurate generalizations because it is likely to be representative of the population. In experiment 2, we investigate whether our observations of generalization number, accuracy, and confidence are robust to a small sample. We used an identical study design and data and stimuli generation procedure to compare the generalizations of participants when exposed to stimuli depicting a smaller sample size of 50 (see Figure 2 right).

### 4.1. Methods

With the exception of the following changes, the methods used in our second experiment were identical to those used in the first.

#### 4.1.1. Data Generation and Sample Size

We randomly sampled 50 records from the larger sample size of 1000. We used the same synthetic dataset for each participant, resulting in a dataset of 50 samples with 12 total attributes (4 quantitative, 3 nominal, 5 ordinal), and 6 correlated attribute pairs (see Data Generation).

**Preregistration:** We pre-registered our conditions, analysis (including the coding and Bayesian analysis of dichotomous generalizations) and data collection criteria on the Open Science Framework (https://osf.io/zx7we/) before collecting any data.

#### 4.1.2. Participants

We recruited 90 participants from MTurk, using the same criteria in Experiment 1. We applied the same exclusion criteria, until we had a total of 90 complete sets of generalizations from participants after exclusions. During the collection process, 4 participants were excluded because more than half their generalizations consisted of no generalizations and 8 were excluded because the majority of their generalizations were not about the data they saw, per our pre-registered exclusion criteria. We denied those who had previously completed the experiment from retaking it. We did not exclude participants based on statistical or chart familiarity. In an exit survey, all 90 participants self-reported their familiarity with statistics on a five point scale. Of these, three (3%) participants reported that they "used statistics often or were experts", six (7%) that they "took [statistics] in college and sometimes use it", 11 (11%) that they "took it in college but rarely use statistics" and 70 (78%) that they have "very little experience, and never took a course". All 90 participants reported their usage of "charts and graphs" on a five point scale. Eight (9%) participants reported that they "never" use charts, 34 (38%) that they "rarely (less than once a month)" use charts, 29 (32%) they use charts "Sometimes (1-5 times per month)", 16 (18%) "Often (1 - 5 times per week)", and 3 (3%) participants use charts "Very often (about everyday)".

## 4.2. Results

We obtained 90 full sets of generalizations that passed our pre-registered exclusion criteria. We omitted from analysis 12 participants who completed the study but did not pass the pre-registered criteria. The average time to complete the task was 24.1 min (sd: 12.5 min), with an average trial completion time of 120.9 sec (sd: 108.8 sec).

### 4.2.1. Generalization Frequency and Type

Participants submitted a total of 1,880 generalizations. Per our preregistration, we omitted 86 generalizations that did not appear to be about the data or were misinterpretations of the visualizations that were shown (24 disaggregated, 29 disaggregated with means, 33 mean), and 186 trials where the participant did not express any generalization (68 disaggregated, 60 disaggregated with means, 58 mean). This process left us with 1,608 generalizations (86%).

We observed the fewest generalizations in the mean aggregation condition with 500 (31.1%), with the two disaggregated conditions producing similar numbers (with mean: 558, 34.7%; without: 550, 34.2%). Relative to the number of generalizations produced in Experiment 1, where the sample size was 20x larger, participants in the disaggregated and disaggregated with mean conditions had slightly larger reductions in the amount of generalizations they produced in Experiment 2: participants who viewed disaggregated views produced 550 generalizations compared to 607 (9.4% less), those who viewed disaggregated with mean views produced 558 generalizations vs 608 (8.2% less), and those who viewed mean aggregated views produced 528 generalizations vs 500 (5.3% less). Across data-type combinations we observed the most generalizations for a one nominal, one quantitative data-type combination (646; 40.2%), followed by two quantitative variables(491; 30.5%) and univariate data (471; 29.3%).

Of our generalizations in Experiment 2, 648 (40.3%) described distribution shape, 420 (26.1%) described a mean or difference in means, 276 (17.2%) described a correlation, 262 (16.3%) described a rank, and only 2 (0.1%) concerned variance. As in Experiment 1, we unsurprisingly observe more shape generalizations when data is disaggregated by default (60.1% vs 43.9% for disaggregation with means vs 13.6% for mean aggregation). When data includes some mark for the mean, we observe more mean generalizations (10.3% for disaggregated vs 45.4% for mean aggregation vs 24.4% for disaggregation with means) and more rank generalizations (13.8% for disaggregated vs 24.2% for mean aggregation vs 11.6% for disaggregation with means). When analyzing the types of generalizations participants generated on only the first trial, we observe the same ranking by frequency of generalization types.

### 4.2.2. Accuracy

Of the generalizations we analyzed, a lower proportion of 51.6% were accurate compared to 66.4% in Experiment 1, suggesting that not all participants were adequately sensitive to the relationship between sample size and estimate reliability. An average participant was accurate for 52.4% of their generalizations, but this rate varied greatly across participants (min: 12.9%, max: 85.7%, sd: 14.3%).

Accuracy rates were similar across aggregation strategies: disaggregated views produced a rate of 49.5%, mean aggregated views producing a rate of 49.8%, and disaggregated with means overlaid produced a rate of 53.2%.

We specified a Bayesian hierarchical binomial model identical to that of Experiment 1 to assess accuracy by aggregation strategy. We report results as the distribution of posterior odds ratio estimates. Figure 7(a) presents the results of our accuracy model. While we observed lower accuracy rates overall for Experiment 2, we again find no reliable effect of aggregation strategy on accuracy.

In addition, we calculated accuracy by data-type combination. We see the same pattern in Experiment 1: the one nominal, one quantitative variable data-type combination produced a much lower accuracy rate of 39.8% (95% CI: [38.1%, 41.5%]). Views that display 2 quantitative variables produced a rate of 65.6% (95% CI: [63.3%, 68.0%]), and univariate data-types produced a rate of 57.7% (95% CI: [55.4%, 60.0%]).

### 4.2.3. Confidence

We found that participants were at least 90% confident for 517 (32.1%) generalizations, and in 11 (0.7%) generalizations reported a confidence of 0, similar rates to the results of Experiment 1.

To examine the effect of aggregation strategy on confidence, we use the same pre-registered Bayesian hierarchical model for confidence as in Experiment 1, controlling for the random effects of participant ID and dataset ID. Figure 7(b) presents posterior mean estimates for effects of disaggregation and disaggregation with mean conditions relative to the mean aggregation condition. On average, viewing a stimulus in the disaggregated condition reduced confidence by 5.9 points (95% CI [-8.4, -3.3]. Viewing a stimulus disaggregated with the mean overlaid reduced confidence on average by 0.7 points (95% CI [-1.9, 3.2]), though not reliably. These results reflect a similar confidence pattern found in Experiment 1, where the participants expressed less confidence when the mean mark is not annotated.

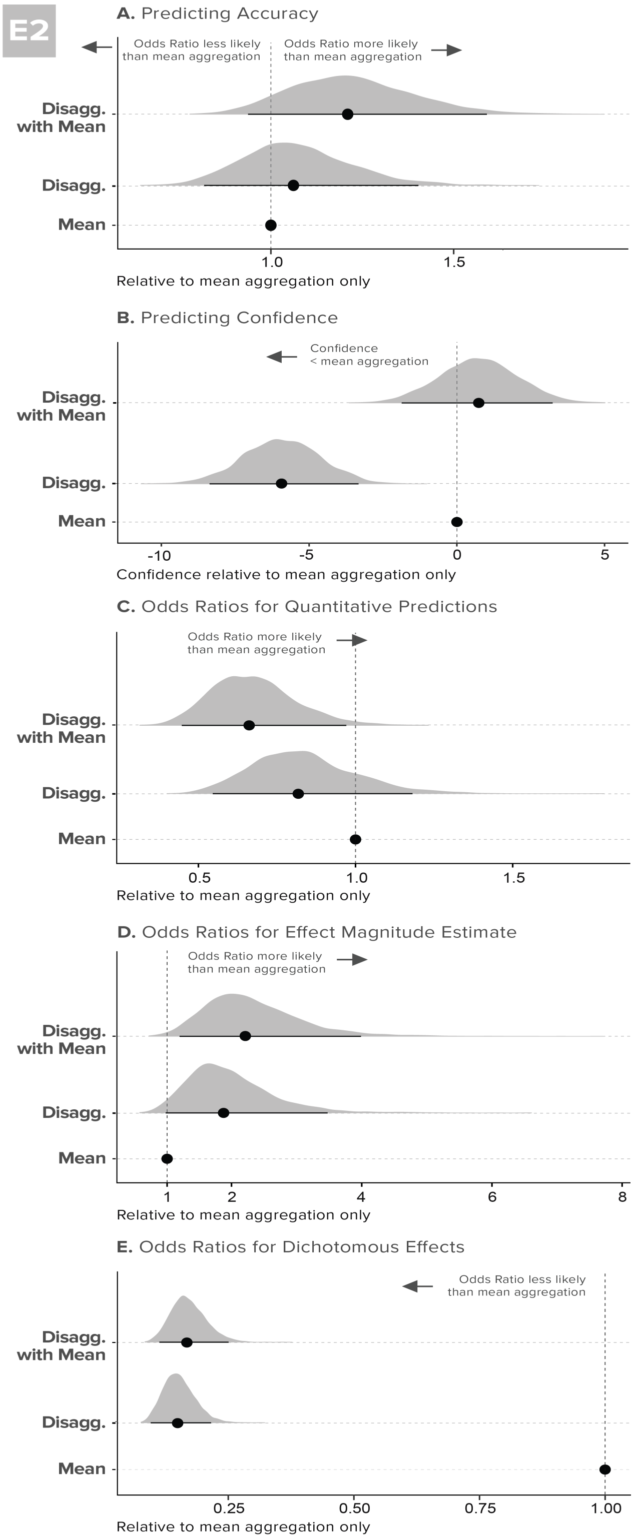


**Figure 7:** *Posterior estimates in E2 of effects of aggregation strategy on the probability that (a) a participant's generalization is correct, (b) a participant reports lower confidence, (c) a participant's generalization expresses a quantitative prediction, (d) a participant's generalization references effect magnitude and (e) a participant's generalization references a dichotomous effect. Effects in all models except confidence are expressed in terms of odds ratios relative to viewing a mean aggregation. Intercept and estimate effect of trial are available in Supplemental Material.*

Following the results of Experiment 1, we conduct a pre-registered analysis of dichotomous thinking. Relative to Experiment 1, we observe slightly higher rates of dichotomous generalizations: 37.9% for disaggregation, 38.0% for disaggregation with mean, and 55.2% with mean aggregation. We assessed the reliability of these differences in aggregation strategy by specifying a hierarchical model identical to our dichotomous model for Experiment 1 (Figure 7(e)). We found that relative to mean aggregation, views where data is disaggregated reliably reduced the probability of dichotomous statements: disaggregation views were only 0.17x (95% CI: [0.11, 0.24]) as likely to lead to dichotomous generalizations and disaggregation with mean views were 0.18x (95% CI: [0.12, 0.26]) as likely. These effects are consistent with, but larger than, the corresponding reductions by 0.43x and 0.57x in Experiment 1, suggesting that particularly when samples are small, plotting disaggregated data can be an important way to signal caution to untrained viewers.

## 5. Discussion

We set out to explore what sorts of generalization behaviors might be affected by the use of different aggregation strategies among untrained users of visualizations. Several aspects of our results suggest interesting avenues for future work around the impact of aggregations on users' confidence and the nature of their generalizations.

First, while aggregation strategy did not appear to significantly affect accuracy rates, in both studies we observed that generalizations about combinations of a nominal variable and a quantitative variable were considerably more likely to lead to inaccurate generalizations than other combinations. Visualization system designers may want to explore ways of warning viewers of such combinations that while tempting to make, comparisons may not be reflective of the population.

While we observed no reliable differences in accuracy, across both a large (Experiment 1) and small (Experiment 2) sample, disaggregating data without a mean mark led to less confidence in generalizations by 3 to 6 points on average. Considering well-known biases like overconfidence [LFP81], more cautious conclusions are likely to benefit analysts by helping them avoid false positives. That said, given that the confidence scale we used was 100 points, the size of this difference suggests that overconfidence is unlikely to be completely offset aggregation strategy. Other debiasing strategies may be needed.

We considered whether portraying an effect in dichotomous "present or absent" terms (e.g., "there is an effect of ad campaign on visits"), a type of thinking about effects that has been sharply criticized in the wake of a replication crisis occurring in several fields [HFKJ06], could be a useful metric for differentiating behavior from aggregation strategies. Indeed, we found that the probability that a viewer would describe an effect in dichotomous terms was only about one fifth to one half as high when they viewed a disaggregated view rather than a single summary mark showing the mean. Along the same lines, we saw a bigger reduction in the number of generalizations produced between large and small sample when participants used disaggregated views relative to mean aggregation only.

These results shed some empirical light on Correll and Gleicher's [CG15] question about how views that require ensemble processing affect confidence in data. It appears that displaying marks of mean aggregation pose a risk that untrained users will overestimate the reliability of superficial generalizations in which patterns either exist or do not. We find it promising that even when users are relatively unfamiliar with statistics and visualization, they appear to be sensitive to cues like sample size afforded by disaggregated views. When analysis settings are unlikely to involve very large datasets and users may not be trained in analysis, using disaggregation may lead to more conservative conclusions.

Our experiments intentionally traded off external validity for control over the stimuli participants viewed and statistical power. According to self-reported experience with statistics and visualizations, the majority of participants in our experiments were novices. Future work should undertake study of how effect size interpretations manifest in realistic exploratory data analysis sessions with users that range in experience. Given the increase in reliance on data in a number of domains, further exploration into what features of a visualization environment viewers are sensitive to and how visualizations can best communicate effect size and support estimation in more personalized ways is well warranted.

### 5.1. Evaluating Exploratory Data Analysis

Our work builds on Zgraggen et al.'s [ZZZK18] recent study of how susceptible a group of student analysts were to the multiple comparisons problem. Differences between our results and theirs may provide insight into causes of inaccurate generalizations and methodologies for studying bias in data-driven generalizations.

For instance, there is a considerable disparity between the false positive rate observed in their study (73.8%) versus ours (33.0% to 34.2% or 41.0% to 49.3% depending on aggregation strategy and sample size).

There are a number of reasons why our results may paint a more positive picture of novice analysts' false positive rates. First, we conducted a controlled study in which participants did not generate views themselves. It is possible that confirmation bias causes people to be more likely to identify effects (including unsupported effects) in views that they generated themselves based on some question or interest. We also embedded ground truth correlations (which were always between 0.4 and 0.8 in the data we showed participants), whereas Zgraggen et al.'s ground truth included any non-zero correlation. It is possible that participants in their study stated that variables with very small correlations were null effects when in fact they were not. Their sample size was also significantly smaller than ours, potentially contributing noise; they recorded a total of 155 generalizations where we accumulated a total of 1,743 and 1,609 for experiment 1 and 2 respectively. Finally, an experimenter effect may have led their participants in their study to generate inferences simply because they thought that was what they were expected to do. While this may also be true in our setting, the lack of an in-person observer may have made our participants feel less pressure to produce inferences for views where they did not see a clear pattern.

Our results have implications for future studies of inference through exploratory data analysis. That participants in our study did not exhibit subjective confidence levels that matched the requested level of reliability in our prompt suggests that it may not be possible to prescribe a confidence level and expect analysts to use it as a filter for their inferences. While the type of relative subjective confidence that we elicited could potentially be compared to results of a statistical test of an inference (see, e.g., [CG14]), we believe that the ambiguity of confidence as a construct makes confidence a noisy signal at best [HQC*19]. As an alternative, visualization researchers have suggested understanding threats to the reliability of conclusions drawn during exploratory data analysis more broadly as an example of Bayesian inference [HH18] or a garden of forking paths in which "model overfitting" (overconfidence in trends observed in a sample) can be mitigated using regularization or visual bias corrections [PK18]. Our experiment results lead us to believe that a better approach to evaluating the integrity of exploratory data analysis may be to elicit intervals for any stated effects directly from the analyst. For example, if an analyst describes a small effect of ad campaign on number of purchases, they could be asked to describe their best guess of the size of the effect as an interval. A challenge may arise if the analyst is not accustomed to thinking about uncertainty intervals. However, coarser (e.g., multiple choice) descriptions of potential effect sizes are possible, or graphical elicitation interfaces similar to those intended for gathering untrained users' prior and posterior beliefs that have recently been demonstrated as a means to evaluate visualizations [**kim2019**]. Importantly, more fine-grained approaches to representing subjective probability could support more reliable testing of inferences against a ground truth.

### 5.2. Limitations

Naturally, our study has some limitations, most notably that the untrained participants we recruited from Mechanical Turk may be less accustomed to working with data than the novice users of visualization systems for which aggregation defaults are an important choice. The visual marks used to convey the mean differ across our three aggregation strategy conditions, with the mean aggregation condition using a bar rather than a horizontal tick. It is possible that this difference in mark contributed additional noise or bias. It is also possible that our use of mean aggregation in two strategy conditions made participants more likely to draw generalizations about the mean. Comparing the first trial distribution across generalization types of disaggregation to the full trial results in Figure 5, it appears that with more trials, those viewing disaggregated views did become more likely to report mean generalizations. The breakdown of generalization types by aggregation strategy should be taken as evidence of relative but not absolute differences. In addition, the methods we employed to translate generalizations into testable statements of various types, while systematic and preregistered, still have many degrees of freedom. The reported accuracy levels should not be interpreted as absolute.

## 6. Conclusion

Little empirical work in visualization attempts to study how visual aggregation (or other visualization decisions) impact the types of conclusions people draw from summary visualizations. We presented the results of two pre-registered experiments that compare statistically untrained visualization users' generalizations made from three different aggregation strategies: disaggregating the data, mean aggregation of the data and disaggregating the data with mean marks overlaid. Our results provide an initial characterization of how aggregation strategy affects the number of, focus of, accuracy of, and confidence of generalizations produced by untrained users.

## References

[ACG14] Albers, Danielle, Correll, Michael, and Gleicher, Michael. "Task-driven evaluation of aggregation in time series visualization". *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2014, 551–560 2.

[Alv11] Alvarez, George A. "Representing multiple objects as an ensemble enhances visual cognition". *Trends in cognitive sciences* 15.3 (2011), 122–131 2.

[Ama18] Amazon. *Amazon Mechanical Turk*. 2018. URL: https://www.mturk.com/ 3.

[AO07] Alvarez, George and Oliva, Aude. "The representation of ensemble visual features outside the focus of attention". *Journal of Vision* 7.9 (2007), 129–129 2.

[ARH12] Aigner, Wolfgang, Rind, Alexander, and Hoffmann, Stephan. "Comparative evaluation of an interactive time-series visualization that combines quantitative data with qualitative abstractions". *Computer Graphics Forum*. Vol. 31. Wiley Online Library. 2012, 995–1004 2.

[BFWC05] Belia, Sarah, Fidler, Fiona, Williams, Jennifer, and Cumming, Geoff. "Researchers misunderstand confidence intervals and standard error bars." *Psychological methods* 10.4 (2005), 389 3, 4.

[BHG*11] Bautista, Susana, Hervás, Raquel, Gervás, Pablo, et al. *How to Make Numerical Information Accessible: Experimental Identification of Simplification Strategies*. 2011 7.

[BW08] Bachthaler, Sven and Weiskopf, Daniel. "Continuous scatterplots". *IEEE Trans. on Visualization and Comp. Graphics* 14.6 (2008), 1428–1435 3.

[CA76] Clark, William AV and Avery, Karen L. "The effects of data aggregation in statistical analysis". *Geographical Analysis* 8.4 (1976), 428–438 2.

[CG14] Correll, Michael and Gleicher, Michael. "Error bars considered harmful: Exploring alternate encodings for mean and error". *IEEE Trans. on Visualization and Comp. Graphics* 20.12 (2014), 2142–2151 3, 4, 13.

[CG15] Correll, Michael and Gleicher, Michael. "Implicit Uncertainty Visualization: Aligning Perception and Statistics". *Proc. of the 2015 Workshop on Visualization for Decision Making Under Uncertainty*. Oct. 2015. URL: http://graphics.cs.wisc.edu/Papers/2015/CG15 2, 13.

[CH17] Correll, Michael and Heer, Jeffrey. "Regression by eye: Estimating trends in bivariate visualizations". *Proc. of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM. 2017, 1387–1396 1, 2.

[Coe02] Coe, Robert. "It's the effect size, stupid: What effect size is and why it is important". *Annual Conference of the British Educational Research Association* (2002) 3.

[Cor18] Corporation, Microsoft. *Power BI Desktop, Version 2.62.5222.761*. 2018. URL: https://powerbi.microsoft.com/en-us/desktop/ 3.

[Cum14] Cumming, Geoff. "The New Statistics: Why and How". *Psychological Science* 25.1 (2014), 7–29. URL: https://doi.org/10.1177/0956797613504966 4.

[DOV11] Diemand-Yauman, Connor, Oppenheimer, Daniel M, and Vaughan, Erikka B. "Fortune favors the (): Effects of disfluency on educational outcomes". *Cognition* 118.1 (2011), 111–115 10.

[EF10] Elmqvist, Niklas and Fekete, Jean-Daniel. "Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines". *IEEE Trans. on Visualization and Comp. Graphics* 16.3 (2010), 439–454 2.

[EfIM82] Efron, Bradley, for Industrial, Society, and Mathematics, Applied. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia, Pa. : Society for Industrial and Applied Mathematics, 1982. ISBN: 0898711797 (pbk.) 4.

[FWM*18] Fernandes, Michael, Walls, Logan, Munson, Sean, et al. "Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making". *Proc. of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, 144 4.

[Gal95] Gal, Iddo. "Statistical tools and statistical literacy: The case of the average". *Teaching Statistics* 17.3 (1995), 97–99 3.

[GBFM16] Gschwandtneir, T., Bögl, M., Federico, P., and Miksch, S. "Visual Encodings of Temporal Uncertainty: A Comparative User Study". *IEEE Trans. on Visualization and Comp. Graphics* 22.1 (Jan. 2016), 539–548. ISSN: 1077-2626. DOI: 10.1109/TVCG.2015.2467752 2.

[GCNF13] Gleicher, Michael, Correll, Michael, Nothelfer, Christine, and Franconeri, Steven. "Perception of average value in multiclass scatterplots". *IEEE Trans. on Visualization and Comp. Graphics* 19.12 (2013), 2316–2325 2.

[HB15] Hubert-Wallander, Bjorn and Boynton, Geoffrey M. "Not all summary statistics are made equal: Evidence from extracting summaries across time". *Journal of vision* 15.4 (2015), 5–5 2.

[HFKJ06] Hoekstra, Rink, Finch, Sue, Kiers, Henk AL, and Johnson, Addie. "Probability as certainty: Dichotomous thinking and the misuse ofp values". *Psychonomic Bulletin & Review* 13.6 (2006), 1033–1037 4, 10, 12.

[HH18] Hullman, Jessica and Heer, Jeffrey. *Multiple Perspectives on the Multiple Comparisons Problem in Visual Analysis*. 2018. URL: https://medium.com/hci-design-at-uw/multiple-perspectives-on-the-multiple-comparisons%20-problem-in-visual%20-analysis-df7493818bb 13.

[HQC*19] Hullman, Jessica, Qiao, Xiaoli, Correll, Michael, et al. "In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation". *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2019). URL: http://idl.cs.washington.edu/papers/uncertainty-eval-survey 4, 13.

[HRA15] Hullman, Jessica, Resnick, Paul, and Adar, Eytan. "Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering". *PloS one* 10.11 (2015), e0142444 2, 4.

[Hul16] Hullman, Jessica. "Why evaluating uncertainty visualization is error prone". *Proc. of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*. ACM. 2016, 143–151 4, 9.

[HW09] Haberman, Jason and Whitney, David. "Seeing the mean: ensemble coding for sets of faces." *Journal of Experimental Psychology: Human Perception and Performance* 35.3 (2009), 718 2.

[HZ84] Hasher, Lynn and Zacks, Rose T. "Automatic processing of fundamental information: the case of frequency of occurrence." *American Psychologist* 39.12 (1984), 1372 2.

[IIS*14] Isenberg, Petra, Isenberg, Tobias, Sedlmair, Michael, et al. *Toward a deeper understanding of Visualization through keyword analysis*. Research Report RR-8580. INRIA, Aug. 2014. URL: https://hal.inria.fr/hal-01055309 4.

[Inc18] Inc, TIBCO Software. *TIBCO Spotfire, Version 7.14*. 2018. URL: https://spotfire.tibco.com/ 3.

[KKHM16] KAY, MATTHEW, KOLA, TARA, HULLMAN, JESSICA R, and MUNSON, SEAN A. "When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems". *Proc. of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM. 2016, 5092–5103 4.

[KNKH18] KALE, ALEX, NGUYEN, FRANCIS, KAY, MATTHEW, and HULLMAN, JESSICA. "Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data". *IEEE Trans. on Visualization and Comp. Graphics* (2018) 2, 4.

[LBI*12] LAM, H., BERTINI, E., ISENBERG, P., et al. "Empirical Studies in Information Visualization: Seven Scenarios". *IEEE Trans. on Visualization and Comp. Graphics* 18.9 (Sept. 2012), 1520–1536. ISSN: 1077-2626. DOI: 10.1109/TVCG.2011.279 4.

[LFP81] LICHTENSTEIN, SARAH, FISCHHOFF, BARUCH, and PHILLIPS, LAWRENCE D. *Calibration of probabilities: The state of the art to 1980*. Tech. rep. DECISION RESEARCH EUGENE OR, 1981 12.

[LKW16] LEIB, ALLISON YAMANASHI, KOSOVICHEVA, ANNA, and WHITNEY, DAVID. "Fast ensemble representations for abstract visual impressions". *Nature communications* 7 (2016), 13186 2.

[McE15] MCELREATH, RICHARD. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, 2015. ISBN: 9781482253443 8.

[McE16] MCELREATH, RICHARD. *Rethinking: An R Package for Fitting and Manipulating Bayesian Models v1.56*. 2016 8.

[MPOW17] MICALLEF, L., PALMAS, G., OULASVIRTA, A., and WEINKAUF, T. "Towards Perceptual Optimization of the Visual Design of Scatterplots". *IEEE Trans. on Visualization and Comp. Graphics* 23.6 (June 2017), 1588–1599. ISSN: 1077-2626. DOI: 10.1109/TVCG.2017.2674978 2.

[Neu12] NEUMAN, WILLIAM LAWRENCE. *Basics of Social Research: Qualitative and Quantitative Approaches (3rd Edition)*. Pearson Publishing, 2012 3, 4.

[PK18] PU, XIAOYING and KAY, MATTHEW. "The garden of forking paths in visualization: A design space for reliable exploratory visual analytics". *Evaluation and Beyond - Methodological Approaches for Visualization*. BELIV 2018. 2018 13.

[Sch13] SCHUMACHER, JASON. *Tableau for Students: Free access to Tableau Desktop*. 2013. URL: https://www.tableau.com/about/blog/2013/3/tableau-students-free-access-tableau-desktop-21617 (visited on 12/19/2019) 1.

[SGS18] SARIKAYA, A, GLEICHER, M, and SZAFIR, DA. "Design factors for summary visualization in visual analytics". *Computer Graphics Forum*. Vol. 37. 2018, 145–156 2.

[SHGF16] SZAFIR, DANIELLE ALBERS, HAROZ, STEVE, GLEICHER, MICHAEL, and FRANCONERI, STEVEN. "Four types of ensemble coding in data visualizations". *Journal of vision* 16.5 (2016), 11–11 1, 2.

[SK06] SHAH, RAJIV C and KESAN, JAY P. "Policy through software defaults". *Proc. of the 2006 international conference on Digital government research*. Citeseer. 2006, 265–272 1.

[Tab18] TABLEAU SOFTWARE, INC. *Tableau Desktop, Version 2018.2*. 2018. URL: https://www.tableau.com/products/desktop 1, 3.

[Tea18a] TEAL, SCOTT. *Journalists: Now Tableau Prep is free for you*. 2018. URL: https://public.tableau.com/en-us/s/blog/2018/05/journalists-now-tableau-prep-free-you (visited on 12/19/2019) 1.

[Tea18b] TEAM, STAN DEVELOPMENT. *Package 'rstan'*. 2018 8.

[TK74] TVERSKY, AMOS and KAHNEMAN, DANIEL. "Judgment under Uncertainty: Heuristics and Biases". *Science* 185.4157 (1974), 1124–1131. ISSN: 00368075, 10959203. URL: http://www.jstor.org/stable/1738360 3.

[Tuk77] TUKEY, JOHN W. *Exploratory data analysis*. Vol. 2. Reading, Mass., 1977 4.

[WMA*16] WONGSUPHASAWAT, KANIT, MORITZ, DOMINIK, ANAND, ANUSHKA, et al. "Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations". *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2016). URL: http://idl.cs.washington.edu/papers/voyager 1, 3, 5.

[ZDZ*17] ZHAO, ZHEGUANG, DE STEFANI, LORENZO, ZGRAGGEN, EMANUEL, et al. "Controlling false discoveries during interactive data exploration". *Proc. of the 2017 ACM International Conference on Management of Data*. ACM. 2017, 527–540 4.

[ZZZK18] ZGRAGGEN, EMANUEL, ZHAO, ZHEGUANG, ZELEZNIK, ROBERT, and KRASKA, TIM. "Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis". *Proc. of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, 479 3, 4, 6, 7, 9, 13.