Synergy between research on ensemble perception, data visualization, and statistics education: A tutorial review

Lucy Cui¹ · Zili Liu¹

Accepted: 17 November 2020 / Published online: 3 January 2021
© The Psychonomic Society, Inc. 2021

Abstract

In the age of big data, we are constantly inventing new data visualizations to consolidate massive amounts of numerical information into smaller and more digestible visual formats. These data visualizations use various visual features to convey quantitative information, such as spatial position in scatter plots, color saturation in heat maps, and area in dot maps. These data visualizations are typically composed of ensembles, or groups of related objects, that together convey information about a data set. Ensemble perception, or one's ability to perceive summary statistics from an ensemble, such as the mean, has been used as a foundation for understanding and explaining the effectiveness of certain data visualizations. However, research in data visualization has revealed some perceptual biases and conceptual difficulties people face when trying to utilize the information in these graphs. In this tutorial review, we will provide a broad overview of research conducted in ensemble perception, discuss how principles of ensemble encoding have been applied to the research in data visualization, and showcase the barriers graphs can pose to learning statistical concepts, using histograms as a specific example. The goal of this tutorial review is to highlight possible connections between three areas of research—ensemble perception, data visualization, and statistics education—and to encourage research in the practical applications of ensemble perception in solving real-world problems in statistics education.

Keywords Ensemble perception · Data visualization · Graphical perception · Statistics education

Our visual system is collecting summary information from our environment all the time, whether it be scanning an arrangement of boxed strawberries to select the “best” box to purchase, scanning a crowd of people exiting an auditorium to decide which exit is the least crowded, or scanning an audience to assess how your performance was received. Our ability to extract summary statistics of visual features, like mean color of a box of strawberries, mean direction of a crowd, or mean facial expression from an audience, helps us navigate our environment and make decisions that best serve us. This ability has been studied by researchers for decades under the term “ensemble perception.” While summary statistics can be extracted via other modalities, like audition (e.g., E. A. Piazza, Sweeny, Wessel, Silver, & Whitney, 2013), this tutorial review will only discuss ensemble perception in the visual modality.

While most of the field have historically concentrated on our ability to estimate mean (see Whitney & Yamanashi Leib, 2018, for a recent review), as of more recently, the study of variance perception and the use of ensemble coding in interpreting data visualizations have been on the rise. Both areas of research have implications for statistics education. For one, variance is one of the most difficult statistical concepts for students to understand and reason accurately about (Reid & Reading, 2008), yet it is a summary statistic that our visual system is sensitive to (Bronfman, Brezis, Jacobson, & Usher, 2014; Haberman, Lee, & Whitney, 2015; Morgan, Chubb, & Solomon, 2008; Solomon, 2010; Tong, Ji, Chen, & Fu, 2015). Next, while a lot of the graphical perception literature (e.g., Shah, Freedman, & Vekiri, 2005; Shah & Hoeffner, 2002) focuses on simpler graphs, such as bar graphs and line graphs for interaction plots (e.g., Shah & Freedman, 2011); data visualizations and the graphs used to teach core concepts in statistics, such as correlation with scatter plots and distributions with histograms, use many more graphical marks (e.g., bars, dots, lines).

In the first section of this tutorial review, we will provide an overview of the research conducted in ensemble perception on mean and variance perception. We will summarize the factors

✉ Lucy Cui
lucy.cui@ucla.edu

¹ University of California, Los Angeles, 502 Portola Plaza, Los Angeles, CA 90095, USA

that influence our capacity for mean and variance perception and discuss the relationship between the two capacities. The list of factors discussed is not meant to be exhaustive. While researchers study many factors, like spatial grouping/layout (e.g., Yildirim, Ögreden, & Boduroglu, 2018) and viewing distance (e.g., Tiurina & Utochkin, 2019), this tutorial review will focus on the factors related to statistical distributions that are most relevant to research in data visualization and statistics education. Next, we will provide an overview of the research conducted on data visualizations, focusing on studies that relate to ensemble perception or statistics education. Finally, we will provide an overview of the research conducted in statistics education on students' reasoning about mean and variance, as well as their misinterpretations of histograms.

The goal of this tutorial review is to draw connections between the three research areas in order to inspire research into investigating ways to optimize graphs for statistics education. The scope of this tutorial review has been narrowed based on this goal, and our review of each research area is not meant to be exhaustive. Our intended audience is any researcher in ensemble perception, data visualization, or statistics education who may be interested in contributing to or applying research from the other research areas.

Ensemble perception

There is bountiful sensory information around us—too much for our limited cognitive resources to encode everything. How do our brains cope? Researchers have proposed that our brains cope through condensing the information into smaller pieces of summary information. This ability to extract summary information (i.e., summary statistics such as mean) from a scene, or array of items, is referred to as *ensemble perception*. An *ensemble* (referred to as an array in experiments) refers to a set of objects, features, or configurations that have some meaningful or consistent relationship amongst its members and an underlying statistical distribution of some visual feature (Whitney & Yamanashi Leib, 2018). These ensembles could be composed of countable objects or refer to uncountable features of an object or scene (e.g., texture on a grass field). Some examples of ensembles include berries on a bush, animals in a herd, and people in a crowd. Ensemble perception, operationally defined, is our ability to reproduce statistical moments from ensembles or discriminate between the statistical moments of multiple ensembles (Whitney & Yamanashi Leib, 2018).

Statistical moments

Statistical moments measure characteristics of the shape of a distribution. There are different levels of these statistical moments, each abstracting information regarding the distribution

one order higher than the previous one. In the field of statistics, there are four statistical moments that are commonly used as well as higher-order statistics (i.e., 5th, 6th; Hsu & Lawley, 1940) that are rarely used due to difficulty in interpretation. The four statistical moments discussed below are statistical definitions. Whether humans are sensitive to all four and to what extent still remains an unanswered question.

The first statistical moment is the one we are most familiar with—*mean*, or average. This is also the same level that most researchers study ensemble perception. The second statistical moment is *variance*, or an aggregate measure of how much individuals deviate from the mean of a group—the study of which is more recently popularized by, for example, Morgan et al. (2008). The third statistical moment is *skewness*, a measure of the degree and direction of asymmetry in a distribution. Few studies on the third statistical moment were found within visual ensemble perception literature (Chetverikov, Campana, & Kristjánsson, 2016; Oriet & Hozempa, 2016). However, research on people's sensitivities to skewness is more prevalent in the risk-taking and financial decision-making literature (e.g., Åstebro, Mata, & Santos-Pinto, 2015; Kraus & Litzenberger, 1976). Finally, the fourth statistical moment is *kurtosis*—the measure of the flatness, peakedness, or tailedness of a distribution (Westfall, 2014). Research on the visual ensemble perception of the third and fourth statistical moments is scarce to none. Due to this scarcity, this tutorial review will only discuss mean and variance.

Individual representations

Foremost, all statistical moments are calculated using individual data points from a distribution and represent different relationships between those individual data points. Before discussing people's perceptions of the first two statistical moments, we should discuss the relationship ensemble perception has with individual items in an array. Ensemble perception is thought to economize the limited capacity of our visual system by not requiring full encoding of each individual item in an array (Chong & Treisman, 2003).

Other important components of the operational definition of ensemble perception juxtaposes ensemble representations with representations of individual objects. For one, single item recognition is not required for ensemble coding, and it is often the case that individuals in an array cannot be reported (e.g., Ward, Bear, & Scholl, 2016). What is more, ensemble representations can be more vivid and accurate than that of individuals. For example, size discrimination is more accurate for arrays of same-sized circles than for individual circles (Chong & Treisman, 2003). This effect might be appreciated as follows. Estimating the size of a circle necessarily entails measurement errors. Having available multiple samples of the same-sized circle at multiple spatial locations provides an opportunity to measure the same circle repeatedly. The

uncertainty of the mean of such measurements (or the standard error) is by definition smaller than that of any single measurement. Hence, discrimination sensitivity is higher with multiple circles than with a single circle on each side. Secondly, ensemble representations can be extracted as fast or faster than recognitions of individual objects or search for an object (e.g., Haberman & Whitney, 2009). For example, people perceive characteristics of crowd behavior (i.e., average speed and average direction of headings) better than that of individuals (i.e., the speed and direction of a particular individual in a crowd; Sweeny, Haroz, & Whitney, 2013). One way to summarize an ensemble of individuals is numerosity.

Numerosity, or sample size

Numerosity, or quantity, provides summary information about our environment. Being able to approximate number or count objects has undeniable ecological value, and our ability to do so has its evolutionary roots (e.g., Brannon & Merritt, 2011). However, researchers of ensemble perception do not usually list numerosity as a summary statistic (cf. Khvostov & Utochkin, 2019; Solomon & Morgan, 2018; Valsecchi, Stucchi, & Scocchia, 2018; “On the other hand, estimation involves an approximation of the number of different items or a sense of ensemble statistics”: p. 1, Chong & Evans, 2011). Our ability to extract information about numerosity from our environment is studied mostly by researchers of numerical cognition. In the field of numerical cognition, approximate number system (ANS) refers to the cognitive system that supports the approximation of quantities (see Odic & Starr, 2018, for a review). The acuity of this system, measured through nonsymbolic numerical comparison tasks (i.e., which array has more items), develops throughout childhood (M. Piazza & Izard, 2009) and can be improved through training even in adulthood (Cochrane, Cui, Hubbard, & Green, 2019). While we are able to very accurately estimate quantities less than four without counting, an ability referred to as *subitizing* (see Katzin, Cohen, & Henik, 2019, for a review), our enumeration of quantities above 10 follow Weber’s law (Revkin, Piazza, Izard, Cohen, & Dehaene, 2008). Mathematically, numerosity is involved in the calculation of the mean, the perception of which is popular in ensemble perception research.

Mean perception

The ability to extract the mean value of a visual feature from a set of items spans across low-level features, such as size (e.g., Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013; Ariely, 2001; Corbett & Melcher, 2014; Corbett, Wurnitsch, Schwartz, & Whitney, 2012; Im & Halberda, 2013; Luo & Zhao, 2018; Myczek & Simons, 2008; Tiurina & Utochkin, 2019; Tokita, Ueda, & Ishiguchi, 2016), line length (Bauer, 2017; Utochkin, Khvostov, & Stakina, 2018), orientation

(Solomon, 2010; Utochkin et al., 2018; Witt, 2019), and hue (Maule & Franklin, 2015, 2016; Michael, de Gardelle, & Summerfield, 2014; Tong et al., 2015), to high-level features, such as emotion and gender (e.g., Haberman & Whitney, 2007, 2009), facial expressions (e.g., Griffiths, Rhodes, Jeffery, Palermo, & Neumann, 2018; Li et al., 2016; Wolfe, Kosovicheva, Leib, Wood, & Whitney, 2015), and lifelikeness (e.g., Yamanashi Leib, Kosovicheva, & Whitney, 2016). Size is perhaps one of the most studied features, and is tested with various stimuli: typically, with dots/circles (e.g., Ariely, 2001; Chong & Treisman, 2003), but also with concrete illustrations of strawberries and lollipops (Yang, Tokita, & Ishiguchi, 2018).

Out of all four statistical moments, one may argue that the ability to extract the mean of a distribution from a visual scene is the most relevant for everyday life, and most ecologically grounded. For example, the ability to represent size and color distributions of berries would have been very useful when our hunter-gatherer ancestors were scanning for bushes that had the biggest, ripest, and most berries. When encountering other tribes and assessing whether they posed a threat, the ability to quickly assess the distribution of emotions, such as friendly or angry expressions, would support a tribe’s survival and facilitate befriending other tribes.

Factors that influence mean perception As previously mentioned, adaptation to an array biases participants’ mean perception of a subsequent array (e.g., Corbett et al., 2012). For example, Corbett and Melcher (2014) had participants adapt to patches of small-sized and large-sized dots in opposite regions of the screen (left/right or top/bottom) and asked participants to judge which of the two test dots (one test dot per tested region) is bigger. Participants had the tendency of perceiving the subsequently viewed test arrays as having an average that is biased towards the opposite end of the feature spectrum of the adapted array. Ying, Burns, Lin, and Xu (2019) found that adapting to a group of unattractive faces made participants judge subsequent faces to be more attractive than they would without the adaptation.

Other factors that influence mean perception can be divided into the items within the ensembles and how those items compare to other items in the ensemble, or characteristics that describe the ensemble’s composition or feature distributions. As for the items within the ensembles, set size has been shown to influence perceptual averaging performance. For small set sizes of two to three items, viewers show a lot of uncertainty, taking longer to estimate the mean of the ensemble and often being very inaccurate—their estimates deviating more from the true mean value than estimates of larger set sizes. Typically, viewers perform better with more items in the ensemble (Allik et al., 2013; Baek & Chong, 2020; Haberman & Whitney, 2010). However, up to a certain large set size, accuracy of mean estimation no longer improves (Dakin, 2001; Robitaille & Harris,

2011), leading some researchers to believe that only a subset of items are actually used for estimating the mean, with the number of incorporated items being constant regardless of set size (Simons & Myczek, 2008), or exponentially decreasing as the set size increases (Dakin, 2001). However, some studies have found relatively constant sensitivity to the mean with increasing set size (Alvarez, 2011; Ariely, 2001; Chong & Treisman, 2005) with a few cases where averaging ability actually decreases with an increasing set size (Ji & Pourtois, 2018, when large variance exists in the ensemble).

As for the factors relating to an ensemble's composition, measures related to the ensemble's extreme values seem to play a role in influencing mean estimations. For example, an ensemble's range changes the difficulty of a mean estimation task. Specifically, Maule and Franklin (2015) found that increasing the range of colors used, measured by the unit of just-noticeable-differences (JND; Witzel & Gegenfurtner, 2013), decreased accuracy and speed of averaging hue (e.g., Experiment 1: "green" to "blue"; Experiment 2: at the most extreme, mean of equally spaced [in JND] color hues between "pink" and "light blue" is "yellowish green"). Post hoc analysis of whether number of color categorical boundaries explain a decrease in performance with increased range revealed that it is the perceptual differences between items that affected performance.

While an ensemble can have a huge range, extreme values may be discounted if there are too few of them. Those items are regarded as outliers. Instead of being incorporated into the mean estimation, outliers seem to be discarded or less weighted when estimating the mean, as with the case of emotional outliers (Haberma & Whitney, 2010). The perceived means, in turn, better reflect the mean of the majority of the array's items and are not pulled towards the outliers as one would expect from exact computation. This stands in stark contrast to what students typically do when asked to estimate the average of numbers, which is to overweight the outliers (Anderson, 1968). This tendency to underweight outliers in our visual environment may make sense from an ecological perspective. One or two angry people in a crowd ought not change one's decision on how to deal with a foreign crowd of people, who is otherwise consistently friendly. Instead, one has many opportunities to build good relations with the new group and just need to avoid a select few where conflicts may arise.

People may have a preference for the mean, as with the literature on perceived attractiveness, suggesting that people with average facial features are more attractive than people with more unique/idiosyncratic facial features (e.g., Langlois & Roggman, 1990). For example, de Fockert and Wolfenstein (2009) investigated the extraction of mean identity by showing participants a set of four faces and then asking participants whether a test face was present in the previously seen set. It turns out that participants responded yes to faces that were the morphed mean (of all four faces and not previously presented)

more often than to faces that was actually presented in the previously seen set. This result highlights how important the mean is to the brain, so much so that it created a false memory of the mean being in the set. It also suggests that our memory for the mean is clearer and more accessible than any one of the individuals (Posner & Keele, 1970). In another study, when shown a sequence of items and then asked which of two stimuli was a member of the set, participants preferentially chose the members that were the mean or close to it, and this was the case for various types of stimuli: circle size, line orientation, and circle brightness (Khayat & Hochstein, 2018). Finally, not only does the mean get classified as being part of a previously seen set but the mean also seems to influence the memory of items in that set. Utochkin and Brady (2020) found that participants' memory for individual orientations was biased towards the mean. The false memory that the mean was previously seen in the set may relate to why students believe that the average of a data set exists within that set (Batanero, Cobo Merino, & Diaz, 2003). Students may bring their previous experiences with them into a statistics course and use those experiences to relate to the curriculum. Therefore, the misconception that the mean exists in a data set may reflect general beliefs or assumptions about the world—that the prototype of a category exists as a real and previously seen object.

Finally, mean perception is affected by the variance of an ensemble. We will discuss this relationship in more detail later, but first, we will discuss variance perception and factors that influence it.

Variance perception

The perception of the mean has been shown to be influenced by factors that have some relationship with variance, such as range, existence of outliers, and heterogeneity (e.g., Chong & Treisman, 2003; Marchant, Simons, & de Fockert, 2013), leading researchers to manipulate variance directly and study variance perception within ensemble perception. There is evidence that our visual system can extract the variance of an array. This has been seen in the reproduction of the variance of a previously seen array, such as with facial emotion (Haberma et al., 2015), and discrimination of the variances in two arrays, such as with orientation (Morgan et al., 2008; Solomon, 2010; Yang et al., 2018), gray hues (Tong et al., 2015), color hues (Bronfman et al., 2014), and motion direction (Suárez-Pinilla, Seth, & Roseboom, 2018).

Our sensitivity to variance may have been an evolutionary advantage. Just as there is preference for the mean, such as for the mean face (de Fockert & Wolfenstein, 2009), people also have preferences for low variance. People are more likely to categorize themselves as part of a low variance group if their reported emotion is close to the group average and part of a high variance group if their reported emotion is much different from the group average (Goldenberg, Sweeny, Shpigel, &

Gross, 2020). For each of these cases, the chosen group optimizes one's chance for acceptance within the group (i.e., to fitting in, not standing out, respectively).

Unfortunately, there is limited literature on variance perception. This could be due in part to the mean appearing to be the more evolutionarily useful summary statistic to extract visually, as compared with the auditory modality, where the mean may be less significant. For the auditory modality, it may have been more evolutionarily advantageous to develop an ability to extract other summary statistics, like variance and skewness, for situations like assessing the number of approaching herds and inferring the level of threat when it is hard to see.

Factors that influence variance perception Given the paucity of literature on variance perception, there is limited knowledge on the factors that influence variance perception. As with mean perception, adapting to variance biases subsequent variance perception away from the adapted stimulus (as with orientation and color hue: Maule & Franklin, 2019). Spatial locations and spatial relationships between individual items seem to affect variance perception. For example, participants have more difficulty discriminating between small variance differences when the items are crowded in the display (Solomon, 2010). Which individual items are available near the fixation point also seems to affect variance perception. For example, in Lau and Brady (2018), items with extreme values, such as minimums and maximums that give a sense of the range of features, were either displayed close to or far from fixation. When these extreme values were close to fixation, participants seemed to rely on the range to estimate variance because they performed worse when there was incongruence between range and variance. Whereas, when the extreme values were far away from fixation, participants' variability discrimination was not affected by range, possibly because they relied on the range of values near fixation and did not gather complete information about range.

Relationship between mean and variance

The relationship between the perception of mean and the perception of variance is still relatively unclear. How do the two relate to each other? Do they draw from similar mechanisms? Are they dependent on each other? Theoretically, we know that neither summary statistic—mean or variance—gives us information about the other without knowing the individuals that make up the set, so assessing one will not replace the assessment of the other. More importantly, we cannot infer whether two distributions are likely significantly different from each other without both pieces of information. While students' visual system may incorporate both mean and variance into determining whether two groups are the same, the need for both summary statistics (i.e., the statistical terms) is

not so clear to students, who often think that just the mean is enough to determine whether two data sets came from separate populations (Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007). Just like these two pieces of information work together to give us information about the distributions, it also appears that mean perception and variance perception support each other.

Keeping one of the two parameters constant seems to benefit the perception of the other. This is particularly the case for keeping the variance constant when having participants perceive the average visual feature of an array. For example, participants respond to average discrimination tasks faster when they are first exposed to an array with the same variance as the subsequent array that they need to extract the average from (Michael et al., 2014). Constant variance also serves as a buffer, as sensitivity to the mean is less influenced by the shape of the statistical distribution (e.g., uniform, normal, bimodal; Allik et al., 2013; Chong & Treisman, 2003; Haberman & Whitney, 2009). It is interesting that our visual system is robust against the shape of distributions, compared with how much the shape of a distribution affects students' assessment of variance in histograms (Cooper & Shore, 2008; Kaplan, Gabrosek, Curtiss, & Malone, 2014; Meletiou-Mavrotheris & Lee, 2002). This difference suggests that the representation of data in histograms is not intuitive to our visual system.

Varying one of the parameters seems to reduce performance in perceiving the other. For example, increasing the variance of an array reduces sensitivity to the mean of the array (Dakin, 2001; Fouriez, Rubinfeld, & Capstick, 2008; Haberman et al., 2015; Im & Halberda, 2013; Morgan et al., 2008; Solomon, Morgan, & Chubb, 2011). Likewise, varying the mean of arrays was found to reduce participants' sensitivity to variance and increase perceived variance (Tong et al., 2015). In this particular study, participants compared two 9×9 grids of gray tones and were asked to choose the grid with more grayscale variability. Participants performed better when the mean gray tone matched between the two grids than when they were different.

While it appears that mean perception and variance perception support one another, it would be inaccurate to claim that the two perceptions depend on each other. For one, perception of the mean heading of a crowd is tolerant to variability of direction and speed of the crowd (Sweeny et al., 2013), which may be a specific characteristic of higher-level visual features. Second, some effects that provide evidence that people do process variance, such as the aftereffect from adapting to variance, are durable to changes in mean (Norman, Heywood, & Kentridge, 2015).

Research investigating the relationship between mean perception and variance perception directly are scarce. Recently, a study found an interaction between mean and variance representations, where adapting to high variance made

discrimination of mean orientation more precise, and adaptation to mean orientation of 15° , which resulted in a tilt after-effect, led to overestimation of orientation variance (Jeong & Chong, 2020). The authors concluded that mean and variance representations are interrelated because perceived variance changed with perceived mean orientation.

Data visualization

Bar graphs, line graphs, and pie graphs are some of the oldest and most commonly used graphs (see Friendly, 2008, for history of data visualizations). However, modern technology has enabled us to collect more data and thus, necessitated innovations in graphs. In order to better convey trends and patterns in large data sets, researchers have been developing and testing new visualizations. *Graphical perception* refers to our ability to accurately interpret data values and different statistics (e.g., mean, variance, range) from visualizations (e.g., Cleveland & McGill, 1984; Saket, Srinivasan, Ragan, & Endert, 2018). Historically, research on graphical perception focused on people's abilities to accurately read data values from different graphs and to compare data values within a graph type. Data values are represented in graphs through *marks* (e.g., dots, line, bars), and the decoding of these marks are discussed through visual features (e.g., position, length). Though the reading of some marks may be intuitively attributed to certain visual features (e.g., dots = position, line = length), we later discuss that this is not necessarily the case for all tasks.

In early work, dots were used to study position, lines for lengths, and typically circles or rectangles for area. Cleveland and McGill (1986) used these marks for their pioneering work in graphical perception. In one of their studies, participants compared positions, lengths, angles, slope, and area by estimating the percentage of a target to a standard (Cleveland & McGill, 1986). Participants' performance from most accurate to least accurate was position, lengths, angles, and slopes, and finally areas, using significant differences to distinguish between rank levels. Heer and Bostock (2010) were able to replicate these rankings in a crowdsourced experiment. Subsequently, research on graphical perception looked to vision science for inspiration on frameworks and methodology to better quantify people's abilities, explain results, and give design recommendations.

Applications of vision science

Researchers often decompose graph reading into "elementary perceptual processes," such as detecting the orientation of a line for line graphs, the height of a bar in bar graphs, or the size of an angle in pie graphs (Larkin & Simon, 1987; Simkin & Hastie, 1987). It is upon these elementary perceptual

processes that researchers try to explain the advantages and disadvantages of using different graphs and data visualizations, and to make design suggestions.

Researchers incorporated psychophysics to create models to explain human performance and develop metrics to predict the effectiveness of a generated visualization. Experiments utilized more trials and adaptive staircase methods (e.g., Rensink & Baldridge, 2010), as well as different task types: discrimination trials (e.g., Rensink & Baldridge, 2010) and magnitude (re)production trials (e.g., Saket et al., 2018) to measure precision. Saket et al. (2018) found that participants' accuracy at adjusting interactive marks to match a reference followed the same ranking as in Cleveland and McGill (1986), except for a significant difference between length and angle: position, length, distance, angle, curve, area, and then shading.

The different marks used in graphs, such as dots, lines, and rectangles, have different levels of precision in their representations (e.g., Bair, 2005; Palmer, 2002). Starting from the higher end of precision, we have dots that represent spatial locations, then lines for lengths, and then rectangles for areas (e.g., Bair, 2005; Palmer, 2002). This order matches the accuracy levels found in comparing data values using the respective graph marks (Yuan, Haroz, & Franconeri, 2019). People are also better at estimating the average of the diameter (length) rather than the area of circles, and better at judging average diameter and average areas than judging cumulative diameter and cumulative areas (Raidvee, Toom, Averin, & Allik, 2020). Because discriminability of total area worsens as the number of items increases, these authors suggest that bubble maps use circle diameter as the dimension that is proportional to data values, since data values in bubble maps that are proportional to area may be less accurately perceived. Indeed, it is difficult to compare circles by area in bubble charts (Heer & Bostock, 2010), and bubble charts can be used to misrepresent data (Kong, Heer, & Agrawala, 2010).

More focus was put on measuring people's sensitivity to differences in visual features of marks. This focus allowed researchers to develop just-noticeable-difference formulas (e.g., Rensink & Baldridge, 2010; Szafrir, 2018) and quantitative metrics to advise on using visual features more effectively (e.g., as with color in Szafrir, 2018, and parameters of scatter plots in Micallef, Palmas, Oulasvirta, & Weinkauff, 2017). Modeling techniques are used to deduce the visual features used to accurately reading visualizations (e.g., Jardine, Ondov, Elmqvist, & Franconeri, 2020) and to isolate the contribution of different visual features (e.g., scale, density, curvature) on graphical perception (e.g., scatter plots: Sedlmair, Tatu, Munzner, & Tory, 2012; high-dimensional data visualizations: Bertini, Tatu, & Keim, 2011). These models then inform design decisions and recommendations.

Data value perception

The precision of reading values from a bar graph depends also on context. In most cases, it has been well documented that participants have a biased impression when judging heights of bars in bar graphs (Jarvenpaa & Dickson, 1988; Kosslyn, 2006; Peebles, 2008; Zacks, Levy, Tversky, & Schiano, 1998). Compared with line graphs and Kiviart charts (also called radar charts or spider charts; see Fig. 1 for an example), values in bar graphs are systematically underestimated (Peebles, 2008). However, bar graphs do prevail to line graphs, but only in a particular context. When relationships between three variables need to be conveyed, bar graphs are preferable to line graphs, since participants make fewer errors when interpreting those bar graphs (e.g., Ali & Peebles, 2013; Peebles & Ali, 2009).

Researchers have tried to infer what visual feature of marks people use to compare data values in different graphs based on their experimental results. For bar graphs, participants could be comparing the spatial locations of the top of bars, the height of bars, or the area of bars (assuming equal width). Yuan et al. (2019) concluded that people use a spatial position encoding rather than a length encoding because their participants performed as well at comparing pairs of bars in bar graphs as comparing pairs of data points in dot plots. Compared with pie graphs, bar graphs can be read faster and more accurately (Simkin & Hastie, 1987). This difference in performance may be due to the fact that any of the possible encodings of bars (spatial location, height, or area) can be compared directly, whereas pie angles are more difficult to compare with each other because the angles do not share a common starting

orientation as compared with the common starting level of bar heights. Results from Huestegge and Pötzsch (2018) support this explanation, since participants performed better (shorter response times and lower error rates) on proportion judgments with rectangles in tree maps than with slices in pie graphs. Bar graphs and tree maps provide some common point of reference for comparing their components. Bars in bar graphs start from the same baseline and rectangles in tree maps typically share some dimension with adjacent rectangles. The structure and organization of rectangles in tree maps may be essential to their readability. When squares or rectangles of extreme aspect ratios were used in tree maps, area estimation became less accurate (Kong et al., 2010). Eye-tracking data suggests another explanation for the performance difference between tree maps and pie graphs: greater ease with scanning in tree maps than in pie graphs (Huestegge & Pötzsch, 2018). Like with tree maps, there are other visual cues people could use to make proportion judgments in pie graphs. Proportion judgments in pie graphs could be made using area, arc length, or angle of pie slices. Skau and Kosara (2016) tested this possibility by designing graphs that separated the encodings of area, arc length, and angle. Performance was negatively impacted by providing only angle information, but not by providing only arc length or area. Even when area was hard to decode, like with thin donut graphs, performance remained good. These results led the authors to conclude that arc length is the most important encoding and angle least important. This makes sense from an evolutionary standpoint, since circumference and arc length may be more meaningful and commonly assessed, such as tracking the phases of the moon, whereas objects that

Fig. 1 Simplified examples of common data visualizations

resemble pie graphs may not naturally occur. Further investigation into the essential components of pie graphs revealed the importance of its off-centered representation of proportions (Kosara, 2019). The variations of pie graphs that used centered shapes to represent proportions produced more error and consistent overestimation, which Kosara (2019) believed was due to increased difficulty in make area comparisons.

Incorporation of ensemble perception

Research on ensemble perception can be applied to human–computer interactions to convey information more effectively. Data visualizations encompass a lot of the visual features studied in ensemble perception, such as dot size in bubble maps and orientation in weather maps (see Fig. 1 for examples). There is an undeniable parallel between perceiving summary statistics from visual arrays typical in ensemble perception, such as line length, orientation, and (circle) size; and perceiving summary statistics from objects typical of graphs, such as points, lines, circles, and rectangles. In fact, we observe similar findings with graphs as with similar features in ensemble perception studies. In particular, the factors that influence performance when comparing mean values in two visual arrays also influence performance when comparing mean values between two graphs. For example, in bar graphs, the number of bars and the amount of variation in their height influence participants' ability to discriminate between means (Fouriez et al., 2008). This is similar to differences in performance observed with different set sizes for visual arrays (i.e., greater uncertainty when there are only a few items in an array vs. many; Dakin, 2001). It is also similar to decreased performance when there is a lot of variation amongst items in the array (e.g., Chong & Treisman, 2003; Marchant et al., 2013). Ensemble perception plays a role in making insightful observations when reading graphs and making effective design decisions for presenting information. For example, one may quickly estimate the mean of a graphs' objects (e.g., dots, circles, bars) in order to create a reference point for making comparisons across other objects, which would aid in identifying key points of interest. Along the same line, presenters can use this knowledge to present key information in an intuitive way, thereby guiding the viewer to the relevant observations.

Ensemble coding tasks Researchers in data visualization have drawn upon ensemble perception research to categorize and understand the processes involved in interpreting data visualizations. Szafir, Haroz, Gleicher, and Franconeri (2016) proposed four different types of ensemble coding in data visualizations, which describe the different visual aggregations that a viewer may perform with visualized data: summary, identification, pattern recognition, and segmentation.

Summary tasks can be assessing numerosity in tagged text or assessing mean in bubble maps or weather maps. Identification tasks include locating outliers and extrema, and assessing range. Pattern recognition refers to trend assessment, such as correlation in scatter plots and silhouettes in weather maps. Segmentation refers to organizing data into subsets, such as clustering or categorization. Viewers' accuracy when performing these tasks may depend on the visual features used in the data interpretation.

Decomposing graph reading into elementary perceptual processes and attributing advantages and disadvantages of different graphs to different visual features (e.g., spatial location, length, area) is not as straightforward when the task requires integrating across or summarizing trends from the data (Carswell, 1992; Ratwani, Trafton, & Boehm-Davis, 2008). Likewise, the recommendation of a visualization for highlighting trends is not straightforward. The effectiveness of a visualization depends on the task (e.g., Albers, Correll, & Gleicher, 2014; Saket, Endert, & Demiralp, 2018). Saket et al. (2018) compared people's speed and accuracy at tasks, such as finding outliers, clusters, max/min, and estimating means and correlations, between tables, lines graphs, bar graphs, scatter plots, and pie graphs. They made separate recommendations depending on which statistical property one wants to highlight. For example, they recommended bar graphs for clusters, line graphs for correlations, and scatter plots for outliers.

Just as mean perception is a popular research topic in ensemble perception, people's ability to judge mean values from graphs is a popular research topic in data visualizations. Researchers have assessed mean perception in commonly used graphs—namely, scatter plots, bar graphs, and line graphs. When asked to judge the mean values in bar graphs and scatter plots, participants' estimates were systematically lower for bar graphs, but not for scatter plots (Godau, Vogelgesang, & Gaschler, 2016). The authors attributed this finding to the fact that scatter plots have data values represented at different spatial locations while bar graphs represent the same data values, but in a less precise dimension: area. It is because of this imprecision, the authors believe, that made participants systematically underestimate mean values from bar graphs. People are exceptionally good at judging the average value in scatter plots, even with increasing set size, redundant and conflicting encodings and additional sets (Gleicher, Correll, Nothelfer, & Franconeri, 2013). This ability could be explained by the fact that centroid localization—the ability to accurately locate the center of gravity in a dot cluster (Badcock & Westheimer, 1985; D. Whitaker & Walker, 1988; Wright, Morris, & Krekberg, 2011) is a visual hyperacuity, where performance on tasks have substantially lower thresholds than the grain of the receiving layer (Westheimer, 2010). When asked to judge whether a value was larger or smaller than the average of the graph, participants only systematically underestimated values for bar

graphs, but did not have a systematic bias with Kiviat charts (also called radar charts or spider charts; see Fig. 1 for an example) or line graphs (Peebles, 2008).

With which visual features (e.g., spatial position, length, area) people use to encode marks for summary tasks may be different than those used for comparing data values. Yuan et al. (2019) concluded that participants use a spatial position encoding when comparing data values in a bar graph. However, their results from a summary task suggest that participants use a different encoding. For the summary task, participants were asked to compare the average values between two groups of bars, whether the two groups have the same number (i.e., six vs. six) or different number (i.e., six vs. 10) of bars. The authors concluded that participants are using length encodings for their task. For one, participants' performance was worse than what would be expected if they still used spatial locations to estimate the averages. Secondly, the comparison condition of unequal number of bars helped discount area encoding as a possible interpretation. Because unequal number of bars makes greater cumulative area a poor proxy for greater average, we can be more certain that they did average. However, averaging here may be a more unnatural task than summation because the adjacency of the bars may encourage the perception of one object instead of an ensemble of objects.

Given the ubiquitous use of scatter plots and histograms for teaching statistics, the findings discussed above should be put into consideration when lesson planning and kept in mind when addressing students' misconceptions. Beyond being able to extract summary statistics and identify trends and patterns in graphs, statistics students also need to learn how to interact with inferential uncertainty.

Insights from uncertainty perception

Uncertainty perception refers to people's perception of variance in data visualizations and their judgment of the (un)certainly of trends or predictions displayed in data visualizations. Many real-world problems require uncertainty to be represented in data visualizations. For example, viewers need to be reminded that weather and storm forecasts are predictions and that there is some level of uncertainty. Conveying uncertainty in data visualizations is a difficult problem for visualization designers, who have used very diverse approaches. For example, in the context of meteorology, contour lines have been used to convey atmospheric pressure predictions (Stephenson & Doblas-Reyes, 2000) and arranged to showcase median, spread, and outliers for temperature fields (R. T. Whitaker, Mirzargar, & Kirby, 2013). An ensemble of spatial positions with concentric confidence intervals has been used for hurricane prediction plots (L. Liu, Mirzargar, Kirby, Whitaker, & House, 2015). However, usage of spatial spread can mislead viewers. For example, when an uncertainty cone is used for hurricane forecasts, viewers had misperceptions

that storm size increases with time and overestimated the likelihood that a storm will occur within the cone (Broad, Leiserowitz, Weinkle, & Steketee, 2007). L. Liu et al. (2017) used confidence interval bands (colored superimposed circles), splats, and icons to display storm predictions. They found that icon visualizations were better than the others for time-specific visualizations. Participants also estimated risk (an integration of likelihood and intensity) more accurately with icon visualizations.

These studies suggest that ensembles of discrete marks convey information about variance and uncertainty better than a (continuous) mark that summarizes the data set. The benefit of using ensembles of marks may derive from our abilities for ensemble perception. Each mark maps to a data point, the possible outcomes are visible, and the variance can be extracted via ensemble perception. On the other hand, for summary displays, a more abstract notion of variance or uncertainty is used, and the viewer needs to make inferences about the originating distribution of data points using the representation of variance or uncertainty (e.g., as from error bars in bar graphs). These displays may also require more basic knowledge about statistics. The advantage of using ensembles of discrete over continuous marks that represent data sets relates to the common finding that people reason about probabilities better when probabilities are presented as discrete events (see McDowell & Jacobs, 2017, for meta-analysis).

The literature on uncertainty perception in data visualizations could be a valuable contribution to resolving some of the problems in statistics education. Results from research on uncertainty perception may highlight the undeniable visual aspect of learning statistics and predispositions people have when reading data visualizations. Inferential statistics is commonly taught in introductory statistics courses (Castro Sotos et al., 2007; Lavigne, Salkind, & Yan, 2008). Results of hypothesis testing are often reported using confidence intervals or shown in a bar graph where error bars represent standard errors. However, research shows that error bars and confidence intervals are often misinterpreted. People tend to find it more likely for data points to fall within the error bar inside the bars of bar graphs than within the error bar outside the bars (Newman & Scholl, 2012)—an effect the authors attribute to object perception. People also assume that overlapping error bars mean non-significant difference (Cumming, 2009) or all values within the error bars are equally likely (Ibrekk & Morgan, 1987). Though many of the participants in these studies were crowdsourced or had little or no statistics background, experts can also have problems with error bars. For example, Belia, Fidler, Williams, and Cumming (2005) recruited authors from leading journals in psychology, behavioral neuroscience, and medicine to adjust means and error bars in a graph to represent statistical significance. The researchers found that the experts had misconceptions about how error bars relate to statistical significance and how confidence intervals and error bars differ.

Error bars and confidence intervals may be difficult to interpret because they represent the variation of the population parameter and, as such, appear narrower in variation compared with other representations of variation. On the other hand, prediction intervals represent a range of predicted values for new observations and relate more directly to individuals in the population. When confidence intervals (CIs) are used in a graph to show two treatment options, participants tended to overstate the superiority of a treatment option, be more willing to pay for that option, and understate the variability of outcomes (Hofman, Goldstein, & Hullman, 2020). However, when the axes of the CI graph were rescaled to match those of the prediction interval (PI) alternative, in which variation of individual outcomes are represented, these biases were reduced (Hofman et al., 2020). This rescaling made the perceived difference between treatment options smaller, which translated to less certainty that the two options were significantly different. In that study, the visualizations that encoded the variation in individual outcomes: prediction interval and hypothetical outcome plots (HOPs: an animation of discrete outcomes) yielded more accurate judgments from participants.

Error bars do not represent the underlying distribution. Because of how error bars look, many people think that error bars represent an interval of uniform probability (Ibrekk & Morgan, 1987). Commonly recommended alternatives to error bars include the violin plot and gradient plot (e.g., Correll & Gleicher, 2014). Violin plots are believed to be easier to interpret correctly because they encode probability density as mark width (Barrowman & Myers, 2003; Kampstra, 2008). Gradient plots encode uncertainty using degree of transparency/opacity (Correll & Gleicher, 2014). These visualizations capitalize on our capacity to average line length and hues, respectively. Correll and Gleicher (2014) compared visualizations that encoded margin of error either binarily (error bars, box plots) or continuously (violin and gradient plots) and found that people were overconfident about their judgments of effects when using error bars compared with the other plots. Hullman, Resnick, and Adar (2015) compared people's abilities to estimate properties of distributions (mean, probability of outcome above a threshold, probability of outcome between two thresholds) from error bars, violin plots, and HOPs. They found that people had more accurate estimates for multivariate distributions of two and three variables (either independent or correlated) when using HOPs. The authors believed HOPs were helpful because they allowed the use of counting and probabilities to infer properties of distributions. However, one drawback of HOPs was that participants had worse estimates of mean when the distribution had high variance. Since HOPs display discrete outcomes over time and we have the capacity of ensemble perception across time (e.g., Hubert-Wallander & Boynton, 2015), it is not surprising that Hullman et al. (2015) found this, since estimation of mean has

also been found to be worse when arrays have higher variance in the ensemble perception literature (Dakin, 2001; Fouriez et al., 2008; Haberman et al., 2015; Im & Halberda, 2013; Morgan et al., 2008; Solomon et al., 2011).

The proposed alternatives to error bars incorporate information about point estimates and probabilities. When displaying uncertainty in visualization, people make more optimal decisions when they have both point estimates and probabilities (Joslyn & LeClerc, 2013). Both pieces of information are important. Hullman et al. (2015) find that separating the visual marks that encode the underlying data from those that encode uncertainty may encourage participants to favor information about central tendency and consider probabilistic information less. Thus, it is recommended that visualizations of uncertainty be *intrinsic* to the representation instead of an extrinsic annotation of a distribution's property (e.g., error bars or marks that locate mean, median, or mode; Kay, Kola, Hullman, & Munson, 2016). For example, in a density plot, mode and probability density are intrinsic to each other because the mode is represented by the peak of the density. Kay et al. (2016) compared people's probabilistic estimates using density plots, stripe plots, and dot plots (20 or 100 dots). They found that people's probability estimates were less variable when dot plots used easily countable quantities (20 dots), and that performance using dot plots with 100 dots was similar to density plots. The authors believed the countability of the dots made it easier to reason about hypothetical outcomes. These dot plots may be beneficial because they tap into our capacity for numerosity, which is less precise at larger numbers like 100 (Landy, Silbert, & Goldin, 2013) and for reasoning about probabilities in terms of frequencies.

Our survey of data visualization literature aims to highlight the importance of considering visual influences on reading graphs and identifying trends and patterns within them. One main takeaway is the limitations of graphical perception of bar graphs, which have direct implications for understanding students' misconceptions about histograms that are discussed in the next section. Other takeaways include the misperceptions of the graphical representations of inferential uncertainty that are typically used in statistics classrooms (i.e., error bars, confidence intervals) and the assistance that ensemble representation has on accurate assessment of variance and uncertainty.

Statistics education

In the age of big data, it is more important than ever to have statistical literacy when entering the job market. Colleges prepare the next generation of workers by requiring one or more statistics courses in such majors as psychology, economics, and biology. However, many undergraduates still struggle to apply what they learned in their introductory statistics courses for real life problem-solving (e.g., Castro Sotos et al., 2007;

delMas, Garfield, Ooms, & Chance, 2007; Lavigne et al., 2008). One contributing factor to students' inability to apply these statistical concepts may be shallow or misunderstanding of statistical concepts that arise from not reading graphs properly. A lot of statistics instruction depends on the use of graphs, such as scatter plots, bar graphs, and most importantly, histograms. While most students likely encountered scatter plots and bar graphs in their K–12 education, most students are likely encountering histograms for the first time in an introductory statistics course.

In our discussion of literature on statistics education, we will focus on histograms because they are a fundamental component of statistics education and a commonly studied graph by researchers in statistics education. Histograms was also selected because they are not intuitive to our visual system as with scatter plots and is, therefore, a research topic that could benefit from insights from visual perception. While researchers in statistics education often take a cognitive perspective of understanding the difficulties students have with histograms, we want to start a discussion on possible perceptual difficulties. While many graphs may be difficult for students to interpret (e.g., Ali & Peebles, 2013), our discussion on histograms serves as an example of the types of conceptual problems that may arise from perceptual difficulty.

An example of using graphs: Histograms

The field of statistics aims to transform concrete or qualitative characteristics of the world into quantifiable measures and convert this raw data into numerical or visual summaries that help make inferences and conclusions about the world. This transformation of the concrete and observable into the abstract and intangible is perhaps the most challenging for students to understand and grasp fully. Transnumeration, or transformations that change data representation, has been deemed to be a fundamental framework for statistical reasoning and better understanding of variation, distribution, and other important statistical concepts (Wild & Pfannkuch, 1999). One example of these transformations is the histogram, which converts individual data points into a graphical representation.

Histograms are used as a graphical representation of the shape and variability of a distribution and a visual tool for explaining many statistical concepts, such as spread and skewness. The understanding of histograms has also been regarded as necessary for developing the understanding of density curves of distributions (delMas, Garfield, & Ooms, 2004). In fact, histograms, as a graphical representation, provides perhaps the most natural transition to teaching the concept of density curves as they already use area to represent proportions (delMas et al., 2004). The following are common *misconceptions* about histograms:

- (1) Histograms are (like) bar graphs (Biehler, 1997; Cooper & Shore, 2010; Meletiou-Mavrotheris & Lee, 2010).
- (2) Histograms display raw data, with each bar representing an individual observation rather than a grouped set of data (Lee & Meletiou-Mavrotheris, 2003).
- (3) Histograms display two-variables like scatter plots or time sequence plots (Lee & Meletiou-Mavrotheris, 2003).
- (4) The y -axis (where frequency is displayed) is used when finding the mean and mode of the distribution (i.e., mode is the y -value of the highest bar in the histogram; Kaplan et al., 2014).
- (5) Differences in the heights of bars (y -axis) can be used to compare the variation of two histograms (Lee & Meletiou-Mavrotheris, 2003).
- (6) Flatter histograms have fewer variable data (Cooper & Shore, 2008; Kaplan et al., 2014; Meletiou-Mavrotheris & Lee, 2002).

These misconceptions about how to read histograms are especially alarming because being able to correctly interpret histograms is integral to understanding key concepts in statistics, such as central tendency. Misconceptions about histograms may contribute to misconceptions about summary statistics.

Misconceptions about summary statistics

Central tendency Students have misconceptions about measures of central tendency: mode (Huck, 2009), median (Cooper & Shore, 2008), and mean (Olani, Hoekstra, Harskamp, & Van der Werf, 2010). When asked to identify the three measures of central tendency from a histogram, where frequency information is displayed on the y -axis, students make some common errors that stem from misinterpreting the information portrayed by the axes (reading the y -axis as data values and not frequency) and by the bars.

Some students misinterpret mode, or the data values that appear the most often in a data set, as the height of the highest bar, giving the y -value as the mode (Huck, 2009; Ismail & Chan, 2015). Others give the number that appears the most in the axis labels (Ismail & Chan, 2015).

For the median, students know that it refers to the midpoint of some numbers, but some students give the median of the frequencies (or the middle bar), while other students give the median of the numbers that appear on the x -axis (Cooper & Shore, 2008; Ismail & Chan, 2015). Some make procedural errors, like knowing to divide some number by two to find a midpoint, but choose an irrelevant set of numbers to use and/or forget a step or two. For example, they may count the number of data values (i.e., irrelevant set of numbers) on the x -axis and divide that by two or add the frequencies and divide that by two, but forget to use that

answer to find the corresponding data value (Ismail & Chan, 2015). These kinds of errors showcase the complete misinterpretation of a histogram. Those that correctly add up all the frequencies and find the data value corresponding to the midpoint of the sorted data values are the minority: less than 10% in Ismail and Chan (2015).

Out of the three measures of central tendency, the mean is the most computationally intensive and perhaps the most dependent on understanding what exactly is portrayed in the histogram. It requires a series of multiplications (each pair of data value and its frequency) and additions and finally a division (number of data points). Ismail and Chan (2015) observed eleven different computational errors. These include not using multiplication, using multiplication but with irrelevant numbers or using irrelevant numbers as divisors. These errors all stem from misinterpreting the axes and misunderstanding what is portrayed by bars.

Mean Precollege and college students have many misconceptions about the concept of average (e.g., Cai & Moyer, 1995; Mevarech, 1983; Pollatsek, Lima, & Well, 1981; Strauss & Bichler, 1988). In particular, it appears that while students know how to compute the average when given a set of numbers, they do not necessarily know how to apply the concept of average in word problems. For example, in Cai & Moyer (1995), middle school students were given a word problem: there are five people who have blocks, given the average number of blocks across the five people and the number of blocks that four out of five total people have, how many blocks does the fifth person have? Some students incorrectly add up the number of blocks from the four people and the average number and divide by five (Cai & Moyer, 1995). Misconceptions about the concept of average can be remedied through instruction. After instruction, Batanero, Cobo Merino, and Diaz (2003) found significant improvement in the understanding that the mean value may not exist as a data point and how to use the mean algorithm to solve word problems like the one described above.

Variance Variability refers to the variable attribute of an entity (e.g., that a population has variability in feature x), whereas variation, or variance, assesses or quantifies that attribute (e.g., as a summary statistic; Reading & Shaughnessy, 2004). The understanding of variability and variance is very important for statistical reasoning, as it lays the foundation to understanding distributions and statistical inference, such as hypothesis testing. However, students often have misconceptions about variability and variance, and reason about them incorrectly.

One common mistake involves the misinterpretation of histograms—looking at the y -axis for information about variability instead of the x -axis (Lee & Meletiou-Mavrotheris, 2003). As one might expect, misinterpreting the information portrayed in a histogram makes it very difficult for students to

reason about variability correctly. For example, some students (~27%) may use the fact that the heights of bars in a bell-shaped histogram are more variable than those of another bell-shaped histogram to conclude that the former has greater variability (Cooper & Shore, 2008). Some students mistakenly connect the concept of variability with a concept more familiar to them: range. For example, some students (~20%) misunderstand ‘spread’ (commonly referring to variance) to be the range of data values on the x -axis and conclude that any two histograms with the same range also have the same amount of variability regardless of other features of the distribution (Cooper & Shore, 2008).

Another common mistake is using irrelevant features of a distribution to make conclusions about a distribution’s variability. For example, some students focus on the differences between the heights of bars in histograms, believing that variability depends on the number of data points, or on the irregularity of the shape of the distribution (Inzunza, 2006). One of the reasons why students use irrelevant features to reason about distributions may be this: Novices have a tendency to be drawn to irrelevant features of the graph not only for histograms, but also for bar graphs (e.g., Ali & Peebles, 2013; Okan, Galesic, & Garcia-Retamero, 2016).

Just as students have difficulty using relevant features to reason about a distribution’s variability, students also struggle with reasoning about a distribution’s standard deviation. In particular, they have a difficult time identifying what information from a histogram they can use to assess standard deviation and how standard deviation relates to other measures. For example, when high school students were shown two histograms, one uniformly distributed across x -values and another bell shaped, some students used total frequency or the mean to assess whether two graphs have the same standard deviation (Chan & Ismail, 2013). Using such features of a distribution to reason about its standard deviation suggests a deep misunderstanding of what standard deviation represents. Student’s lack of conceptual understanding of standard deviations is well documented (e.g., Garfield & Ben-Zvi, 2008). It is commonly the case that students know how to calculate the standard deviation from a data set but do not know how to interpret the number as a result of the computation (Garfield & Ben-Zvi, 2008). These misconceptions about variance are concerning because many researchers agree that the concepts of variability and distribution support each other in students’ learning of statistics (Bakker, 2004; Ben-Zvi, 2003; delMas & Liu, 2005; Makar & Confrey, 2003, 2005; Pfannkuch & Reading, 2006; Wild, 2006).

Perceptual barriers to reading histograms

Though we do not deny the possibility that students’ difficulties with histograms may have predominantly cognitive roots, we do think it is important to consider perceptual explanations

because this is typically an overlooked perspective amongst researchers in statistics education. While a lot of discussed misconceptions about histograms seem to stem from not knowing which axis to use to find certain information, it is possible that the confusion is not that students do not remember but that the dimensions are so counterintuitive that they have trouble parsing the information that is there, falling back on what feels more intuitive—read it like a bar graph.

Perceptual difficulties should not be overlooked because they can even affect statistics experts. Statistics experts (e.g., graduate students, professors, researchers) seem to be susceptible to influences from the vertical dimension as well. This was suggested by worse performance (i.e., more errors, slower response times) consistent with using the *height heuristic* (i.e., histograms with taller bars have greater means) for comparing histograms with the same mean and shape of distribution, but with scaled up heights of bars, but not with other types of histogram pairs (Lem, Onghena, Verschaffel, & Van Dooren, 2014). Even with experts, we see some level of perceptual interference with statistical reasoning. Lem et al. (2014) found that experts had slower response times when height information could bias one to an incorrect response (incongruent trials) than when height information was sufficient for the correct response (congruent trials). This pattern of slower response times for incongruent trials is characteristic of the Stroop effect. The Stroop effect has been found for incongruence between the magnitude of numbers and their font size in numerical magnitude comparison tasks (e.g., when comparing 3 with 5, slower to select 3 as being the smaller number if it is printed in larger font; Henik & Tzelgov, 1982). In Lem et al. (2014), choosing the correct answer in the histogram comparisons may have required the suppression of irrelevant or distracting information and reasoning beyond the perceptually intuitive answer. However, that study also highlighted the importance of conceptual understanding—the undergraduates did not perform any better when they were given unlimited time to make the comparison and warned about potential fallacies in graphs.

Literature on ensemble perception and data visualization could provide a basis for understanding what makes learning about distributions from histograms so difficult. Research with bar graphs suggest that people treat bars like objects and judge data values that are within the bars as more likely to be in the data set than values outside of the bars (Newman & Scholl, 2012). Likewise, people may be treating bars within histograms like objects. When looking upon a collection of rectangular bars in a histogram, we may automatically perceive the average height of the bars, which provides no meaningful information about the distribution and is an inappropriate interpretation of the histogram. Existing perceptual tasks for interpreting histograms may require people to carry out a series of explicit, nonautomatic steps. For example, in order to locate the mean and median within a histogram, we may need

to reconceptualize the bars, perceive cumulative area, and perform a mental division of that space. Unfortunately, the perception of cumulative area is less accurate than the perception of the mean (Raidvee et al., 2020) and perhaps less intuitive for histograms. Some researchers (e.g., Kahneman, 2011) have postulated that the imprecision of perceiving cumulative area may be created through taking the perceived average, a measurement that is more accurate, and using multiplication to reconstruct cumulative area. This explanation presumes that the observer can multiply or divide sensory magnitudes, which have been suggested by some experimental results (Torgerson, 1961). Regardless of the explanation, histograms may be difficult to interpret correctly for many reasons: bias to the vertical, perceiving bars as separate objects representing one data point each, and mostly importantly, difficulty of cumulative area perception, or the relevant dimension for correct histogram interpretation.

Inspiration for solutions

One of the greatest mysteries regarding students' difficulties with learning statistics is how foreign they find the statistical concepts being taught. It could not be because they have not interacted with summary statistics, like mean and variance, or experienced distributions of examples in their everyday experiences of the world. Their visual systems automatically extract information regarding mean and variance to represent their environment. But somehow, when asked to reason explicitly about means, variances, and distributions, they lose all intuition. Perhaps some of this disconnect can be explained by the level of abstraction used in the teaching of statistics, which deals with numbers and formulas rather than relatable experiences.

One way to approach this problem is through a grounded cognition perspective on education: that environment and bodily experiences are important for developing cognitive processes (e.g., Barsalou, 1999). By creating a learning experience that is close to natural everyday experience, we may engage the intuition our visual system has about summary statistics and support the learning of statistics. Simulations and animations have been shown to be helpful for new statistics students (Neumann, Neumann, & Hood, 2011; Wang, Vaughn, & Liu, 2011). The usefulness of animations for learning statistics is unsurprising given that animation was helpful for conveying information about uncertainty to people without instruction (Hofman et al., 2020).

Simulations have also been effective at correcting misconceptions and improving understanding of correlation (T.-C. Liu, Lin, & Kinshuk, 2010). While computer simulations of sampling distributions have been shown to be effective for teaching the concept (e.g., Jamie, 2002; Lane, 2015), the use of hands-on activities prior to the use of computer simulations yielded better learning (Hancock & Rummerfield, 2020). There is also work leveraging our capacity for ensemble

perception on improving statistical reasoning. For example, Yu, Goldstone, and Landy (2018) improved students' statistical reasoning with sample size, variance, and difference between means by using an experientially grounded learning intervention that involved representing data points as different colored balls and using animation to illustrate the factory production of these balls.

Using these previous successes as inspiration and insights from ensemble perception and data visualization, we propose potential solutions for students' difficulties with histograms in the corresponding future directions subsection.

Future directions

This tutorial review highlighted the gaps in ensemble perception literature, the recent usage of ensemble coding principles within the data visualization literature, and the hindrance histograms may pose to students' learning and reasoning within statistics education literature. Within the ensemble perception literature, investigations on mean perception are abundant but those of variance perception are scarce. While the first study on variance perception was published more than a decade ago, it has not gained its deserved traction amongst researchers. Insights drawn from studying variance perception could provide context to the existing literature on mean perception, and in turn help us understand mean perception better. It may make more sense to study some visual features because of their ecological relevance (e.g., color hue for food, and facial expressions for social interactions) and therefore they should be given higher priority due to their potential for insights into human vision and behavior.

While researchers of data visualization have long manipulated graphical features to study their effects (e.g., 2D vs. 3D: Zacks et al., 1998; vertical vs. horizontal bars: Fischer, Dewulf, & Hill, 2005), the use of ensemble perception literature can provide more theoretical grounding for manipulating visual features in future studies and extend the literature to include overlooked or understudied visual features (e.g., dot size, color saturation) within graphs. Researchers can draw many similarities between the tasks used in ensemble perception research and those performed when interpreting trends from data visualizations (e.g., locating the centroid, or average spatial location, of a dot cloud: Gleicher et al., 2013), and use those similarities to hypothesis-test the effectiveness of different data visualizations. They could also directly test people's performance on these visual tasks and how related perceptual fluency is to trend detection and accurate data interpretation. Given the listed opportunities to gain new insight, we hope that researchers continue to apply ensemble perception literature to research in data visualization and that this approach becomes more popular.

More importantly, the knowledge from ensemble perception can be used to propose new visual representations for conveying statistical properties about data and for teaching statistical concepts in the classroom. A major consideration for design decisions lies within our natural perceptual abilities, opting to align our perceptual strengths with the type of visual aids used and the types of visual tasks involved and to avoid our perceptual weaknesses or imprecisions. These perceptual abilities should be taken into consideration when creating graphs for scientific communication and education purposes, but not without consideration of the perceptual task (e.g., comparing values, finding averages, identifying outliers) as well. Visual aids should be designed to be as natural and straightforward for our visual system to perform the task of interest as possible, leaving the difficulty of interpretation to conceptualizations and reasoning rather than to perception.

The literature on ensemble perception and data visualizations also provide a context for understanding the difficulties students face when learning statistical concepts, especially since statistics education depends on the use of graphs. Some barriers to learning and reasoning about statistics may stem from automatic perceptual processes, such as being drawn to extreme values or representing sets of visual stimuli with summary statistics (e.g., the average height of bars within a histogram is more automatic than the estimation and manipulation of cumulative area).

Research on ensemble perception has significant theoretical and practical implications for more applied areas of study, such as data visualization and statistics education. Researchers pursuing more applied research questions can harness knowledge from ensemble perception to inspire new lines of research and new approaches to solving existing problems. Some integration of knowledge within these three areas of study has the potential to reveal new and impactful insights that would be of interest to other areas. The following are some specific suggestions for future research in each of the three areas, drawing connections to the other areas where appropriate.

In ensemble perception

The discovery that people are sensitive to variance in visual displays is surprising. Outside of a statistics classroom or context, people do not usually comment on the variance between things, even though they use "average" as a descriptive adjective often. Inside a classroom, students struggle to fully grasp the concept of variance, often unable to reason about variance accurately. Yet our visual systems seem to be able to extract, reproduce, and discriminate between variances. This is puzzling as well as amazing. Therefore, research that can demystify our ability to perceive variance deserves much attention. But first, is the information being processed really variance or do we rely upon proxies that our visual systems have learned to use due to their reliability in natural environments (i.e., high

correlation between variance and the proxies)? Recent work suggests that people rely upon range to make judgments about variance because performance became worse when there was incongruency between range and variance (Lau & Brady, 2018). Future research should continue to test other explanations for accurate performance. If we conclude that our visual system is in fact performing computations of the variance, are all items within an ensemble equally likely to be incorporated into variance estimation? Or do we use smart subsampling strategies (as with mean: e.g., Marchant et al., 2013)? What assumptions are our visual system making? Most visual features in the world are normally distributed, is this a prerequisite for being able to perform variance discrimination tasks? Or would we be able to perform equally as well on unnatural or asymmetric distributions (e.g., gamma distributions)? For example, two arrays can have the same variance, but wildly different compositions of values and shapes of distributions. Finally, there needs to be more work on how mean perception and variance perception relate to each other, if at all. Do they rely upon similar mechanisms or share connections in their representations? If cued for a variance discrimination task, would we be able to access representations of mean after the stimulus is no longer in view? Or would performance significantly suffer?

In data visualization

Results from ensemble perception research can inspire research questions and practical solutions for data visualization and statistics education. Given our efficiencies at extracting summary statistics from different visual features, would certain data visualizations be more effective at conveying certain messages (e.g., highlighting variation in a data set) because of the visual features used? Would the optimal data visualization depend on whether the message is about relations/concentration (e.g., via color saturation) or quantity (e.g., via numerosity or size)? These design considerations are worth exploring because low-level visual features, such as color, can make a difference in the speed of interpreting trends in heat maps (e.g., dark is more: Silverman, Gramazio, & Schloss, 2016). However, performing visual tasks in ensemble perception experiments is not exactly synonymous to performing visual tasks with data visualizations and educational graphs like histograms. For one, typically, objects in ensembles are separable (e.g., shapes are not touching each other or overlapping; cf. Tong et al., 2015), while the objects used in data visualizations and statistics education (e.g., circles in bubble maps, rectangles in heat maps and histograms) are connected/laying adjacent or overlapping, which may make ensemble perception more difficult. Second, the graphical objects (e.g., axis labels, legends, text within the graph) may be distracting for performing summary tasks (e.g., perceiving mean or variance), so perceptual fluency may not

always translate to accurate data interpretations. While research on ensemble perception could give suggestions of visual features to use for data visualizations, it is still necessary to test the effectiveness empirically because of the inherent differences in the tasks and the incorporation of higher-level cognition in interpreting the data.

One data visualization that deserves a lot more attention from researchers in data visualization is histograms. While literature from statistics education has highlighted the common misconceptions students have regarding histograms, we do not know how accurately histograms can be perceived when read properly. Can people accurately estimate the mean and median from histograms, assuming the distributions are unimodal? Does the skewedness of the distribution affect people's ability to make these estimations? Perhaps estimating and segmenting cumulative area across bars of different heights is difficult. Would people do better if they were provided a density plot instead? These questions are important because if we are to use these graphs to teach statistical concepts, we need to know what students are capable of *seeing* in the graphs.

In statistics education

Histograms are essential to teaching statistical concepts related to distribution, but are commonly misinterpreted by students. While histograms are intended to be used as a visual aid for statistical reasoning, they are more often a barrier to it. Histograms could be redesigned to play to our perceptual strengths, like perceiving numerosity and averages, and prevent the misattribution of the bars to objects. This is not to say that the current version of the histogram needs to be replaced completely, but the problem could be addressed at different stages to see what is the minimal intervention necessary to make judgments from histograms more intuitive and less error prone. We can teach the statistical concepts related to distributions using dot frequency graphs (see Fig. 2a) and use this simpler visual representation for statistical reasoning exercises. Since prior research have shown that simulations and animations are helpful for learning statistics (Jamie, 2002) and estimating variance (e.g., Hofman et al., 2020), we could also try an animated version of this graph. Once extracting the relevant information from these graphs becomes more natural, we can introduce the bars as binning the previously presented individual data points into adjacent groups (see Fig. 2b). Although the use of dot frequency graphs may seem elementary, not all students have seen or used dot frequency graphs by the time they enter college. Dot frequency graphs also make good entry-level graphs and provide a natural transition into using bars in histograms. The most obvious limitation of these graphs is their smaller capacity to deal with large data sets due to our representation of numbers becoming less precise as we jump orders of magnitudes (e.g., from representing 10s to 100s or 1,000s of data points; Landy et al., 2013). Some

Fig. 2 Example of suggested alternative to histograms, utilizing our ability to perceive numerosity and average color hue. Variations of (a) can be tested to optimize effectiveness: having columns of dots be closer together or touching, have space in between dots in each column so they are not touching, having subtle hue differences within columns for more

fine-grained representation of data values and for better transition to the concept of binning data values. A hybrid between dot frequency plots and histograms, such as in (b), can be used when transitioning from dot frequency plots to histograms. (Color figure online)

structured perceptual training may speed up the transitions between visual aids, the adjustment against perceptual biases and the maintenance of perceptual fluency with the graphs.

It is important that perceptual training is accompanied with explanations of corresponding concepts. The lack of explicit instruction and self-reflection can have consequences. While extensive training in one's estimation of correlation from scatter plots was successful at curbing perceptual biases, students were unable to convert their perceptual learning into explicit knowledge, unaware of the contributions of the x -and- y -value variances to correlation and thus not able to identify the formula for correlation (Cui, Massey, & Kellman, 2018). In that study, despite being shown an example of a scatter plot with a negative correlation prior to training, students failed to transfer their improved estimation ability to negative correlation plots from positive correlation plots.

There are many exciting directions for research in the three reviewed research areas: ensemble perception, data visualization, and statistics education. While research in these three areas can certainly progress without the others, each area stands to benefit from being aware of the others, for inspiration, direction, and/or application.

Conclusion

The field of ensemble perception has significant implications and practical applications for research on data visualizations and statistics education. The visual features studied in ensemble perception (e.g., size, hue) are used in common data visualizations (e.g., heat map, tree map, bubble map; see Fig. 1) and common graphs used to teach statistics (e.g., dots in scatter plots, rectangles in histograms). Researchers have recently

started to leverage the knowledge generated from research on ensemble perception to assess the effectiveness of different data visualizations on delivering key information and relevant trends to viewers, and to give design recommendations backed by research (e.g., Gleicher et al., 2013; Szafr et al., 2016). The application of ensemble perception literature is scarcely used in the domain of statistics education, but deserves more attention as it has been shown to be successful in improving statistical reasoning (Yu et al., 2018).

By comparing findings from ensemble perception and from statistics education, we can appreciate the similarities and differences in the factors that influence the perception and the conceptualization of mean and variance. Some of the misconceptions about mean and variance could be better understood when we look at how we naturally interact with that summary information. The mean of an ensemble is particularly meaningful for category learning, where we summarize a group of instances and store that summary to use as a reference for later categorization. What is stored is physical object(s) representing that category. This may be why we have a tendency to report an object with average visual features to be part of the previously encountered set of objects even though it did not exist (de Fockert & Wolfenstein, 2009) and why students believe that the average of a data set exists within that set (Batanero et al., 2003).

Unlike the mean, the variance of a set cannot be an object. Instead, what people may be doing is storing a set of objects that represent the variance in some way. Lau and Brady (2018) propose that people may be using range as a proxy for variance. Unlike variance, range can be stored as physical objects: minimum and maximum. Range also seems to be meaningful and intuitive to students. Many ways in which students misinterpret histograms involve range in some

fashion, such as using the range of x and y values to find the median of a distribution (Ismail & Chan, 2015). We can also anticipate there being differences because ensemble perception involves physical objects whereas the formal statistics used in classrooms and research involve mostly numbers. For example, when estimating the average of a set, people have a tendency to overweight outliers with a set of numbers, but underweight outliers with a set of dots (Anderson, 1968). While our visual system may have a tendency to disregard outliers (e.g., Haberman & Whitney, 2010), the explicit assessment of numbers in a set may make outliers more salient, pulling one's estimate of the mean towards the outlier.

Relating how people perceive summary statistics from a scene to how students think about summary statistics should not be overlooked because statistics education relies on graphs to demonstrate key statistical concepts and a lot of these graphs represent data via ensembles of objects (e.g., bars, dots). It is important to put (ensemble) perception into consideration when designing visual aids and curriculum for statistics classrooms because people are susceptible to perceptual biases and some features are easier or more intuitive to process than others (e.g., bars in bar graphs are perceived as objects, and therefore their size, or area and length, is easier to encode than spatial position—though not accurately; Yuan et al., 2019). Identifying perceptual biases matter because perceptual biases interfere with the development of conceptual understanding. Instead, we should think about how we can use perceptual features to support the learning of statistical concepts. That is where knowledge from ensemble perception comes in. A better understanding of ensemble coding of visual features commonly used in graphs can help us identify why certain graph types are hard to interpret or easy to misinterpret, as well as how to avoid such misinterpretations. This knowledge could also be used to design visual aids that have better learning outcomes than preexisting ones, potentially changing the way statistics is taught.

Open practices statement This manuscript has no associated data and no experiments to preregister.

References

- Albers, D., Correll, M., & Gleicher, M. (2014). Task-driven evaluation of aggregation in time series visualization. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, 551–560. Toronto, Ontario, Canada: ACM Press. <https://doi.org/10.1145/2556288.2557200>
- Ali, N., & Peebles, D. (2013). The effect of gestalt laws of perceptual organization on the comprehension of three-variable bar and line graphs. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 55(1), 183–203. <https://doi.org/10.1177/0018720812452592>
- Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research*, 83, 25–39. <https://doi.org/10.1016/j.visres.2013.02.018>
- Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15(3), 122–131. <https://doi.org/10.1016/j.tics.2011.01.003>
- Anderson, N. H. (1968). Averaging of space and number stimuli with simultaneous presentation. *Journal of Experimental Psychology*, 77(3, Pt.1), 383–392. <https://doi.org/10.1037/h0025950>
- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2), 157–162. <https://doi.org/10.1111/1467-9280.00327>
- Ástebro, T., Mata, J., & Santos-Pinto, L. (2015). Skewness seeking: Risk loving, optimism or overweighting of small probabilities? *Theory and Decision*, 78(2), 189–208. <https://doi.org/10.1007/s11238-014-9417-4>
- Badcock, D. R., & Westheimer, G. (1985). Spatial location and hyperacuity: The centre/surround localization contribution function has two substrates. *Vision Research*, 25(9), 1259–1267. [https://doi.org/10.1016/0042-6989\(85\)90041-0](https://doi.org/10.1016/0042-6989(85)90041-0)
- Baek, J., & Chong, S. C. (2020). Distributed attention model of perceptual averaging. *Attention, Perception, & Psychophysics*, 82(1), 63–79. <https://doi.org/10.3758/s13414-019-01827-z>
- Bair, W. (2005). Visual receptive field organization. *Current Opinion in Neurobiology*, 15(4), 459–464. <https://doi.org/10.1016/j.conb.2005.07.006>
- Bakker, A. (2004). Reasoning about shape as pattern in variability. *Statistics Education Research Journal*, 3(2), 64–83.
- Barrowman, N. J., & Myers, R. A. (2003). Raindrop plots: A new way to display collections of likelihoods and distributions. *The American Statistician*, 57(4), 268–274. <https://doi.org/10.1198/0003130032369>
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4), 577–660. <https://doi.org/10.1017/S0140525X99002149>
- Batanero, C., Cobo Merino, B., & Diaz, C. (2003). Assessing secondary school students' understanding of averages. *Proceedings of the 3rd Conference of the European Society for Research in Mathematics Education*, 1–9.
- Bauer, B. (2017). Perceptual averaging of line length: Effects of concurrent digit memory load. *Attention, Perception, & Psychophysics*, 79(8), 2510–2522. <https://doi.org/10.3758/s13414-017-1388-4>
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers Misunderstand Confidence Intervals and Standard Error Bars. *Psychological Methods*, 10(4), 389–396. <https://doi.org/10.1037/1082-989X.10.4.389>
- Ben-Zvi, D. (2003). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, 3(2), 42–63.
- Bertini, E., Tatu, A., & Keim, D. (2011). Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2203–2212. <https://doi.org/10.1109/TVCG.2011.229>
- Biehler, R. (1997). Software for Learning and for Doing Statistics. *International Statistical Review*, 65(2), 167–189. <https://doi.org/10.1111/j.1751-5823.1997.tb00399.x>
- Brannon, E. M., & Merritt, D. J. (2011). Evolutionary foundations of the approximate number system. In S. Dehaene & E. M. Brannon (Eds.), *Space, time and number in the brain: Searching for the foundations of mathematical thought* (pp. 207–224). Elsevier. <https://doi.org/10.1016/B978-0-12-385948-8.00014-1>
- Broad, K., Leiserowitz, A., Weinkle, J., & Steketee, M. (2007). Misinterpretations of the “cone of uncertainty” in Florida during the 2004 hurricane season. *Bulletin of the American Meteorological Society*, 88(5), 651–668. <https://doi.org/10.1175/BAMS-88-5-651>

- Bronfman, Z. Z., Brezis, N., Jacobson, H., & Usher, M. (2014). We see more than we can report: “cost free” color phenomenality outside focal attention. *Psychological Science*, 25(7), 1394–1403. <https://doi.org/10.1177/0956797614532656>
- Cai, J., & Moyer, J. (1995). *Beyond the computational algorithm: Students' understanding of the arithmetic average concept*. 3, 144–151. Recife, Brazil: Universidade Federal de Pernambuco.
- Carswell, C. M. (1992). Choosing Specifiers: An Evaluation of the Basic Tasks Model of Graphical Perception. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 34(5), 535–554. <https://doi.org/10.1177/001872089203400503>
- Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, 2(2), 98–113. <https://doi.org/10.1016/j.edurev.2007.04.001>
- Chan, S. W., & Ismail, Z. (2013). Assessing Misconceptions in Reasoning About Variability Among High School Students. *Procedia - Social and Behavioral Sciences*, 93, 1478–1483. <https://doi.org/10.1016/j.sbspro.2013.10.067>
- Chetverikov, A., Campana, G., & Kristjánsson, Á. (2016). Building ensemble representations: How the shape of preceding distractor distributions affects visual search. *Cognition*, 153, 196–210. <https://doi.org/10.1016/j.cognition.2016.04.018>
- Chong, S. C., & Evans, K. K. (2011). Distributed versus focused attention (count vs estimate): Distributed versus focused attention. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(6), 634–638. <https://doi.org/10.1002/wcs.136>
- Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, 43(4), 393–404. [https://doi.org/10.1016/S0042-6989\(02\)00596-5](https://doi.org/10.1016/S0042-6989(02)00596-5)
- Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, 45(7), 891–900. <https://doi.org/10.1016/j.visres.2004.10.004>
- Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387), 531–554. <https://doi.org/10.1080/01621459.1984.10478080>
- Cleveland, W. S., & McGill, R. (1986). An experiment in graphical perception. *International Journal of Man-Machine Studies*, 25(5), 491–500. [https://doi.org/10.1016/S0020-7373\(86\)80019-0](https://doi.org/10.1016/S0020-7373(86)80019-0)
- Cochrane, A., Cui, L., Hubbard, E. M., & Green, C. S. (2019). “Approximate number system” training: A perceptual learning approach. *Attention, Perception, & Psychophysics*, 81(3), 621–636. <https://doi.org/10.3758/s13414-018-01636-w>
- Cooper, L., & Shore, F. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education*, 15(2), 1–13. <https://doi.org/10.1080/10691898.2008.11889559>
- Cooper, L., & Shore, F. S. (2010). The effects of data and graph type on concepts and visualizations of variability. *Journal of Statistics Education*, 18(2), 1–16.
- Corbett, J. E., & Melcher, D. (2014). Characterizing ensemble statistics: Mean size is represented across multiple frames of reference. *Attention, Perception, & Psychophysics*, 76(3), 746–758. <https://doi.org/10.3758/s13414-013-0595-x>
- Corbett, J. E., Wumtsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition*, 20(2), 211–231. <https://doi.org/10.1080/13506285.2012.657261>
- Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 2142–2151. <https://doi.org/10.1109/TVCG.2014.2346298>
- Cui, L., Massey, C. M., & Kellman, P. J. (2018). Perceptual learning in correlation estimation: The role of learning category organization. In T.T. Rogers, M. Rau, X. Zhu, & C.W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp 262–267). Madison, Cognitive Science Society.
- Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine*, 28(2), 205–220. <https://doi.org/10.1002/sim.3471>
- Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A*, 18(5), 1016. <https://doi.org/10.1364/JOSAA.18.001016>
- de Fockert, J., & Wolfenstein, C. (2009). Short article: Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology*, 62(9), 1716–1722. <https://doi.org/10.1080/17470210902811249>
- delMas, R., Garfield, J., & Ooms, A. (2004). *Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions*. Paper presented at the Proceedings of the Fourth International Research Forum on Statistical Reasoning, Literacy, and Reasoning, Auckland, New Zealand.
- delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2), 28–58.
- delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, 4(1), 55–82.
- Fischer, M. H., Dewulf, N., & Hill, R. L. (2005). Designing bar graphs: Orientation matters. *Applied Cognitive Psychology*, 19(7), 953–962. <https://doi.org/10.1002/acp.1105>
- Fouriez, G., Rubinfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics*, 70(3), 456–464. <https://doi.org/10.3758/PP.70.3.456>
- Friendly, M. (2008). A brief history of data visualization. In C.-h. Chen, W. K. K. Härdle, & A. Unwin (Eds.), *Handbook of computational statistics: Data visualization* (Vol. 3, pp. 15–56). Heidelberg, Germany: Springer-Verlag.
- Garfield, J. B., & Ben-Zvi, D. (2008). Learning to reason about variability. In *Developing students' statistical reasoning: Connecting research and teaching practice* (pp. 201–214). Springer.
- Gleicher, M., Correll, M., Nothelfer, C., & Franconeri, S. (2013). Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2316–2325. <https://doi.org/10.1109/TVCG.2013.183>
- Godau, C., Vogelgesang, T., & Gaschler, R. (2016). Perception of bar graphs—A biased impression? *Computers in Human Behavior*, 59, 67–73. <https://doi.org/10.1016/j.chb.2016.01.036>
- Goldenberg, A., Sweeny, T. D., Shpigel, E., & Gross, J. J. (2020). Is this my group or not? The role of ensemble coding of emotional expressions in group categorization. *Journal of Experimental Psychology: General*, 149(3), 445–460. <https://doi.org/10.1037/xge0000651>
- Griffiths, S., Rhodes, G., Jeffery, L., Palermo, R., & Neumann, M. F. (2018). The average facial expression of a crowd influences impressions of individual expressions. *Journal of Experimental Psychology: Human Perception and Performance*, 44(2), 311–319. <https://doi.org/10.1037/xhp0000446>
- Haberman, J., Lee, P., & Whitney, D. (2015). Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision*, 15(4), 16. <https://doi.org/10.1167/15.4.16>
- Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, 17(17), R751–R753. <https://doi.org/10.1016/j.cub.2007.06.039>
- Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, 35(3), 718–734. <https://doi.org/10.1037/a0013899>
- Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception & Psychophysics*, 72(7), 1825–1838. <https://doi.org/10.3758/APP.72.7.1825>

- Hancock, S. A., & Rummerfield, W. (2020). Simulation methods for teaching sampling distributions: Should hands-on activities precede the computer? *Journal of Statistics Education*, 28(1), 9–17. <https://doi.org/10.1080/10691898.2020.1720551>
- Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 203–212). Atlanta, GA: Association for Computing Machinery. <https://doi.org/10.1145/1753326.1753357>
- Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition*, 10(4), 389–395. <https://doi.org/10.3758/BF03202431>
- Hofman, J. M., Goldstein, D. G., & Hullman, J. (2020). How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Honolulu, HI: ACM. <https://doi.org/10.1145/3313831.3376454>
- Hsu, C. T., & Lawley, D. N. (1940). The derivation of the fifth and sixth moments of the distribution of b^2 in samples from a normal population. *Biometrika*, 31(3/4), 238. <https://doi.org/10.2307/2332606>
- Hubert-Wallander, B., & Boynton, G. M. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of Vision*, 15(4), 5. <https://doi.org/10.1167/15.4.5>
- Huck, S. W. (2009). *Statistical Misconception*. London, UK: Psychology Press.
- Huestegge, L., & Pöttsch, T. H. (2018). Integration processes during frequency graph comprehension: Performance and eye movements while processing tree maps versus pie charts. *Applied Cognitive Psychology*, 32(2), 200–216. <https://doi.org/10.1002/acp.3396>
- Hullman, J., Resnick, P., & Adar, E. (2015). Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLOS ONE*, 10(11), e0142444. <https://doi.org/10.1371/journal.pone.0142444>
- Ibrekk, H., & Morgan, M. G. (1987). Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis*, 7(4), 519–529. <https://doi.org/10.1111/j.1539-6924.1987.tb00488.x>
- Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception, & Psychophysics*, 75(2), 278–286. <https://doi.org/10.3758/s13414-012-0399-4>
- Inzunza, S. (2006). *Students' errors and difficulties for solving problems of sampling distributions by means of computer simulation*. Paper presented at the 7th Annual International Conference on Teaching Statistics.
- Ismail, Z., & Chan, S. W. (2015). *Malaysian students' misconceptions about measures of central tendency: An error analysis* (pp. 93–100). Pahang, Malaysia. <https://doi.org/10.1063/1.4907430>
- Jamie, D. M. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, 10(1), 4. <https://doi.org/10.1080/10691898.2002.11910548>
- Jardine, N., Ondov, B. D., Elmqvist, N., & Franconeri, S. (2020). The Perceptual Proxies of Visual Comparison. *IEEE Transactions on Visualization and Computer Graphics*, 26(1), 1012–1021. <https://doi.org/10.1109/TVCG.2019.2934786>
- Jarvenpää, S. L., & Dickson, G. W. (1988). Graphics and managerial decision making: Research-based guidelines. *Communications of the ACM*, 31(6), 764–774. <https://doi.org/10.1145/62959.62971>
- Jeong, J., & Chong, S. C. (2020). Adaptation to mean and variance: Interrelationships between mean and variance representations in orientation perception. *Vision Research*, 167, 46–53. <https://doi.org/10.1016/j.visres.2020.01.002>
- Ji, L., & Pourtois, G. (2018). Capacity limitations to extract the mean emotion from multiple facial expressions depend on emotion variance. *Vision Research*, 145, 39–48. <https://doi.org/10.1016/j.visres.2018.03.007>
- Joslyn, S., & LeClerc, J. (2013). Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, 22(4), 308–315. <https://doi.org/10.1177/0963721413481473>
- Kahneman, D. (2011). *Thinking fast and slow*. New York, NY: Farrar, Straus & Giroux.
- Kampstra, P. (2008). Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software*, 28(Code Snippet 1). <https://doi.org/10.18637/jss.v028.c01>
- Kaplan, J. J., Gabrosek, J. G., Curtiss, P., & Malone, C. (2014). Investigating student understanding of histograms. *Journal of Statistics Education*, 22(2), 1–30.
- Katzin, N., Cohen, Z. Z., & Henik, A. (2019). If it looks, sounds, or feels like subitizing, is it subitizing? A modulated definition of subitizing. *Psychonomic Bulletin & Review*, 26(3), 790–797. <https://doi.org/10.3758/s13423-018-1556-0>
- Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5092–5103). San Jose, CA: ACM. <https://doi.org/10.1145/2858036.2858558>
- Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, 18(9), 23. <https://doi.org/10.1167/18.9.23>
- Khvostov, V. A., & Utochkin, I. S. (2019). Independent and parallel visual processing of ensemble statistics: Evidence from dual tasks. *Journal of Vision*, 19(9), 3. <https://doi.org/10.1167/19.9.3>
- Kong, N., Heer, J., & Agrawala, M. (2010). Perceptual guidelines for creating rectangular treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 990–998. <https://doi.org/10.1109/TVCG.2010.186>
- Kosara, R. (2019). Circular part-to-whole charts using the area visual cue. *Proceedings of the Eurographics/IEEE VGTC Symposium on Visualization (EuroVis)*, 13–17. <https://doi.org/10.2312/EVS.20191163>
- Kosslyn, S. M. (2006). *Graph Design for the Eye and Mind*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195311846.001.0001>
- Kraus, A., & Litzenberger, R. H. (1976). Skewness preference and the valuation of risk assets*. *The Journal of Finance*, 31(4), 1085–1100. <https://doi.org/10.1111/j.1540-6261.1976.tb01961.x>
- Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. *Cognitive Science*, 37(5), 775–799. <https://doi.org/10.1111/cogs.12028>
- Lane, D. M. (2015). Simulations of the sampling distribution of the mean do not necessarily mislead and can facilitate learning. *Journal of Statistics Education*, 23(2), 6. <https://doi.org/10.1080/10691898.2015.11889738>
- Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, 1(2), 115–121. <https://doi.org/10.1111/j.1467-9280.1990.tb00079.x>
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1), 65–100. <https://doi.org/10.1111/j.1551-6708.1987.tb00863.x>
- Lau, J. S.-H., & Brady, T. F. (2018). Ensemble statistics accessed through proxies: Range heuristic and dependence on low-level properties in variability discrimination. *Journal of Vision*, 18(9), 3. <https://doi.org/10.1167/18.9.3>
- Lavigne, N. C., Salkind, S. J., & Yan, J. (2008). Exploring college students' mental representations of inferential statistics. *The Journal of Mathematical Behavior*, 27(1), 11–32. <https://doi.org/10.1016/j.jmathb.2007.10.003>
- Lee, C., & Meletiou-Mavrotheris, M. (2003). Some difficulties of learning histograms in introductory statistics. *Proceedings of the 2003 Joint Statistical Meeting, Section on Statistical Education* (pp. 2326–2333).

- Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2014). Interpreting histograms. As easy as it seems? *European Journal of Psychology of Education*, 29(4), 557–575. <https://doi.org/10.1007/s10212-014-0213-x>
- Li, H., Ji, L., Tong, K., Ren, N., Chen, W., Liu, C. H., & Fu, X. (2016). Processing of individual items during ensemble coding of facial expressions. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.01332>
- Liu, L., Boone, A. P., Ruginski, I. T., Padilla, L., Hegarty, M., Creem-Regehr, S. H., ... House, D. H. (2017). Uncertainty visualization by representative sampling from prediction ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 23(9), 2165–2178. <https://doi.org/10.1109/TVCG.2016.2607204>
- Liu, L., Mirzargar, M., Kirby, R. M., Whitaker, R., & House, D. H. (2015). Visualizing time-specific hurricane predictions, with uncertainty, from storm path ensembles. *Computer Graphics Forum*, 34(3), 371–380. <https://doi.org/10.1111/cgf.12649>
- Liu, T.-C., Lin, Y.-C., & Kinshuk. (2010). The application of Simulation-Assisted Learning Statistics (SALS) for correcting misconceptions and improving understanding of correlation: Effects of SALS-based learning. *Journal of Computer Assisted Learning*, 26(2), 143–158. <https://doi.org/10.1111/j.1365-2729.2009.00330.x>
- Luo, A. X., & Zhao, J. (2018). Capacity limit of ensemble perception of multiple spatially intermixed sets. *Attention, Perception, & Psychophysics*, 80(8), 2033–2047. <https://doi.org/10.3758/s13414-018-1572-1>
- Makar, K., & Confrey, J. (2003). Moving the context of modeling to the forefront. In C. Lee (Ed.), *Reasoning about variability: A collection of current research studies*. Dordrecht, The Netherlands: Kluwer Academic Publisher.
- Makar, K., & Confrey, J. (2005). Using distributions as statistical evidence in well-structured and ill-structured problems. *Reasoning about distribution: A collection of current research studies*. Presented at the Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy, Brisbane, Australia.
- Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica*, 142(2), 245–250. <https://doi.org/10.1016/j.actpsy.2012.11.002>
- Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision*, 15(4), 6. <https://doi.org/10.1167/15.4.6>
- Maule, J., & Franklin, A. (2016). Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism. *Journal of the Optical Society of America A*, 33(3), A22. <https://doi.org/10.1364/JOSAA.33.000A22>
- Maule, J., & Franklin, A. (2019). Adaptation to variance generalizes across visual domains. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0000678>
- McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, 143(12), 1273–1312. <https://doi.org/10.1037/bul0000126>
- Meletiou-Mavrotheris, M., & Lee, C. (2002). Teaching students the stochastic nature of statistical concepts in an Introductory statistics course. *Statistics Education Research Journal*, 1(2), 2–37.
- Meletiou-Mavrotheris, M., & Lee, C. (2010). Investigating college-level introductory statistics students' prior knowledge of graphing. *Canadian Journal of Science, Mathematics and Technology Education*, 10(4), 339–355. <https://doi.org/10.1080/14926156.2010.524964>
- Mevarech, Z. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics*, 14, 415–429.
- Micallef, L., Palmas, G., Oulasvirta, A., & Weinkauff, T. (2017). Towards perceptual optimization of the visual design of scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 23(6), 1588–1599. <https://doi.org/10.1109/TVCG.2017.2674978>
- Michael, E., de Gardelle, V., & Summerfield, C. (2014). Priming by the variability of visual information. *Proceedings of the National Academy of Sciences of the United States of America*, 111(21), 7873–7878. <https://doi.org/10.1073/pnas.1308674111>
- Morgan, M., Chubb, C., & Solomon, J. A. (2008). A “dipper” function for texture discrimination based on orientation variance. *Journal of Vision*, 8(11), 9–9. <https://doi.org/10.1167/8.11.9>
- Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics*, 70(5), 772–788. <https://doi.org/10.3758/PP.70.5.772>
- Neumann, D. L., Neumann, M. M., & Hood, M. (2011). Evaluating computer-based simulations, multimedia and animations that help integrate blended learning with lectures in first year statistics. *Australasian Journal of Educational Technology*, 27(2). <https://doi.org/10.14742/ajet.970>
- Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, 19(4), 601–607. <https://doi.org/10.3758/s13423-012-0247-5>
- Norman, L. J., Heywood, C. A., & Kentridge, R. W. (2015). Direct encoding of orientation variance in the visual system. *Journal of Vision*, 15(4), 3. <https://doi.org/10.1167/15.4.3>
- Odic, D., & Starr, A. (2018). An introduction to the approximate number system. *Child Development Perspectives*, 12(4), 223–229. <https://doi.org/10.1111/cdep.12288>
- Okan, Y., Galesic, M., & Garcia-Retamero, R. (2016). How people with low and high graph literacy process health graphs: Evidence from eye-tracking: Graph literacy and health graph processing. *Journal of Behavioral Decision Making*, 29(2/3), 271–294. <https://doi.org/10.1002/bdm.1891>
- Olani, A., Hoekstra, R., Harskamp, E., & Van der Werf, G. (2010). Statistical reasoning ability, self-efficacy, and value beliefs in a reform based university statistics course. *Electronic Journal of Research in Education Psychology*, 8(22). <https://doi.org/10.25115/ejrep.v9i23.1427>
- Oriet, C., & Hozempa, K. (2016). Incidental statistical summary representation over time. *Journal of Vision*, 16(3), 3. <https://doi.org/10.1167/16.3.3>
- Palmer, S. E. (2002). Perceptual organization in vision. In H. Pashler (Ed.), *Stevens' handbook of experimental psychology* (p. pas0105). Hoboken, NJ: John Wiley & Sons, Inc. <https://doi.org/10.1002/0471214426.pas0105>
- Peebles, D. (2008). The effect of emergent features on judgments of quantity in configural and separable displays. *Journal of Experimental Psychology: Applied*, 14(2), 85–100. <https://doi.org/10.1037/1076-898X.14.2.85>
- Peebles, D., & Ali, N. (2009). *Differences in comprehensibility between three-variable bar and line graphs* (pp. 2938–2943). Mahwah, NJ: Erlbaum.
- Pfannkuch, M., & Reading, C. (2006). Reasoning about distribution: A complex process. *Statistics Education Research Journal*, 5(2), 4–9.
- Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science*, 24(8), 1389–1397. <https://doi.org/10.1177/0956797612473759>
- Piazza, M., & Izard, V. (2009). How humans count: Numerosity and the parietal cortex. *The Neuroscientist*, 15(3), 261–273. <https://doi.org/10.1177/1073858409333073>
- Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics*, 12(2), 191–204. <https://doi.org/10.1007/BF00305621>

- Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, 83(2, Pt.1), 304–308. <https://doi.org/10.1037/h0028558>
- Raidvee, A., Toom, M., Averin, K., & Allik, J. (2020). Perception of means, sums, and areas. *Attention, Perception, & Psychophysics*. <https://doi.org/10.3758/s13414-019-01938-7>
- Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, 14(1), 36–49. <https://doi.org/10.1037/1076-898X.14.1.36>
- Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In Dani Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201–226). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/1-4020-2278-6_9
- Reid, J., & Reading, C. (2008). Measuring the development of students' consideration of variation. *Statistics Education Research Journal*, 7(1), 40–59.
- Rensink, R. A., & Baldrige, G. (2010). The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, 29(3), 1203–1210. <https://doi.org/10.1111/j.1467-8659.2009.01694.x>
- Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, 19(6), 607–614. <https://doi.org/10.1111/j.1467-9280.2008.02130.x>
- Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision*, 11(12), 18–18. <https://doi.org/10.1167/11.12.18>
- Saket, B., Endert, A., & Demiralp, C. (2018). Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(7), 2505–2512. <https://doi.org/10.1109/TVCG.2018.2829750>
- Saket, B., Srinivasan, A., Ragan, E. D., & Endert, A. (2018). Evaluating interactive graphical encodings for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(3), 1316–1330. <https://doi.org/10.1109/TVCG.2017.2680452>
- Sedlmair, M., Tatu, A., Munzner, T., & Tory, M. (2012). A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 31(3, Pt. 4), 1335–1344. <https://doi.org/10.1111/j.1467-8659.2012.03125.x>
- Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes: Topics in cognitive science. *Topics in Cognitive Science*, 3(3), 560–578. <https://doi.org/10.1111/j.1756-8765.2009.01066.x>
- Shah, P., Freedman, E. G., & Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 426–476). Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9780511610448.012>
- Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, 14(1), 47–69. <https://doi.org/10.1023/A:1013180410169>
- Silverman, A., Gramazio, C., & Schloss, K. (2016). The dark is more (Dark+) bias in colormap data visualizations with legends. *Journal of Vision*, 16(12), 628. <https://doi.org/10.1167/16.12.628>
- Simkin, D., & Hastie, R. (1987). An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association*, 82(398), 454–465. <https://doi.org/10.1080/01621459.1987.10478448>
- Simons, D. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Perception & Psychophysics*, 70(7), 1335–1336. <https://doi.org/10.3758/PP.70.7.1335>
- Skau, D., & Kosara, R. (2016). Arcs, angles, or areas: Individual data encodings in pie and donut charts. *Computer Graphics Forum*, 35(3), 121–130. <https://doi.org/10.1111/cgf.12888>
- Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, 10(14), 19–19. <https://doi.org/10.1167/10.14.19>
- Solomon, J. A. & Morgan, M. J. (2018). Calculation efficiencies for mean numerosity. *Psychological Science*, 29(11), 1824–1831. <https://doi.org/10.1177/0956797618790545>
- Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, 11(12), 13–13. <https://doi.org/10.1167/11.12.13>
- Stephenson, D. B., & Doblas-Reyes, F. J. (2000). Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus A*, 52(3), 300–322. <https://doi.org/10.1034/j.1600-0870.2000.d01-5.x>
- Strauss, S., & Bichler, E. (1988). The development of children's concepts of the arithmetic average. *Journal for Research in Mathematics Education*, 19(1), 64–80.
- Suárez-Pinilla, M., Seth, A. K., & Roseboom, W. (2018). Serial dependence in the perception of visual variance. *Journal of Vision*, 18(7), 4. <https://doi.org/10.1167/18.7.4>
- Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance*, 39(2), 329–337. <https://doi.org/10.1037/a0028712>
- Szafir, D. A. (2018). Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), 392–401. <https://doi.org/10.1109/TVCG.2017.2744359>
- Szafir, D. A., Haroz, S., Gleicher, M., & Franconeri, S. (2016). Four types of ensemble coding in data visualizations. *Journal of Vision*, 16(5), 11. <https://doi.org/10.1167/16.5.11>
- Tiurina, N. A., & Utochkin, I. S. (2019). Ensemble perception in depth: Correct size-distance rescaling of multiple objects before averaging. *Journal of Experimental Psychology: General*, 148(4), 728–738. <https://doi.org/10.1037/xge0000485>
- Tokita, M., Ueda, S., & Ishiguchi, A. (2016). Evidence for a global sampling process in extraction of summary statistics of item sizes in a set. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00711>
- Tong, K., Ji, L., Chen, W., & Fu, X. (2015). Unstable mean context causes sensitivity loss and biased estimation of variability. *Journal of Vision*, 15(4), 15. <https://doi.org/10.1167/15.4.15>
- Torgerson, W. S. (1961). Distances and ratios in psychophysical scaling. *Acta Psychologica*, 19, 201–205. [https://doi.org/10.1016/S0001-6918\(61\)80072-3](https://doi.org/10.1016/S0001-6918(61)80072-3)
- Utochkin, I. S., & Brady, T. F. (2020). Individual representations in visual working memory inherit ensemble properties. *Journal of Experimental Psychology: Human Perception and Performance*, 46(5), 458–473. <https://doi.org/10.1037/xhp0000727>
- Utochkin, I. S., Khvostov, V. A., & Stakina, Y. M. (2018). Continuous to discrete: Ensemble-based segmentation in the perception of multiple feature conjunctions. *Cognition*, 179, 178–191. <https://doi.org/10.1016/j.cognition.2018.06.016>
- Valsecchi, M., Stucchi, N., & Scocchia, L. (2018). Repulsive serial effects in visual numerosity judgments. *Perception*, 47(7), 780–788. <https://doi.org/10.1177/0301006618775235>
- Wang, P.-Y., Vaughn, B. K., & Liu, M. (2011). The impact of animation interactivity on novices' learning of introductory statistics. *Computers & Education*, 56(1), 300–311. <https://doi.org/10.1016/j.compedu.2010.07.011>
- Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals?: The role of statistical perception in determining whether awareness overflows access. *Cognition*, 152, 78–86. <https://doi.org/10.1016/j.cognition.2016.01.010>
- Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. *R.I.P. The American Statistician*, 68(3), 191–195. <https://doi.org/10.1080/00031305.2014.917055>

- Westheimer, G. (2010). Visual acuity and hyperacuity. In M. Bass (Ed.), *Handbook of optics. Vol. 3: Vision and vision optics* (3rd ed., Ch. 5). New York, NY: McGraw Hill.
- Whitaker, D., & Walker, H. (1988). Centroid evaluation in the vernier alignment of random dot clusters. *Vision Research*, 28(7), 777–784. [https://doi.org/10.1016/0042-6989\(88\)90024-7](https://doi.org/10.1016/0042-6989(88)90024-7)
- Whitaker, R. T., Mirzargar, M., & Kirby, R. M. (2013). Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2713–2722. <https://doi.org/10.1109/TVCG.2013.143>
- Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, 69(1), 105–129. <https://doi.org/10.1146/annurev-psych-010416-044232>
- Wild, C. J. (2006). The concept of distribution. *Statistics Education Research Journal*, 5(2), 10–26.
- Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223–248. <https://doi.org/10.1111/j.1751-5823.1999.tb00442.x>
- Witt, J. K. (2019). The perceptual experience of variability in line orientation is greatly exaggerated. *Journal of Experimental Psychology: Human Perception and Performance*, 45(8), 1083–1103. <https://doi.org/10.1037/xhp0000648>
- Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of Vision*, 13(7), 1–1. <https://doi.org/10.1167/13.7.1>
- Wolfe, B. A., Kosovicheva, A. A., Leib, A. Y., Wood, K., & Whitney, D. (2015). Foveal input is not required for perception of crowd facial expression. *Journal of Vision*, 15(4), 11. <https://doi.org/10.1167/15.4.11>
- Wright, J. M., Morris, A. P., & Kregelberg, B. (2011). Weighted integration of visual position information. *Journal of Vision*, 11(14), 11–11. <https://doi.org/10.1167/11.14.11>
- Yamanashi Leib, A., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications*, 7(1), 13186. <https://doi.org/10.1038/ncomms13186>
- Yang, Y., Tokita, M., & Ishiguchi, A. (2018). Is There a Common Summary Statistical Process for Representing the Mean and Variance? A Study Using Illustrations of Familiar Items. *I-Perception*, 9(1), 204166951774729. <https://doi.org/10.1177/2041669517747297>
- Yildirim, I., Ögreden, O., & Boduroglu, A. (2018). Impact of spatial grouping on mean size estimation. *Attention, Perception, & Psychophysics*, 80(7), 1847–1862. <https://doi.org/10.3758/s13414-018-1560-5>
- Ying, H., Burns, E., Lin, X., & Xu, H. (2019). Ensemble statistics shape face adaptation and the cheerleader effect. *Journal of Experimental Psychology: General*, 148(3), 421–436. <https://doi.org/10.1037/xge0000564>
- Yu, J., Goldstone, R. L., & Landy, D. (2018). *Experientially grounded learning about the roles of variability, sample size, and difference between means in statistical reasoning* (pp. 2744–2749). Madison, CI: Cognitive Science Society.
- Yuan, L., Haroz, S., & Franconeri, S. (2019). Perceptual proxies for extracting averages in data visualizations. *Psychonomic Bulletin & Review*, 26(2), 669–676. <https://doi.org/10.3758/s13423-018-1525-7>
- Zacks, J., Levy, E., Tversky, B., & Schiano, D. J. (1998). Reading bar graphs: Effects of extraneous depth cues and graphical context. *Journal of Experimental Psychology: Applied*, 4(2), 119–138. <https://doi.org/10.1037/1076-898X.4.2.119>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.