Juvenile Graphical Perception: A Comparison between Children and Adults

Liudas Panavas
Northeastern University
panavas.l@northeastern.edu

Amy Worth
Ivy After School Program
amyworth.aw@gmail.com

Tarik Crnovrsanin
Northeastern University
t.crnovrsanin@northeastern.edu

Tejas Sathyamurthi
Northeastern University
sathyamurthi.t@northeastern.edu

Sara Cordes
Boston College
sara.cordes@bc.edu

Michelle A. Borkin
Northeastern University
m.borkin@northeastern.edu

Cody Dunne
Northeastern University
c.dunne@northeastern.edu

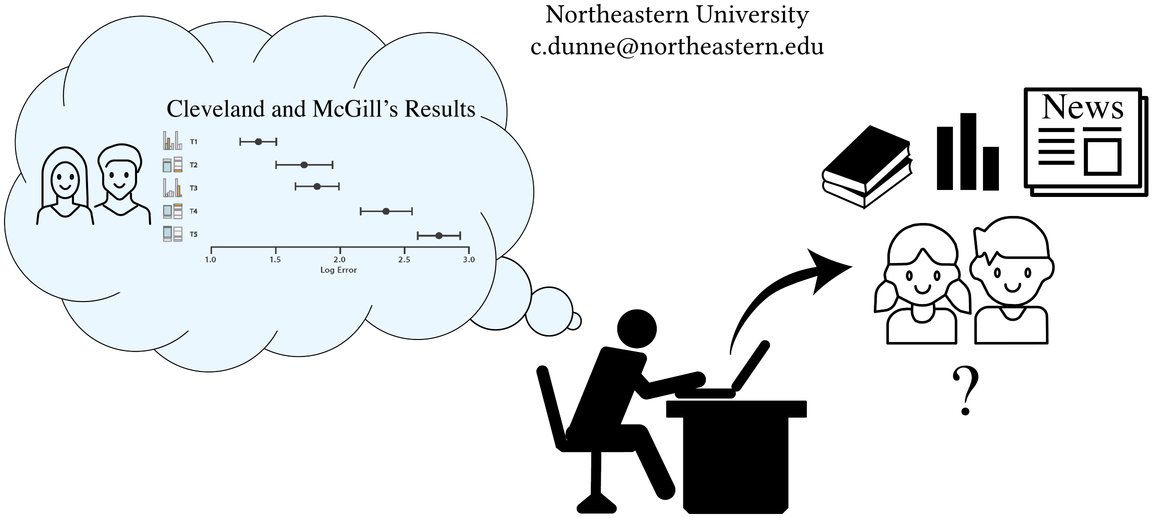


Figure 1: Designers use guidelines empirically tested on adults [48] to create visualizations for children. Our study takes a step towards ensuring those guidelines apply to all ages by comparing graphical perception in children ages 8–12 with adults. Cleveland & McGill results image credit: Tamara Munzner, CC-BY 4.0.

ABSTRACT

Data visualization is pervasive in the lives of children as they encounter graphs and charts in early education and online media. In spite of this prevalence, our guidelines and understanding of how children perceive graphs stem primarily from studies conducted with adults. Previous psychology and education research indicates that children's cognitive abilities are different from adults. Therefore, we conducted a classic graphical perception study on a population of children aged 8–12 enrolled in the Ivy After School Program in Boston, MA and adult computer science students enrolled in Northeastern University to determine how accurately participants judge differences in particular graphical encodings. We record the accuracy of participants' answers for five encodings most commonly used with quantitative data. The results of our controlled experiment show that children have remarkably similar graphical perception to adults, but are consistently less accurate at interpreting the visual encodings. We found similar effectiveness rankings, relative differences in error between the different encodings, and patterns of bias across encoding types. Based on our findings, we provide design guidelines and recommendations for creating visualizations for children. This paper and all supplemental materials are available at https://osf.io/ygrdv.

KEYWORDS

Visualization Literacy, Data Visualization, Graphical Perception

ACM Reference Format:

Liudas Panavas, Amy Worth, Tarik Crnovrsanin, Tejas Sathyamurthi, Sara Cordes, Michelle A. Borkin, and Cody Dunne. 2022. Juvenile Graphical Perception: A Comparison between Children and Adults. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3491102.3501893

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501893

1 INTRODUCTION

Every year millions of children put on their backpacks, get on a bus, and head to school. They open textbooks containing different types of visualizations explaining math, science, or humanities concepts. The design of these visualizations will affect how well students grasp foundational skills and knowledge [46]. Ineffective visualizations may have far-reaching consequences and cause students to miss important concepts or lose interest in the material [31, 39, 61]. Despite these potential consequences, visualization researchers have conducted relatively few studies with children [2, 13, 31, 39]. While teachers *believe* that the visualizations they teach are intuitive for children to read [2, 13, 24], the design guidelines the visualization community has published for practitioners—like textbook authors—are based on conclusions drawn from adult populations. We have little actual *evidence* about children's graphical perception. We need more studies with children to develop age-appropriate visualization design guidelines.

Prior work has focused on how to *teach* visualizations and *cognitive strategies* related to proportional reasoning [2, 24, 64]. Likewise, educational research has made headways into both studying visualization design for children and designing curriculum to introduce the concepts [24, 60]. Developmental psychologists have also studied how children interpret visualizations to better understand their proportional reasoning skills [66]. But these studies do not question the assumptions of basic visualization design guidelines or provide empirical support for how to design visualizations for children.

To begin creating empirical guidelines for children, the study of *graphical perception*—the visual decoding of the quantitative and qualitative information encoded in visualizations [14, 15]—is a good place to start. The seminal work on graphical perception by Cleveland & McGill [14] deconstructed visualizations into their elementary visual encodings and tested adult users' accuracy when using them. Their results led to ranking these encodings forming the foundation of empirically-supported visualization design guidelines [14], which are widely taught in visualization classes. See, for example, fig. 2. Still, the field has not investigated how graphical perception varies with people's age and what perceptual and cognitive processes are involved [12, 14, 54]. Moreover, research in cognitive development shows that children and adults have different visual perception and proportional reasoning skills—two cornerstones of graphical perception [8, 29, 49, 66].

Our study focuses on comparing graphical perception between children aged 8–12 enrolled in the Ivy After School Program program in Boston, MA and adult computer science students enrolled in Northeastern University. By studying children's graphical perception ability, we can start gathering empirical evidence to answer such questions as: What are the elementary visual encoding rankings for children? How accurate are children in their judgements using these encodings? Do children have perceptual biases we need to account for? And how do these results differ from adults? Investigating these questions may help both educators and designers better teach and create visualizations for children by illuminating how children perceive them.

The results of our controlled experiment show that children have remarkably similar graphical perception to adults, but are consistently less accurate at interpreting the visual encodings.

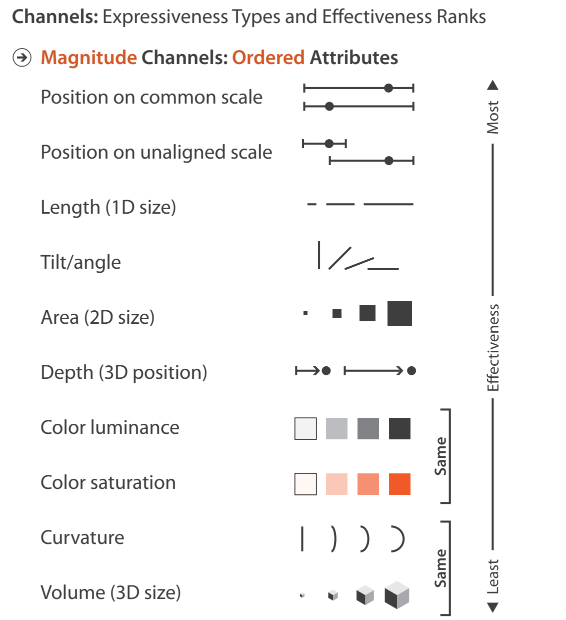


Figure 2: The efficacy ranking of encodings for showing quantitative information in Munzner's book [48], which is used in many visualization courses. Encodings that are more effective generally lead to less error on graphical perception tasks. Image credit: Tamara Munzner, CC-BY 4.0.

In particular, we contribute:

- An empirically-determined effectiveness ranking for the basic visual channels designers can use to build visualizations for children,
- A discussion of the similarities and differences of graphical perception task accuracy between children and adults,
- An understanding of the biases children present when conducting graphical perception tasks, and
- A set of design recommendations for creating visualizations for children.

This paper and all supplemental materials—including study preregistration, stimuli, experiment code, collected data, and analysis code—are freely available at https://osf.io/ygrdv.

2 RELATED WORK

2.1 Children's Use of Data Visualization

The research community broadly accepts that adult data visualization and visualization literacy skills are important [6]. Most user studies and research have also focused their efforts on creating novel or improved guidelines and techniques for adults [5, 26, 27]. Less research has addressed visualization for children and how improved visualization strategies may impact their lives. The relative lack of research is disproportionate to the widespread use and importance of visualization in young students' lives and education [56].

Researchers have examined and quantified the pervasiveness and importance of visualizations during children's early education. Students most frequently encounter visualizations in school when looking at textbooks and other teaching materials. In their review of elementary school textbooks, Alpers et al. found about half of the pages in grade K–4 textbooks contained data-driven graphics [2]. From a survey with teachers, they found around 25% of teaching materials used in the classroom were inherently visual. Visualizations are particularly common throughout educational material related to STEM and are used to reinforce new concepts [13, 37].

In textbooks and other educational materials, visualization has the potential to help students understand concepts in physics, chemistry, earth science, or biology and can teach principles such as collecting and organizing data [37]. Other research has shown that children are better at retaining, learning, and transferring knowledge when shown images rather than just words [71]. Skills, such as the ability to understand fractions, can be supported by fostering children’s ability to visualize proportions [46]. Government agencies have taken notice of these findings and have made teaching and learning visualization techniques a key component of national standards for both mathematics and science [6].

Not only are children seeing visualizations in school, they are increasingly exposed to a wide variety online. As of 2019, children ages 8–12 were consuming nearly 5 hours of screen media per day [57], and they often see the same media as adults. News organizations such as the New York Times have taken note and introduced educational practices such as “What’s Going On In This Picture” to help improve children’s visual literacy [69]. Other authors have published books of infographics to help illustrate concepts such as the human body or animals of the Earth [55].

Though elementary and middle school children often encounter visualizations in their schoolwork and media, the development of these skills is often overlooked. In a survey of 16 teachers of grades K–4, Chevalier et al. found that teachers do not believe enough time is spent on visualization skills [13]. The challenges in teaching visualization literacy can be attributed to many factors. Teachers may not have the necessary resources, or they believe that visualizations are intuitive and do not need to be taught [13, 53]. The result is that students are rarely taught the graphing skills they need for science and struggle to understand more than simple visualizations [41].

Students’ lack of skill, in combination with visualizations in textbooks that are not effectively designed, results in students being turned away from the material presented [61]. Kenney et al. [39] and Grammel et al. [26] found that participants became frustrated when they could not understand visualizations and either ignored the content or switched to a different task. When the visualization was clear or participants felt that their skill was enough to understand the visualization, they experienced positive emotions and wanted to engage with the material [39]. Therefore, depending on the quality of the visualization, students may be turned away or choose to dive deeper into the educational material presented. Due to the broad impact visualizations can make on children’s educational development, it is important that designers present them with effective, understandable visualizations.

2.2 Children’s Graphical Understanding

Though there is more work to do, the last few years has seen a surge of interest in children’s visual literacy [24, 25, 34]. There are now pedagogical guidelines and instructional implications drawn from assessing student visualization comprehension and partnering with education researchers [60]. Visualization has also been used to understand cognitive and visual developments in children [45, 46].

Research on children’s graphical comprehension has largely emphasized whether students have sufficient skill and are taught the necessary lessons to extract information from standard visualization formats. This has been measured through multiple-choice exams of standard visualizations and through assessing visualization construction [20, 52, 72]. Other exams, such as national numeracy exams, have prompted students to answer questions based on information presented in visualizations [43]. These studies examined a wide range of attributes of both visualizations and the students using them, providing educators with progressions for teaching students visualization and guidelines for introducing new topics [24, 60]. However, these pedagogical approaches to understanding visualization literacy and comprehension lack empirical evidence for how children perceive the basic building blocks of visualizations.

Many of the recent studies investigating children’s graphical understanding have been related to visual literacy, defined by Boy et al. as “the ability to use well-established data visualizations (e.g. line graphs) to handle information in an effective, efficient, and confident manner” [7]. This research has taken a variety of forms primarily focused on designing creative ways to improve children’s visual literacy [44]. Researchers have tried various approaches from running workshops with tangible objects to creating games [34, 58]. The games and tools created to teach visual literacy draw on a variety of strategies, from constructionist approaches to role playing [2, 4, 25]. This raises an important issue in visualization literacy, which is how to best teach students about visualization. To do so, teachers and researchers must first be equipped with the information they need—an understanding of how children actually think and decode information from visualizations.

2.3 Visualization Design Guidelines for Children and Adults

Another way of making visualizations more understandable, outside of pedagogical approaches, is by designing visualizations that are likely to be more effective for their intended audience. In one of the seminal works on visualization comprehension, Friel et al. say that graph comprehension may be improved by controlling perceptual demands [24]. Other pedagogical advice presented by Shah et al. [60] states that when educators both construct visualizations and teach visualization construction to children they should “Use the ‘best’ visual dimensions to convey metric information whenever possible.” When putting together pedagogical advice for visualization instructors, both Friel et al. and Shah et al. reference the design guidelines generated from Cleveland & McGill’s [15] study. Though the rankings put forth by Cleveland & McGill have proved more effective than other common design principles (e.g. “Data-to-Ink”) [10] and have been replicated with larger samples [30], they were studied in adult populations and therefore may not apply to children. Therefore, as guidelines stemming from graphical perception tests are a proven and effective way to design visualizations and improve visualization comprehension, it is important to ensure that they are accurate for both children and adults [10].

Graphical perception is a combination of several other perceptual and cognitive tasks—in which performance is likely to vary as a function of cognitive maturity and graphical experience. Two of the major components that compose graphical perception are proportional reasoning and visual perception [17, 66]. Numerous studies on both have shown that children and adults differ in their abilities and perception in a variety of ways [36, 51], including exhibiting various biases of over- and underestimation in proportional reasoning [33, 46]. Hollands & Dyre [32] generalize these biases and demonstrate that children have larger biases when making proportion judgements. Additional work by Jones & Dekker [36] investigated how children perceived and calculated the middle of a point cloud. They found that children were not simply worse at using the strategies adults employed, but instead children chose different but still sensible strategies. This would likely play a role in them performing differently than adults on graphical perception studies questions, as the middle of each encoding is an important reference point. Finally, evidence reveals the acuity of magnitude judgments (such as those for line length or surface area) follow distinct developmental trajectories across childhood, such that children reach adult levels of perceptual acuity in area discrimination earlier than they do for length discrimination [51]. As such, it is conceivable that children may perform relatively better on area compared to length encodings—a finding which would not align with adult rankings.

Beyond perceptual differences, children's error patterns may also differ due to strategic differences. While Cleveland & McGill found in their experiment with high schoolers [16, 17] that the error rate did not depend on technical experience or age, their participants were at the end or near the end of their primary schooling. Adults have significantly more experience with graphs than elementary-aged children—experience that allows for the development of efficient strategies on how to interpret graphs. Children—being graphical novices—may attempt to interpret graphical information using inefficient or entirely inaccurate strategies. There is ample evidence that children initially make robust errors when first dealing with novel representations of quantity. For example, in proportion judgment tasks, children consistently ignore relational information and instead focus on absolute amounts, judging that, e.g., a game spinner with 3 blue and 8 red pieces is more likely to land on blue than one with 2 blue and 2 red pieces because 3 is more than 2 [33, 35]. Moreover, even in the domain of graphical perception, there is some evidence that children's strategies may be different from those of adults. Spence & Krizel [66] found children under the age of 12 or 13 were more likely to interpret pie charts by judging the area or volume of the chart, whereas older children and adults were more likely to focus on the more reliable cues of angle or arc length. Thus, it would not be surprising if children's lack of expertise in interpreting visualizations may lead to distinct encoding rankings. If children's rankings do not align with adults, this would suggest that current guidelines for the development of graphical visualizations for children should be modified in order to promote graphical literacy in early graph learners.

These possibilities, along with research indicating that design guidelines used for adults do not always translate to children [38], prompt us to examine if graphical perception guidelines remain consistent for different age groups.

3 HYPOTHESES

Our experiment, detailed in section 4, has been designed to better understand children's graphical perception and how it compares to adults' graphical perception. Our study examines what are the encoding rankings, accuracy when perceiving each encoding type, and patterns of bias in decoding graphical stimuli. We focus on children ages 8–12, because in this age range their cognitive abilities are mature enough to complete the study, they are less likely to make biased errors due to lack of numerical understanding [64], and it is the primary age for many visualizations to be introduced in their curriculum in the United States [68]. (See section 4.3 for further discussion.)

Our hypotheses are presented here:

Hypothesis 1: Children have varying perceptual accuracy with different visual encodings. As adults have varying perceptual accuracy with different encodings (see, e.g., [14, 15, 30, 48]), we expect that children will also have varying accuracy. Moreover, in section 2.3 we discussed how graphical perception performance is likely to vary as children mature and gain experience using visualization—and, importantly, that this variation with age is not even across encodings. Children reach adult-level performance earlier in some tasks than others [51]. Children also may have less practice using specific charts, such as pie or bubble charts [68]. Concurrently, the math curriculum before this point in a child's education is unlikely to have focused much on concepts like angles [68]. Given how little we know about how children decode different charts, we want to verify that different encodings will indeed have different effectiveness for them. A broad comparison between the encodings can give us direction when creating design guidelines specifically catered to children.

Hypothesis 2: Children's overall elementary perceptual task accuracy will be lower than adults'. Results from prior studies with children (see section 2.3) guide us to expect that children will have larger errors than adults on graphical perception tasks. We believe this because children's proportional reasoning skills are still maturing [8, 46], their graphical perception skill is tied to numerous other cognitive abilities that may mature with more schooling [42], and they use different and sometimes less effective strategies than adults [36]. Our best indicator of a difference between children and adults is the work done by Spence & Krizel [65, 66], which found consistently larger errors for children on proportional reasoning tasks. While these studies compare children and adult accuracy on common visual encodings, they study proportional reasoning—i.e. $\frac{a}{b} = \frac{c}{d}$—as opposed to the graphical perception tasks we focus on which are of the form $\frac{a}{b}$ [8]. From Spence & Krizel's results [65, 66] , as well as the other studies we cite in section 2.3, we expect that the children will have larger errors than adults but do not yet know by how much they differ. Comparing adults to children for each encoding may give us a first glimpse into how children process these encodings differently from adults.

Hypothesis 3: Children will exhibit patterns of bias in errors when making graphical perception judgements. Do children have patterns of over- and underestimation in their answers to graphical perception questions? Previous studies of biases in proportional reasoning [32, 33, 46] indicate that the answer is likely

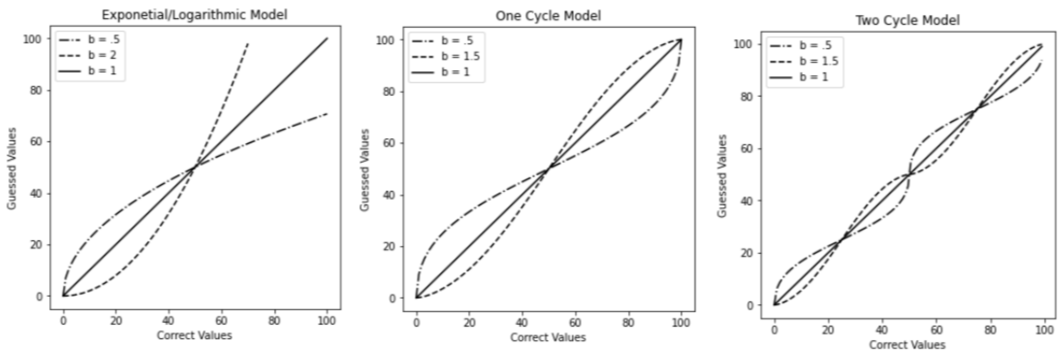


Figure 3: Bias models used to test if patterns of error are present when making graphical perception judgements (section 5.4). The models and their equations were specified by Hollands & Dyre [32]. The curves represent patterns of over- and underestimation of the guessed value compared to the true value in our data. The b value represents the magnitude of the bias, i.e., the farther the b value is from 1, the less accurate the individuals responses are. The center line represents no error.

yes. It is unclear, however, what bias model best describes these errors for general graphical perception tasks. If we can discover how children decode information and where their mistakes are, we can provide design aids that help mitigate cognitive effort and improve accuracy [66]. The foundational work on bias models by Spence & Krizel [66] demonstrated how people leverage certain reference points (endpoints, midpoints, quarters) to make numerical estimates. Hollands & Dyre [32] formalized this theory into the models we use in our analysis, several of which are demonstrated in fig. 3. These models have proved effective in predicting the errors children and adults make on other proportional reasoning tasks [32, 63]. Because of the similarity in nature of our tasks with this prior work, we believe that our participants will also use certain reference points as anchors for their judgements.

4 EXPERIMENTAL DESIGN & METHODS

Our goal in this study is to empirically test the accuracy of individuals on graphical perception tasks and compare the accuracy between age groups. The experiment followed many of the same procedures of previous graphical perception tests, which we describe in detail in section 4.4. We borrowed elements from developmental psychologists studying proportional reasoning to further curate our study for children (8–12 years old). The stimuli and wording of the questions were chosen from existing literature and refined through pilot studies (see fig. 4, section 4.1). To make a valid comparison between the two populations, both adults and children were tested with identical stimuli and question prompts. The independent variables tested were age, elementary perceptual task, and proportion. The dependent variable is the accuracy for each question.

The evaluation was administered to the child participants by an educator they were comfortable working with, while the adults completed the study remotely and asynchronously. Pilot testing verified that our study length, question difficulty, and feedback mechanism were appropriate for the younger population (see section 4.1).

All supplemental materials required to reproduce and replicate the study—including the stimuli, experiment code, collected data, and analysis code—are freely available at https://osf.io/ygrdv. To avoid issues stemming from postdiction such as hindsight bias, overconfidence in post hoc explanations, and underestimating uncertainty, we preregistered our study on the Open Science Framework (OSF) before running the experiment [50]. Our preregistered study design and analysis code is available at https://osf.io/crj2z.

4.1 Stimuli

The stimuli used in our study, illustrated in fig. 4, most closely resemble the stimuli presented in Cleveland & McGill's 1986 paper [16] since they are closest to the perceptual building blocks Cleveland & McGill refer to when initially proposing graphical perception studies [14]. Since we do not know the effect of distractors (additional graphical elements outside of the two tested) or how people use the different encodings in traditional charts (people decode pie charts using area, arc length, and angle), we use the most basic elements we can test [40, 67]. Additionally, the stimuli chosen closely resemble the encoding rankings seen in popular visualization textbooks [16, 48].

The stimuli chosen are rarely seen in such a simple format "in the wild." What they do represent are the basic encodings and building blocks for many types of visualizations. Our goal is to empirically test these encodings and enable other researchers and designers to determine how the encodings come together to create various visualizations. If we tested specific visualization types, i.e., bar charts or pie charts, our results would only be applicable to the visualizations tested. By testing the basic encodings, our results can be better extended to many types of visualizations.

The general premise is that each stimulus is a simple chart showing two data points with a single visual element each. One datum will be the integer **100** and the other datum will be **?**, a question mark. The participants will be asked what is the size of the smaller, knowing that the larger one is **100**. For example, when asked about Position Along a Common Axis, the question would be: "The one marked with a 100 is 100 blocks high. How high is the one marked with a ?. Type your answer below." The question is phrased in this way because the child participants may not not have experience with fractions yet [68]. This format rephrases the question seen in previous graphical perception studies to not bias the results of children who may lack a strong command of fractions. Pilot

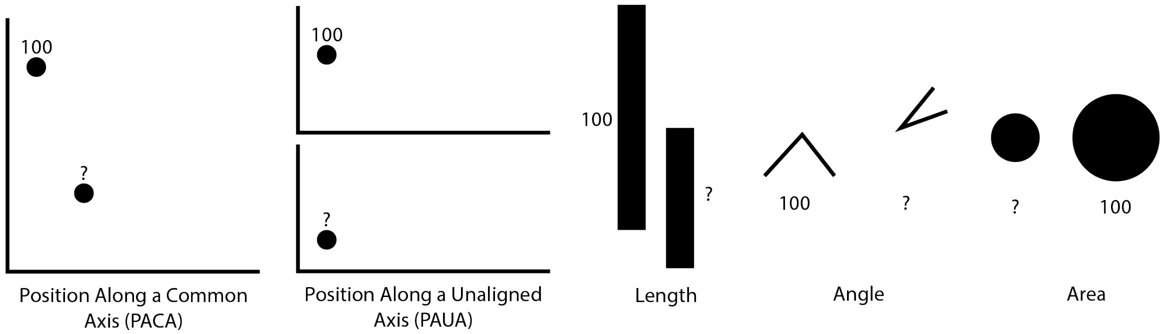


Figure 4: Illustrated versions of the stimuli we used. See https://osf.io/cbf28/ for the 95 actual stimuli (5 encodings × 19 questions). These encodings are commonly used to present quantitative data and ranked as most effective for the task [48]. The ordering of the stimuli and size of the stimuli were randomly generated. Each participant saw the same exact stimuli but in a different order.

testing and speaking with an educator who works closely with the children confirmed the children easily understood this format. One datum in each question was larger than the other, and the order in which the two associated visual elements were placed was randomly set. The stimuli for **?** were created as a proportion of the size of the one marked with **100**. Since previous studies have found that participants tend to answer in multiples of **5** or **10** [67], the proportions presented were [.05, .95] with steps of **.05**—thus 19 stimuli per encoding.

4.2 Encoding Types

In our study, we tested the judgements for five different encodings—(1) Position Along a Common Axis, (2) Position Along an Unaligned Axis, (3) Length, (4) Angle, and (5) Area. These are illustrated in fig. 4. Due to the age of the children, questions related to color hue or luminence would be difficult to explain in an understandable manner, so we did not include these encodings in the study. For further information about the creation of the stimuli, images of all the stimuli used, and exact details on how to recreate them, please refer to the Stimuli section of our preregistration at https://osf.io/crj2z.

4.2.1 Position Along a Common Axis (POCA). The position along a common axis stimuli resembles a dot or scatter plot with a vertical axis for the quantitative data in question. We chose this encoding rather than a bar chart as it may help reduce any perception of length as well.

4.2.2 Position Along an Unaligned Axis (POUA). We want the only major difference between this stimuli and the previous one to be the position of the vertical axis. This encoding is commonly seen in small multiples of scatterplots [70]. The heights of the visualizations are smaller than in the Position Along a Common Axis stimuli to ensure it fits on screen.

4.2.3 Length. The length judgement resembles a stacked bar chart, but only shows a pair of associated middle segments of the overall bar. The two bars cannot start or end near each other, so the encoding remains length and not position along a common axis. The labels are placed to the left and right of the bars' midpoints to try to maintain the encoding as length.

4.2.4 Angle. The stimuli used are pure angles rather than pie charts. In a pie chart, participants use other methods beyond angle to find proportions [40]. We want to keep in line with our previous stimuli of trying to test the graphical element alone with no other elements mixed in. The angles are randomly rotated by some random amount between 0–360°. When the angles all started at 0°, pilot participants were confused since they had difficulty breaking down the mental barrier of abstracting 100 to a 90° angle. Rotating the angle more closely resembles the random direction comparisons that may happen in pie or radial charts.

4.2.5 Area. For our area judgements we compared the area of two circles. We chose circles as the stimuli, as they are often seen as an encoding in node-link visualizations and bubble charts. The results from Cleveland & McGill's 1986 paper show blobs and circles were found to have very similar degrees of accuracy from the participants [16]. The area of the circle scaled linearly with the datum shown, rather than a naïve quadratic radius scaling.

4.3 Participants

After obtaining IRB approval for our study, we recruited from two different populations: children (μ_{age} = 9.91 years, range = 8–12; 19 male, 14 female) attending the Ivy After School Program and adult (μ_{age} = 24, range = 19–29; 13 male, 11 female) computer science students from Northeastern University, both located in the United States. Both groups were recruited via a flyer sent by email. Due to working with only one after-school program we had a connection to, the child participants were mostly second-generation Chinese-American students and from low income families receiving state-issued childcare vouchers. The age of the children ranged from 8–12 and they attended grades 3–5. The adult participants reflect a similar population to the three graphical perception studies conducted by Cleveland & McGill [14–16].

We chose our particular age range for the children based on multiple conversations with a developmental psychologist and the educators at the after-school program. The experts expressed concern that, younger than age 8, the children would not have the attention span nor cognitive abilities to complete the study. Additionally, until age 8, children often make logarithmically-biased errors on number-line estimation tasks [64]. This could bias errors not due to perception but due to a lack of numerical understanding [64]. Furthermore, the "common core" standards in education begin to introduce simple visualizations around Grade 2 [68]. By starting at Grade 3, we can assume that each student would have been introduced to the concept of a visualization and would likely be able to complete our study. We limited maximum age of children in our study to 12 for two reasons. First, 12 is generally considered the maximum age where a child could still be considered an elementary school student. Second, the after-school program only enrolls students 12 and younger. Sampling outside of the after-school program would have added additional complications to the study.

Since we are studying children, there are several factors that we must account for: attention span, knowledge level, and ability to understand directions. These factors dictated many of our design decisions for the study.

4.3.1 Attention Span. It is well known that children have a shorter attention span than adults. The coordinator of the after-school program stated that we could expect the attention span of the children to be approximately 30–45 minutes. We therefore limited the length of the study to less than 45 minutes. In work done by Mohring et al., it took around 10 minutes for children to answer 32 questions involving proportional reasoning in visualizations [46], and, in work conducted by Spence & Krizel, it took around 6 seconds per visualization [66]. With our proposed 95 visualizations to be shown to children, Mohring et al's timing implies a test 32 minutes long and Spence & Krizel's implies 10 minutes. During pilot testing and the actual study, all students were able to complete the study in under 45 minutes with delays and breaks included.

4.3.2 Knowledge level/Ability to understand directions. Students between the ages of 8–12 vary greatly in their cognitive abilities and skills they have learned and know [68]. With children we cannot assume that they will understand the directions or grasp the concepts being provided. Since they may have different ability levels when it comes to understanding the material, we provided practice questions with an alternate visual explanation to help the children understand the question. Additionally, we conducted the experiment synchronously with an educator in the room. Having a person familiar with the children and their behaviors helped assess if the students were comfortable and ready to move on to the data-gathering portion of the study.

4.4 Procedure

The study followed a similar format with children and adults. Both groups had a demographic questionnaire, practice questions, then the actual study. For the children, informed consent was obtained from each child's guardian and assent was obtained from the child themselves before starting the study. The study involved a practice section and then the actual experiment. In the practice section, the participant saw a total of 20 questions, 4 in each of 5 encoding types. In the data-gathering portion of the experiment, each participant saw at total of 95 questions, 19 in each of the 5 encoding types. A study website (code at https://osf.io/y5vsz/) was set up for both children and adults to collect the data. The stimuli presented and format of the study was identical for both children and adults during the data gathering portion of the study.

For the child study, a consent form was sent home asking for the parent or guardian's approval. Once the consent form was returned to the school, study sessions were set up between our educator collaborator and the students. The study was conducted one-on-one in person using the educator's computers.

> *COVID-19 safety statement: This study was run during August 2021 in the United States, in a county that the Center for Disease Control rated as having a substantial community transmission level (50–99 cumulative cases per 100k population or a cumulative NAAT viral test positivity result between 8.0–9.9% in the past 7 days) [11]. As the children were already attending the after-school program in person and the experiment was conducted by only their customary educator, our institution's Institutional Review Board assessed that participating in our study posed no additional risks to the children.*

The children's portion was conducted in person since it allowed for the child to feel comfortable and ask questions, as well as for the experimenter to gauge whether the participant understood the questions. The educator then walked the participant through an example question, building up the question one part at a time to not overwhelm them. The educator then manually clicked through the practice questions with the participant and asked them verbally for their answers. If a participant did not understand the first question in any given encoding type, the educator would display an alternate visual explanation of the question to clarify what was being asked. Every participant that began the study was able to complete the practice questions and the data gathering portion. Upon completion, each participant was awarded a $20 dollar Tango Card which could be redeemed at a store of their choice.

The adults completed their study asynchronously and remotely. The type of device they used and speed at which they completed the study could not be controlled for. A study website was sent to them that included a demographic survey, practice questions, and data-gathering portion. Like the children, each adult participant was awarded a $20 dollar Tango Card.

The data-gathering portion of the study was broken down into blocks with each block assessing one encoding type and containing 19 questions. In between each section, the participants were presented with a screen informing them that they had finished a section and could take a break. The encoding types and proportions were randomized using a counterbalanced Latin square design [74]. The sequence of the encoding types and the order of the proportions presented for each encoding type was assigned based on the participant number. This allowed each participant to see the exact same stimuli but in a different order, reducing any complications that could arise from ordering effects. At the end of the study, a final screen showed the participant their individualized ranking of encodings by effectiveness.

5 ANALYSIS

Our analysis plan and code was preregistered at https://osf.io/crj2z before we collected any data. All data and code used in the final analysis is available at https://osf.io/ygrdv.

5.1 Data preparation

5.1.1 Exclusion Criteria. We ensured that there were no outlier points and no outlier participants that could negatively affect the quality of our statistical analyses. To check that participants understood the questions, we checked to see if the data of correct values and answered values was highly positively correlated. Each task was checked for each individual. The Pearson's R correlation needed to be greater than .7 for the individual's data to be included. Any correlation above a .7 signifies a "High Positive Correlation," as stated by Mukaka [47], which is what we expect to see. Any participant who has a correlation less than .7 in any tasks was removed from the analysis for the study to remain a within subjects study. This cutoff is admittedly dichotomous but reduces errors introduced by human judgement in the analysis.

We also removed all outliers that may have been caused by a misinterpretation of the question. Absolute errors (|actual−guessed|) greater than 50 were removed. An error greater than 50 should not occur due to participant's regular judgement errors. An error of this size would mean a proportion of 20 was judged as 70, or vice versa. An error of 50 puts the participant severely over the halfway proportion reference point.

5.1.2 Distribution Testing. After the data were cleaned and prepared for the statistical analysis, we tested for normality. For each question, the absolute error between guessed and actual value was recorded. The children's data were aggregated by encoding type and tested for normal distributions. We visually tested for normality visually using Q-Q plots and quantitatively using the Shapiro-Wilk test [62]. From both visual inspection and p-values of the Shapiro Wilks test, the data was found to not come from a normal distribution for each encoding type. Therefore, nonparametric statistical tests were used in further analyses.

5.2 Hypothesis 1 Analysis

Hypothesis 1: Children have varying perceptual accuracy with different visual encodings.

Since the child data (1) is continuous, (2) has three or more tasks to compare, (3) is not normally distributed, and (4) comes from a single group, we use the Friedman test to determine if there are differences in error between encoding types. The Friedman test checks if the median absolute error is the same or different for each.

- H_0: Participants will have no difference in median absolute errors.
- H_1: The medians are not all equal, which we expect from prior studies on visual encoding rankings.

Since the Friedman test states only if all the medians are equal or if at least one median is different, we will also need to conduct a post-hoc analysis. We will use the Connover test to determine which encoding types have different absolute errors (|actual−guessed|). We will correct the p-values of the Connover test for multiple hypothesis testing with the Benjamini-Hochberg procedure [3, 18]. Since using p-values as a binary threshold for claiming a scientific finding or statistical significance used in isolation can lead to erroneous beliefs and poor decision making [28, 73], we augment the null hypothesis significance testing with interval estimation of effect sizes [19, 21]. We construct 95% confidence intervals using the mean of the absolute errors for each encoding type. We use the bias-corrected and accelerated (BCa) bootstrapping method for computing them [23]. By visually inspecting the confidence intervals in conjunction with the post-hoc Connover test, the rankings for the different tasks will be apparent.

5.3 Hypothesis 2 Analysis

Hypothesis 2: Children's overall elementary perceptual task accuracy will be lower than adults.

We want to compare the two independent populations of adults and children to see if the dependent variable of absolute error for each encoding type is the same or different. The data used for the analysis is the mean absolute error for each participant for each encoding type. The means are grouped by which population and which encoding type they belong to. Our previous analysis determined that the data is not normally distributed. We followed this analysis by conducting an F test to test if the variances are equal between the adult and child populations for each encoding type. We tested a null hypothesis H_0 that the data has an equal variance at an alpha of .05 and corrected the p-values for multiple hypothesis testing using the Benjamini-Hochberg procedure [3]. We found that there are unequal variances for each encoding type for the adult and child populations.

Because we are testing if there is a difference in a continuous dependent variable between two independent populations, we use the Mann-Whitney U test, which is a measure of the number of times a randomly chosen value from one group exceeds a randomly chosen value from the other group. To interpret the results of the analysis, we can use the language below:

- H_0: the distribution of scores for the two groups are equal.
- With unequal variances, the alternate hypothesis H_1 is that the mean ranks of the two groups are not equal.

In addition to the Mann-Whitney test, we present 95% confidence intervals to help indicate the range of plausible values. The confidence intervals were created for the absolute error means for each participant using bias-corrected and accelerated (BCa) bootstrapped confidence intervals [23].

5.4 Hypothesis 3 Analysis

Hypothesis 3: Children will exhibit patterns of bias in errors when making graphical perception judgements.

To test which bias model is best, we used the Akaike's Information Criterion (AIC) [1]. The AIC tests the amount of information lost by a given model. The model with the lowest AIC score will be considered the highest-quality model.

The rules to interpret the model scores are outlined below, as per Burnham & Anderson [9]. The main test statistic is $\Delta_i = AIC_i - AIC_{min}$, where AIC_i is the ith model and AIC_{min} is the model with the lowest score. The Δ_i scores can be interpreted as:

- if $\Delta_i < 2$, then there is substantial support for the i-th model or the evidence against it is marginal.

- if $2 < \Delta_i < 4$, then there is less support for the i-th model.
- if $4 < \Delta_i < 7$, then there is considerably less support for the i-th model.
- if $\Delta_i > 10$, the model has essentially no support.

The larger Δ_i is, the greater the information loss for that model. Therefore, for our purposes, if the $\Delta_i > 4$ we can say that the information loss for this model is great enough that a different model better supports data gathered. After the bias models are evaluated for the children, the same analysis is run for the adults for comparison.

6 RESULTS

Per our preregistered exclusion criteria at https://osf.io/crj2z, we removed data from four child and three adult participants because their data had low correlation between responses and actual answers—defined as a Pearson's correlation coefficient < 0.7 for at least one of the tasks. We deviated from our preregistered plan in several ways: (1) We corrected a bug for removing outlier points in preregistration section 3.1, "Exclusion Criteria for Individuals," to ensure it matched the plan from section 1 of the preregistration. (2) Because of a mistake with the data collection code, 10 of the 25 adult participants had the exact same ordering of the stimuli presented, which should have been random. (3) Finally, the analysis of the bias models was adjusted to represent the equations set forth by Hollands & Dyre [32] as well as Slusser and Barth [63]. In the original plan (section 1.5, "Bias Models"), a fixed scaling parameter *a* was included in the single-cycle and two-cycle models. This caused the bias models to have unequally spaced reference points, which runs in contradiction to previous literature. The variable was changed to be 100, and the analysis was run using the equally-spaced reference point models. The changes are annotated and described in the analysis code found at https://osf.io/ygrdv.

6.1 Hypothesis 1

From the results of the Friedman test (Fr = 68.93, df = 5,32, p<.001) and interval estimation of effect size in fig. 5A, there is strong evidence that there is a meaningful difference in medians of the absolute error between the encodings. There is strong evidence for both children and adults that the accuracy is *generally* ranked as follows (most accurate to least accurate): (1) Position Along a Common Axis, (2) Length, (3) Position Along an Unaligned Axis, and (4|5) Area and Angle. There was very little support for there being a difference in accuracy between area and angle.

6.2 Hypothesis 2

For our second hypothesis, our evidence—p<.001 for each encoding and the interval estimation of effect size in fig. 5C—strongly supports the claim that children are meaningfully less accurate than adults with each encoding type.

6.3 Hypothesis 3

The results of our tests on the bias models are shown in fig. 6, as per [63]. We have strong evidence that the linear model without constants does not describe the data best for each chart type, with the exception of Length. For Position Along a Common Axis and Angle, no bias model conclusively describes the data best. For Position Along an Unaligned Axis, there is strong evidence that the exponential model describes the data best. Finally, for the Area encoding, there is some evidence to support linear with constants as the model to best support the data.

7 DISCUSSION

D1: Children and adults have similar patterns of graphical perception for the different encodings. The most important of our findings is the similarity in results between children and adults. The effectiveness rankings, relative differences in error between the different encodings, and the patterns of bias across encoding types are markedly similar for both children and adults. The similarity in effectiveness rankings strengthens Cleveland & McGill's [15] claim that the decoding of visualizations is a preattentive visual task. However, since there are distinct biases across tasks for both groups, researchers should not lump the perceptual processes involved in each of these tasks into the same category. Furthermore, the similarity in bias models and patterns of bias across ages suggests these biases are consistent across development. This finding also supports Spence & Krizel's [66] conclusion that children and adults make similar errors when decoding information from visualizations. Thus, the combination of our findings and previous literature leads us to believe that children and adults share the same cognitive processes when decoding visual information.

D2: Children are less accurate than adults on graphical perception tasks. Though there are abundant similarities between the child and adult populations, there is also one clear difference: children are less accurate when decoding visual information than adults. For every task, children's mean error was approximately 2.5 percentage points worse than the adults'. As children mature, they are likely to encounter more visualizations and become more skilled at estimating the sizes and computing ratios, leading to improved graphical perception abilities.

D3: While children show a greater variety of individual differences than adults, both populations show a diversity in encoding rankings and overall accuracy. Aggregated data from the children shows very similar patterns of error as with the adults, but there are individual differences in the rankings. For the 33 child participants, the encodings were ranked in 22 different ways, and for the 24 adult participants, the encodings were ranked in 8 different ways (fig. 5D). These rankings were determined by simply using the mean absolute error aggregated for each task. There is preliminary indication that children display a greater variability on which tasks they are most accurate on. Thus, the aggregation of data misses a larger diversity of thought and decoding that might be present in children.

D4: Both adults and children consistently underestimate values in Position Along an Unaligned Axis tasks. When we examine the bias models for Hypothesis 3 (section 5.4), we can reject our null hypothesis for each of the encodings except Position Along an Unaligned Axis for the children and for all the encodings for adults. Looking at fig. 6B, we can see the tendency for participants to underestimate the correct response. The exponential model (red line) runs below the the linear model (dotted line) indicating a pattern of underestimation. With all other models

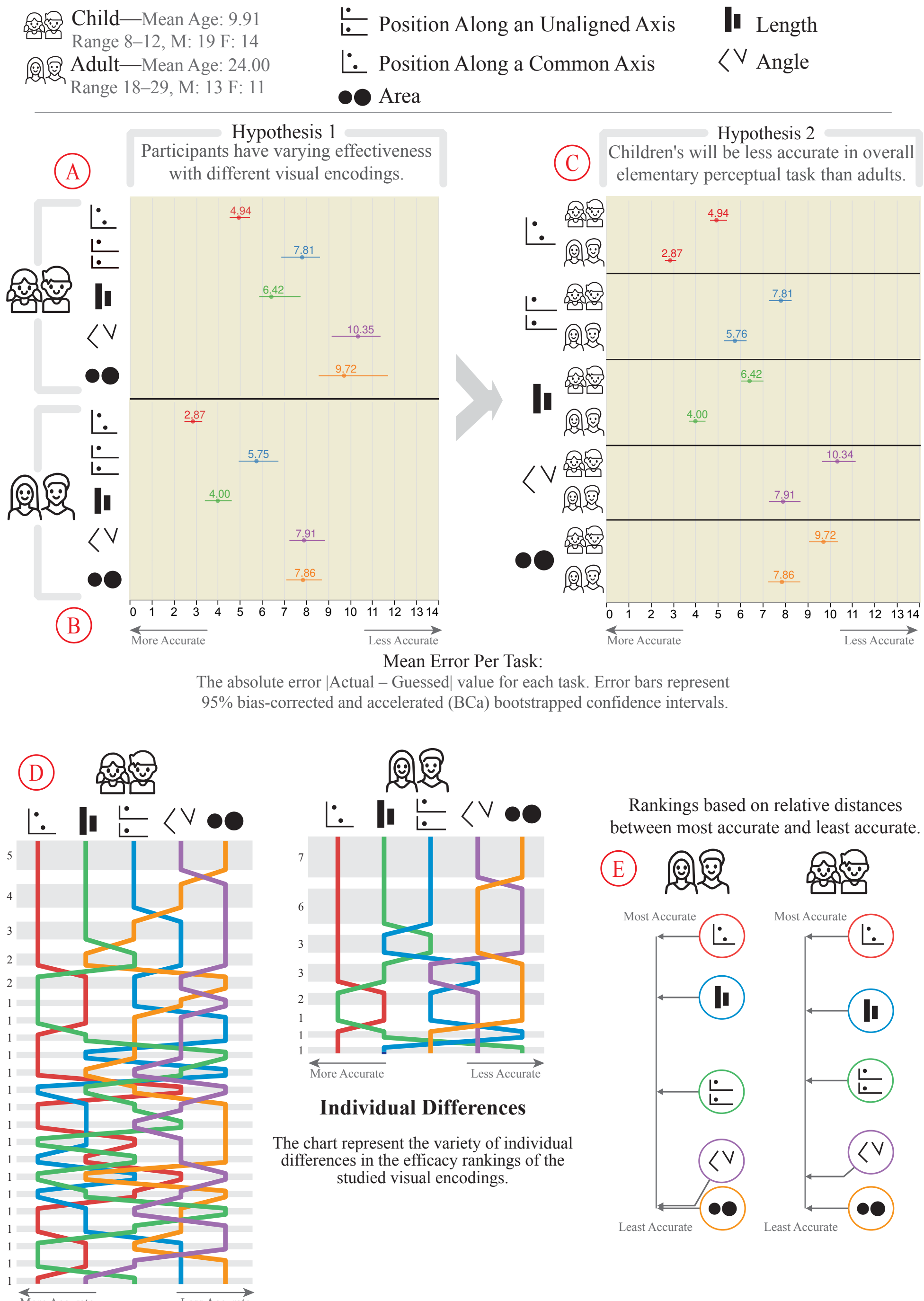
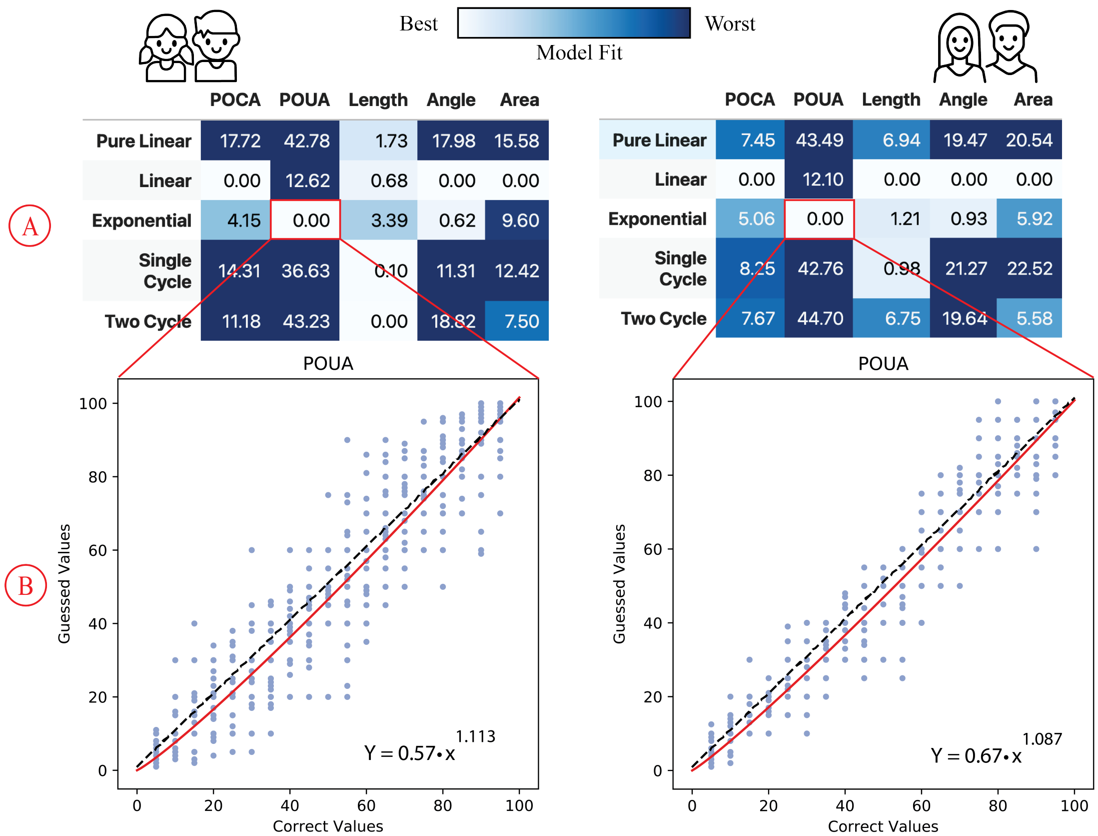


Figure 5: Summative results for Hypothesis 1 and 2 and an exploratory analysis of individual differences in encoding rankings. In A, B, and C the error bars show 95% bias-corrected and accelerated (BCa) bootstrapped confidence intervals [23]. (A *rough* rule of thumb for reading 95% CIs is that if two intervals overlap by less than 1/4 of their average length, then the comparison will have p < .05 [22].) The mean absolute error for each encoding is shown in A for children and B for adults. In C, the previous two charts are rearranged to compare children with adults. Children are clearly less accurate when using each of the encodings. The exploratory analysis included, D, shows the variation in encoding rankings among individual children (left) and adults (right). Each line represents an encoding, ranked left-to-right in increasing mean absolute error for each task. The grey rows are sized to represent the count of individuals with a shared ranking. E.g., the top row shows that 5 children ranked Position Along a Common Axis as most accurate, followed by Length, Position Along an Unaligned Axis, Angle, and lastly Area. The line-row intersections show the encoding ranking for that row. Children displayed a larger variety of individual differences in encoding rankings than adults. Finally, E shows more simply the *overall* rankings we found for adults and children.



Children:

	POCA	POUA	Length	Angle	Area
Pure Linear	17.72	42.78	1.73	17.98	15.58
Linear	0.00	12.62	0.68	0.00	0.00
Exponential	4.15	0.00	3.39	0.62	9.60
Single Cycle	14.31	36.63	0.10	11.31	12.42
Two Cycle	11.18	43.23	0.00	18.82	7.50

Adults:

	POCA	POUA	Length	Angle	Area
Pure Linear	7.45	43.49	6.94	19.47	20.54
Linear	0.00	12.10	0.00	0.00	0.00
Exponential	5.06	0.00	1.21	0.93	5.92
Single Cycle	8.25	42.76	0.98	21.27	22.52
Two Cycle	7.67	44.70	6.75	19.64	5.58

Figure 6: Summative results for Hypothesis 3. Ⓐ The table shows the Δ_i for the Akaike's Information Criterion (AIC) for each model and encoding combination, separately for children (left) and adults (right). The model with the lowest Δ_i is best. See section 5.4 and Burnham & Anderson [9] for details. Though no model is consistently best for either population, both populations share very similar patterns of error. This indicates that they likely use similar cognitive processes when decoding the information. Ⓑ The charts show how the exponential bias model for Position Along an Unaligned Axis fits the participant responses. The clear deviation from the linear model along with the large difference in AIC scores indicates a pattern of underestimation in participant's responses. For both populations, only the exponential model for the Position Along an Unaligned Axis is definitively the best choice to minimize information loss.

for having $\Delta_i > 10$, indicating essentially no support, Position Along an Unaligned Axis is the only encoding where a clear pattern emerges. One possibility is that participants tended toward lower estimates because the reference was always 100 and they were asked for a value under 100. Future research could try to include test values above 100 to see if this holds when asked about values above and below.

D5: Contradicting prior studies, Position Along an Unaligned Axis was less accurate than Length, and there is little evidence for a difference between Area and Angle. Prior studies suggest that Position Along an Unaligned Axis is more accurate than Length [15, 30]—but we found the converse to be true in both populations. Our results further indicated that there was no meaningful difference in task performance between Area and Angle, contradicting established guidance that Angle is more effective [15, 30]. The difference in results likely stems from the stimuli used. Cleveland & McGill set out to test the "perceptual building blocks" of visualizations [14]. Their stimuli, and other studies afterwards, mix several building blocks and add distractors, which may result in the elementary tasks not being accurately examined [30, 40, 67]. We test what we believe are the most basic elements extracted from popular visualization types. Our hope is that by isolating the stimuli to their most basic level we can more accurately rank the elementary building blocks of visualizations. Then designers and researchers can decide from empirical evidence how best to build their visualizations.

7.1 Design Recommendations

We build each of our design recommendations based on the preceding discussion points, which we reference in parenthesis.

DR1 (D1, D2): **We recommend that designers follow the same encoding ranking guidelines for both adults and children.** From our findings, it can be seen that, at least in regards to graphical perception, children use similar cognitive processes as adults. This, however, does not mean that they will be able to effectively interpret visualizations of similar complexity. It is important to keep in mind that children have significantly less experience with visualizations and may not be able to decipher meaning from more complex ones. Moreover, our results indicate that children are approximately 2.5 percentage points less accurate than adults at these tasks.

DR2 (D2, D4, D5): From Hypothesis 2 (section 5.3) it is apparent that children are less accurate when decoding visualizations. **One strategy that may help mitigate children's loss in accuracy is adding gridlines.** Gridlines may improve children's judgment by providing additional reference points. This may be particularly helpful when creating visualizations that rely heavily on Position Along a Common Axis, Position Along an Unaligned Axis, and Length. Work by Hollands & Dyre [32] and Heer & Bostock [30] theoretically and experimentally substantiate this. Heer & Bostock in their third crowdsourced experiment showed that adding gridlines increased perceptual accuracy for Position Along a Common Axis tasks. In Hollands & Dyre's [32] cyclic power model, the more reference points the more the amplitude of the bias decreases. Therefore, adding gridlines and increasing the amount of reference points would decrease participant's perceptual bias, as evidenced by the experimental work done by Heer & Bostock.

DR3 (D3): **Since children's encoding rankings are more varied, visualization creators and teachers may want to consider other factors beyond accuracy in their designs.** This extends to the practical application and interpretation of the results of graphical perception studies. Though the differences between the encodings may be meaningfully different, the differences are still on the smaller side. When considering a visualization, the designer may want to contemplate if the difference in accuracy will affect the user's ability to perceive the patterns in a visualization. Choosing a better encoding that provides slightly more accuracy may not be worth the loss in creativity or user enjoyment [59].

DR4 (D5): **We advise the reader to look deeper at the data behind graphical perception guidelines and future researchers to present these rankings in a way that better mirrors the results.** When the rankings are presented as an ordered list, it gives the impression that one encoding is equidistant from another. In fig. 5E, we present an alternative way to express the results. While this display does not encapsulate the variability of the participants' data, it is a step in the right direction of showing the differences in error. By presenting the results to more closely reflect the findings of our studies, visualization designers may correctly assume that ●● Area and Angle are interchangeable.

7.2 Limitations and Future work

While our work furthers the research on how children perceive visualizations, it is not without limitations. We broaden and confirm the results of graphical perception studies to a population not usually assessed in visualization research [31]. This is a helpful first step, but we continue to encounter the problem of using a narrow sample to generalize to a broader population. The children were all sampled from one after-school program, and adults were all computer science students at one university. The variety of individual differences and spread of results indicates that we need to conduct these studies with broader, more diverse populations to confirm what we found.

To increase the diversity of participants, graphical perception studies could look at a broader age range of children. Our findings show that children are less accurate in graphical perception tasks, but we have not determined when their accuracy becomes comparable to adults. Future work could investigate both younger and older children to see how graphical perception may change or develop over the years. Additional work could also be conducted to help better understand what cognitive processes and skills are involved in improved graphical perception ability.

In addition to diversifying the types of populations studied, future researchers could also broaden the types of studies and analysis done with children to better understand how they perceive visualizations. Our work quantifies the kinds of patterns and errors children and adults make when decoding information from charts. By looking at the patterns of errors for both populations, we hypothesize that they use similar cognitive processes to make their judgements. Future researchers could conduct more qualitative studies to better understand these cognitive processes and identify which visualizations are easier to understand. The educator administering the study noticed that several children easily understood the ●● Area encoding but struggled to process the others. Future qualitative work could explore what data abstractions are more intuitive and comprehensible to children.

To offer a bit of encouragement to future researchers interested in working with children, we leave you with one last observation. When designing studies for children, it can be difficult to think of the many design choices one must make to accommodate them. While this is true, we found that the children in our study were eager to help. Though our study may not be well-characterized as fun, the children were engaged in the study tasks and enjoyed seeing their individualized encoding ranking at the end of the study. We also strongly advise partnering with someone familiar with the students. Having collaborators who know the participants can help in all aspects of the study, from designing the study to gathering and assessing the participants. Other researchers have given similar advice, and this has been associated with a rise in visualization studies involving children [2]. We hope this trend continues as more researchers recognize the importance of studying how children interpret visualizations.

8 CONCLUSION

Children's ability to understand visualizations is critical to their success in building foundational knowledge and skills. And yet visualizations designed for children rely for the most part on graphical perception research conducted with adults. To create better visualizations, we gathered empirical evidence to support design guidelines for children. We take a step towards solving this problem via a graphical perception study with children and adults, investigating how they perceive visualizations. Our results indicate that children process visualizations in very similar ways to adults but have difficulty making judgements as accurately. Our results lead us to propose several guidelines not just for children's visualizations, but also for graphical perception results and interpretation.

We hope that our work will prompt more visualization researchers to conduct studies with this important population of visualization users. While adults may encounter visualizations sporadically, it is likely that every child going to school will encounter visualizations and charts daily. By working with this population and making effective visualizations for them, we can directly impact their lives and create more informed scholars and citizens.

ACKNOWLEDGMENTS

This project could not have succeeded without the help of Matthew Kay, Lane Harrison, and Yiren Ding. Their openness in discussing the methodology and impact of the project helped form the paper. We also greatly appreciate Jane Kokernak's editing assistance.

REFERENCES

[1] Hirotogu Akaike. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*. Springer, 199–213.

[2] Basak Alper, Nathalie Henry Riche, Fanny Chevalier, Jeremy Boy, and Metin Sezgin. 2017. Visualization literacy at elementary school. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5485–5497. https://doi.org/10.1145/3025453.3025877

[3] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[4] Fearn Bishop, Johannes Zagermann, Ulrike Pfeil, Gemma Sanderson, Harald Reiterer, and Uta Hinrichs. 2019. Construct-a-vis: Exploring the free-form visualization processes of children. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 451–460. https://doi.org/10.1109/TVCG.2019.2934804

[5] Rita Borgo, Johannes Kehrer, David HS Chung, Eamonn Maguire, Robert S Laramee, Helwig Hauser, Matthew Ward, and Min Chen. 2013. Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications.. In *Eurographics (State of the Art Reports)*. 39–63. https://doi.org/10.2312/conf/EG2013/stars/039-063

[6] Katy Börner, Adam Maltese, Russell Nelson Balliet, and Joe Heimlich. 2016. Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization* 15, 3 (2016), 198–213. https://doi.org/10.1177/1473871615594652

[7] Jeremy Boy, Ronald A Rensink, Enrico Bertini, and Jean-Daniel Fekete. 2014. A principled way of assessing visualization literacy. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1963–1972. https://doi.org/10.1109/TVCG.2014.2346984

[8] Ty W Boyer, Susan C Levine, and Janellen Huttenlocher. 2008. Development of proportional reasoning: where young children go wrong. *Developmental psychology* 44, 5 (2008), 1478. https://doi.org/10.1037/a0013110

[9] Kenneth P Burnham and David R Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research* 33, 2 (2004), 261–304. https://doi.org/10.1177/0049124104268644

[10] C Melody Carswell. 1992. Choosing specifiers: An evaluation of the basic tasks model of graphical perception. *Human factors* 34, 5 (1992), 535–554. https://doi.org/10.1177/001872089203400503

[11] CDC. 2021. COVID Data Tracker. https://covid.cdc.gov/covid-data-tracker/#county-view

[12] Chaomei Chen. 2005. Top 10 unsolved information visualization problems. *IEEE computer graphics and applications* 25, 4 (2005), 12–16. https://doi.org/10.1109/MCG.2005.91

[13] Fanny Chevalier, Nathalie Henry Riche, Basak Alper, Catherine Plaisant, Jeremy Boy, and Niklas Elmqvist. 2018. Observations and reflections on visualization literacy in elementary school. *IEEE computer graphics and applications* 38, 3 (2018), 21–29. https://doi.org/10.1109/MCG.2018.032421650

[14] William S Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554. https://doi.org/10.2307/2288400

[15] William S Cleveland and Robert McGill. 1985. Graphical perception and graphical methods for analyzing scientific data. *Science* 229, 4716 (1985), 828–833. https://doi.org/10.1126/science.229.4716.828

[16] William S Cleveland and Robert McGill. 1986. An experiment in graphical perception. *International Journal of Man-Machine Studies* 25, 5 (1986), 491–500. https://doi.org/10.1016/S0020-7373(86)80019-0

[17] William S Cleveland and Robert McGill. 1987. Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society: Series A (General)* 150, 3 (1987), 192–210. https://doi.org/10.2307/2981473

[18] William Jay Conover. 1998. *Practical nonparametric statistics*. Vol. 350. John Wiley & Sons.

[19] Geoff Cumming. 2014. The New Statistics: Why and How. *Psychological Science* 25, 1 (2014), 7–29. https://doi.org/10.1177/0956797613504966

[20] Frances R Curcio. 1987. Comprehension of mathematical relationships expressed in graphs. *Journal for research in mathematics education* 18, 5 (1987), 382–393. https://doi.org/10.2307/749086

[21] Pierre Dragicevic. 2016. *Fair Statistical Communication in HCI*. Springer International Publishing, Cham, 291–330. https://doi.org/10.1007/978-3-319-26633-6_13

[22] Pierre Dragicevic. 2016. Fair statistical communication in HCI. In *Modern statistical methods for HCI*. Springer, 291–330.

[23] Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American statistical Association* 82, 397 (1987), 171–185. https://doi.org/10.2307/2289144

[24] Susan N Friel, Frances R Curcio, and George W Bright. 2001. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in mathematics Education* 32, 2 (2001), 124–158. https://doi.org/10.2307/749671

[25] Johannes Gäbler, Christoph Winkler, Nóra Lengyel, Wolfgang Aigner, Christina Stoiber, Günter Wallner, and Simone Kriglstein. 2019. Diagram safari: A visualization literacy game for young children. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 389–396. https://doi.org/10.1145/3341215.3356283

[26] Lars Grammel, Melanie Tory, and Margaret-Anne Storey. 2010. How information visualization novices construct visualizations. *IEEE transactions on visualization and computer graphics* 16, 6 (2010), 943–952. https://doi.org/10.1109/TVCG.2010.164

[27] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. 2013. Lineup: Visual analysis of multi-attribute rankings. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2277–2286. https://doi.org/10.1109/TVCG.2013.173

[28] Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology* 31, 4 (2016), 337–350. https://doi.org/10.1007/s10654-016-0149-3

[29] Charlotte Hanisch, Jürgen Konczak, and Christian Dohle. 2001. The effect of the Ebbinghaus illusion on grasping behaviour of children. *Experimental Brain Research* 137, 2 (2001), 237–245. https://doi.org/10.1007/s002210000655

[30] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 203–212. https://doi.org/10.1145/1753326.1753357

[31] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. Most people are not WEIRD. *Nature* 466, 7302 (2010), 29–29. https://doi.org/10.1038/466029a

[32] JG Hollands and Brian P Dyre. 2000. Bias in proportion judgments: the cyclical power model. *Psychological review* 107, 3 (2000), 500. https://doi.org/10.1037//0033-295X.107.3.500

[33] Michelle A Hurst and Sara Cordes. 2018. Attending to relations: Proportional reasoning in 3-to 6-year-old children. *Developmental psychology* 54, 3 (2018), 428. https://doi.org/10.1037/dev0000440

[34] Elaine Huynh, Angela Nyhout, Patricia Ganea, and Fanny Chevalier. 2020. Designing Narrative-Focused Role-Playing Games for Visualization Literacy in Young Children. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 924–934. https://doi.org/10.1109/TVCG.2020.3030464

[35] Yoonkyung Jeong, Susan C Levine, and Janellen Huttenlocher. 2007. The development of proportional reasoning: Effect of continuous versus discrete quantities. *Journal of Cognition and Development* 8, 2 (2007), 237–256. https://doi.org/10.1080/15248370701202471

[36] Pete R Jones and Tessa M Dekker. 2018. The development of perceptual averaging: Learning what to do, not just how to do it. *Developmental Science* 21, 3 (2018), e12584. https://doi.org/10.1111/desc.12584

[37] Yael Kali and Marcia C Linn. 2008. Designing effective visualizations for elementary school science. *The elementary school journal* 109, 2 (2008), 181–198. https://doi.org/10.1086/590525

[38] Jennifer A Kaminski and Vladimir M Sloutsky. 2013. Extraneous perceptual information interferes with children's acquisition of mathematical knowledge. *Journal of Educational Psychology* 105, 2 (2013), 351. https://doi.org/10.1037/a0031040

[39] Helen Kennedy and Rosemary Lucy Hill. 2018. The feeling of numbers: Emotions in everyday engagements with data and their visualisation. *Sociology* 52, 4 (2018), 830–848. https://doi.org/10.1177/0038038516674675

[40] Robert Kosara. 2019. Evidence for area as the primary visual cue in pie charts. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 101–105. https://doi.org/10.1109/VISUAL.2019.8933547

[41] Kevin Lai, Julio Cabrera, Jonathan M Vitale, Jacquie Madhok, Robert Tinker, and Marcia C Linn. 2016. Measuring graph comprehension, critique, and construction

in science. *Journal of Science Education and Technology* 25, 4 (2016), 665–681. https://doi.org/10.1007/s10956-016-9621-9

[42] Sukwon Lee, Bum Chul Kwon, Jiming Yang, Byung Cheol Lee, and Sung-Hee Kim. 2019. The correlation between users' cognitive characteristics and visualization literacy. *Applied Sciences* 9, 3 (2019), 488. https://doi.org/10.3390/app9030488

[43] Tom Lowrie and Carmel M Diezmann. 2009. National numeracy tests: A graphic tells a thousand words. *Australian Journal of Education* 53, 2 (2009), 141–158. https://doi.org/10.1177/000494410905300204

[44] Andrew Manches and Lydia Plowman. 2021. Smart toys and children's understanding of personal data. *International Journal of Child-Computer Interaction* (2021), 100333. https://doi.org/10.1016/j.ijcci.2021.100333

[45] Wenke Möhring, Andrea Frick, and Nora S Newcombe. 2018. Spatial scaling, proportional thinking, and numerical understanding in 5-to 7-year-old children. *Cognitive Development* 45 (2018), 57–67. https://doi.org/10.1016/j.cogdev.2017.12.001

[46] Wenke Möhring, Nora S Newcombe, Susan C Levine, and Andrea Frick. 2016. Spatial proportional reasoning is associated with formal knowledge about fractions. *Journal of Cognition and Development* 17, 1 (2016), 67–84. https://doi.org/10.1080/15248372.2014.996289

[47] Mavuto M Mukaka. 2012. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal* 24, 3 (2012), 69–71.

[48] Tamara Munzner. 2014. *Visualization analysis and design.* CRC press.

[49] Marko Nardini, Rachael Bedford, and Denis Mareschal. 2010. Fusion of visual cues is not mandatory in children. *Proceedings of the National Academy of Sciences* 107, 39 (2010), 17041–17046. https://doi.org/10.1073/pnas.1001699107

[50] Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2600–2606. https://doi.org/10.1073/pnas.1708274114

[51] Darko Odic. 2018. Children's intuitive sense of number develops independently of their perception of area, density, length, and time. *Developmental Science* 21, 2 (2018), e12533. https://doi.org/10.1111/desc.12533

[52] Rene S Parmar and Barbara R Signer. 2005. Sources of error in constructing and interpreting graphs: A study of fourth-and fifth-grade students with LD. *Journal of learning disabilities* 38, 3 (2005), 250–261. https://doi.org/10.1177/00222194050380030601

[53] Evan M Peck, Sofia E Ayuso, and Omar El-Etr. 2019. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3290605.3300474

[54] Steven Pinker. 1990. A theory of graph comprehension. *Artificial intelligence and the future of testing* (1990), 73–126.

[55] Chiara Piroddi, Federica Fragapane, and Christine Guthry. 2018. *Planet Earth: infographics for discovering our world.* White Star Kids.

[56] Rebecca Pop. 2021. Who Will Shape the Future of Data Visualization? Today's Kids! https://medium.com/nightingale/who-will-shape-the-future-of-data-visualizations-todays-kids-75d1111d346e

[57] Vicky Rideout. 2015. The common sense census: Media use by tweens and teens. (2015).

[58] Puripant Ruchikachorn, Patchar Duangklad, and Thanisara Ruangdej. [n.d.]. A Case Study of Data Visualization and Storytelling Workshop for Middle School Students. ([n. d.]).

[59] Bahador Saket, Alex Endert, and John Stasko. 2016. Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*. 133–142. https://doi.org/10.1145/2993901.2993903

[60] Priti Shah and James Hoeffner. 2002. Review of graph comprehension research: Implications for instruction. *Educational psychology review* 14, 1 (2002), 47–69. https://doi.org/10.1023/A:1013180410169

[61] Priti Shah, Richard E Mayer, and Mary Hegarty. 1999. Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of educational psychology* 91, 4 (1999), 690. https://doi.org/10.1037/0022-0663.91.4.690

[62] S Shaphiro and M Wilk. 1965. An analysis of variance test for normality. *Biometrika* 52, 3 (1965), 591–611. https://doi.org/10.2307/2333709

[63] Emily Slusser and Hilary Barth. 2017. Intuitive proportion judgment in number-line estimation: Converging evidence from multiple tasks. *Journal of Experimental Child Psychology* 162 (2017), 181–198. https://doi.org/10.1016/j.jecp.2017.04.010

[64] Emily B Slusser, Rachel T Santiago, and Hilary C Barth. 2013. Developmental change in numerical estimation. *Journal of Experimental Psychology: General* 142, 1 (2013), 193. https://doi.org/10.1037/a0028560

[65] Ian Spence. 1990. Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance* 16, 4 (1990), 683. https://doi.org/10.1037//0096-1523.16.4.683

[66] Ian Spence and Peter Krizel. 1994. Children's perception of proportion in graphs. *Child Development* 65, 4 (1994), 1193–1213. https://doi.org/10.1111/j.1467-8624.1994.tb00812.x

[67] Justin Talbot, Vidya Setlur, and Anushka Anand. 2014. Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2152–2160. https://doi.org/10.1109/TVCG.2014.2346320

[68] Common Core Standards Writing Team. 2013. Progressions for the Common Core State Standards in Mathematics.

[69] The New York Times. [n.d.]. What's going on in this graph? https://www.nytimes.com/column/whats-going-on-in-this-graph

[70] Edward R Tufte, Nora Hillman Goeler, and Richard Benson. 1990. *Envisioning information.* Vol. 2. Graphics press Cheshire, CT.

[71] Ellen Storey Vasu and Ann C Howe. 1989. The effect of visual and verbal modes of presentation on children's retention of images and words. *Journal of Research in Science Teaching* 26, 5 (1989), 401–407. https://doi.org/10.1002/tea.3660260505

[72] Howard Wainer. 1980. A test of graphicacy in children. *Applied Psychological Measurement* 4, 3 (1980), 331–340. https://doi.org/10.1177/014662168000400305

[73] Ronald L Wasserstein and Nicole A Lazar. 2016. The ASA statement on p-values: context, process, and purpose. https://doi.org/10.1080/00031305.2016.1154108

[74] Evan James Williams. 1949. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry* 2, 2 (1949), 149–168. https://doi.org/10.1071/CH9490149