Journal of Data Science xx (xx), 1–??
July 2022
DOI: 0000

Evaluating Perceptual Judgements on 3D Printed Bar Charts

Tyler Wiederich [??,0] and Susan VanderPlas[??]

[1]Department of Statistics, University of Nebraska-Lincoln, United States of America

Abstract

~~The use of~~ Graphical design principles typically recommend minimizing the dimensionality of a visualization - for instance, using only 2 dimensions for bar charts rather than providing a 3D ~~data visualizations has limitations when the third dimension does not convey any additional information to the viewer. Numerous studies advocate to avoid these types of graphs whenever possible, but these studies are almost entirely focused on the~~ rendering, because this extra complexity may result in a decrease in accuracy. This advice has been oft repeated, but the underlying experimental evidence is focused on fixed 2D projections of ~~the~~ 3D ~~graphs. This paperdescribes the partial replication of a well-known paper in data visualization and its adaptation to focus on~~ charts. In this paper, we describe an experiment which attempts to establish whether the decrease in accuracy extends to 3D ~~printed bar charts. While current results from our studydo not show differences between 2D and~~ virtual renderings and 3D printed charts. We replicate the grouped bar chart comparisons in the 1984 Cleveland & McGill study, assessing the accuracy of numerical estimates using different types of 3D ~~graphs, we will use this study to provide a baseline for future studies dealing with graphics in 3D environments~~and 2D renderings.

Keywords *graphics; 3D bar charts; 3D printing.*

1 Introduction

~~? published a paper that sets up the foundation for~~ Good communication requires both that the information be transmitted correctly and that the intended recipient be able to decode and understand the transmitted information accurately. In order to communicate effectively, we must use graphical forms that accurately convey information relevant to the task in question. In many cases, this means we must understand how accurately people ~~extract quantitative information using~~ can read quantitative information off of charts. While accuracy is not the only quantity of interest in graphical investigations (?), it is an important factor in assessing the utility of many different data graphics.

The accuracy of graphical forms has been studied for almost a century (????), as new ways of representing information evolve, we must revisit old studies to determine whether these representations have the same limitations as previous versions. This is particularly true in areas like graphics which are affected by the immense technological innovation in hardware and software which has taken place since the early 1990s.

1.1 Elementary Graphical Tasks

? established the comparative accuracy of different "elementary perceptual tasks" (EPTs). Elementary Perceptual Tasks, according to these experiments, include assessing graphical elements

*Tyler Wiederich Email: twiederich2@huskers.unl.edu or susan.vanderplas@unl.edu.

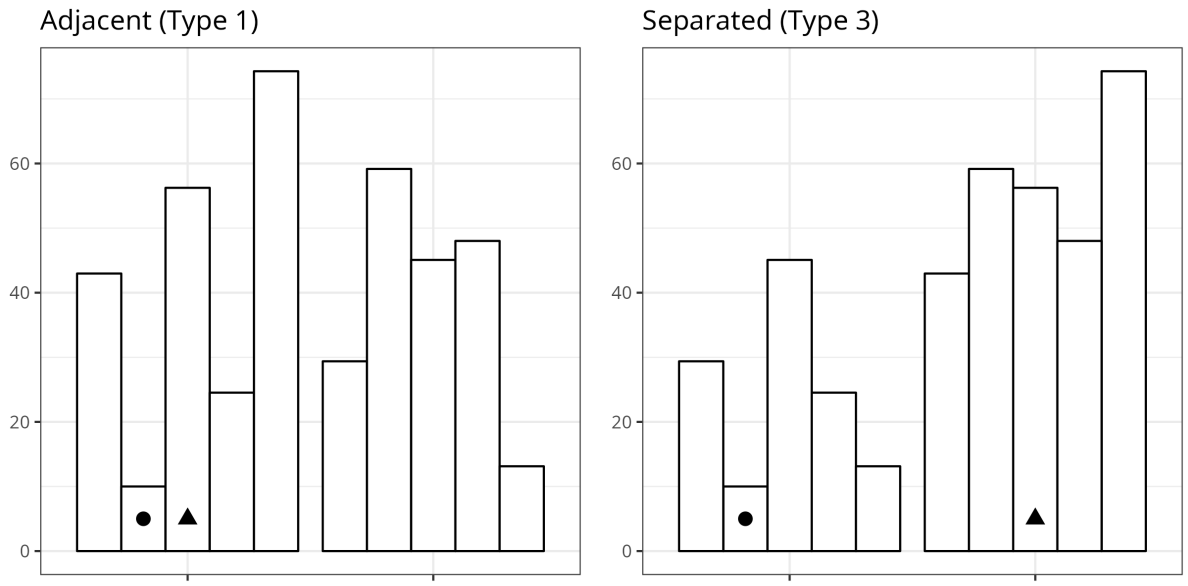
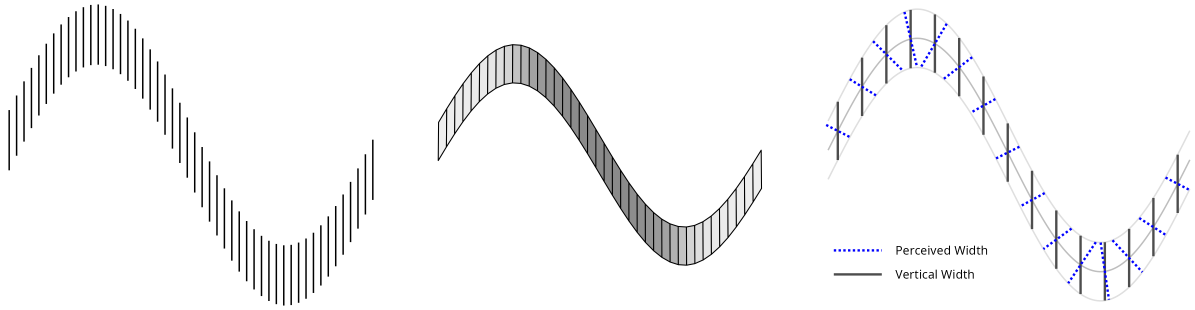


Figure 1: ? used two different types of grouped bar charts: comparisons between adjacent bars, and comparisons between separated bars. It is widely acknowledged that comparisons between separated bars (Type 3 comparisons, in Cleveland & McGill's terminology) are more difficult and error-prone.

such as position along a common scale, length, angle, and volume, and estimating the corresponding numerical value of these representations. ~~They found that estimates of ratios between positions along a common scale had smaller log errors than estimates of ratios between lengths.~~ The study relied entirely on estimation accuracy, which may not always be relevant when extracting information from graphs. For example, estimation is less relevant when ordering values by size. As a result of the Cleveland and McGill (?) study, it is possible to assemble an ordering of perceptual accuracy for the elements of length, position, and angle. ? replicated some parts of ~~?~~ ? in an online setting using Mechanical Turk, ~~largely replicating~~ a platform for crowdsourcing human tasks that has been used for (among other things) assembling machine learning datasets. This replication study largely validated the results of the original study while demonstrating the utility of the Mechanical Turk platform for graphical testing.

~~Further studies have shown that~~ The first experiment in ? (the position-length experiment), used five types of bar charts: two types of grouped bar charts and three types of stacked bar charts. Each chart had two bars marked for comparison; participants were asked to determine which bar was smaller and give the perceived ratio of the smaller bar to the larger bar. ?? shows the two types of grouped bar charts. We are primarily interested in the grouped bar charts (in part because 3D ~~graphs are less accurate at portraying numeric information that 2D graphs (??). In certain contexts and conditions, there is some research suggesting that 3D graphs may better encode information (?)~~printing is not yet inexpensive enough to make moderate-scale stacked bar chart experiments viable), which consisted of two comparison bars which were either adjacent or in separate groups. These grouped bar charts will be referenced as adjacent and separated graph types in this paper, respectively.



(a) Sine Illusion (b) 3D surface with similar features (c) Perceived vs. actual length

Figure 2: An illustration of the sine illusion(**?**), also known as the line-width illusion. All vertical lines are the same length, but the lines in the middle of the curve appear to be much shorter. The illusion results when implicit perceptual corrections useful for perceiving the size of objects with depth are applied to 2D stimuli with no actual depth. As 3D heuristics occasionally cause inaccurate communication when applied to 2D objects, it is reasonable to think that there might be some situations where charts making use of a realistic third dimension might be perceived more accurately than their 2D equivalents.

1.2 3D Graphical Perception

Chart perception is ~~often~~ affected by the human visual system's implicit assumption that visual stimuli are three-dimensional; after all, most of the visual input we process does come from a three-dimensional world, but charts are artificial~~and largely exist~~. In addition, these artificial stimuli are typically created and presented in two dimensions. ~~This occasionally causes problems: the line-width illusion,~~ The artificial scenes we create for data visualization have been documented to cause problems: for instance, the line-width illusion (**???**) has been attributed to implicit 3D perception of ~~two-dimensional~~ 2D stimuli and can affect perception of error bands, candlestick plots, hammock plots, and Sankey diagrams~~(???).~~. The line width illusion causes lines or error bands which are actually the same length to appear shorter at points with large amounts of curvature; a simple example of the classic sine illusion (one version of the line width illusion) is shown in **??**. While this illusion is problematic in two dimensions, when the depth cues are actually present in a three-dimensional situation, the same implicit visual heuristics contribute to accurate size perception (as in the middle panel of **??**). Thus, there is reason to think that perceptual accuracy may be dependent on the realism of the visual display relative to the training of the visual system and the heuristics which are activated in order to make sense of the display.

The use of 3D graphics have been explored in multiple studies, with mixed results. **?** explored ~~the preference of using either~~ user preference for 2D or 3D ~~graphs~~ charts and found that subjects tended to ~~use~~ prefer simpler 2D graphs when tasked with extracting information. **?** compared 2D and 3D graphs presented on paper and on computers~~. Their results showed~~ , showing that the accuracy of subject answers depended on their skill level~~. Novice~~ : novice subjects were more accurate with 2D paper graphs~~and~~ , while experienced managers were more accurate with 3D computer graphs. For both experience levels, participants were more confident in their answers

when using 2D graphsover other presentations of data. There are instances where 2D graphs perform better than 3D graphs, but there are times where 3D graphs may better encode information. provide a more natural way to encode information, at least in theory. For instance, when X and Y are used to represent spatial dimensions, it may be preferable to use a 3D chart to convey numerical information instead of using color, which is perceived much less accurately. **?** highlights the intrinsic attributes of 3D graphs and its the benefits when used appropriately with other 3D elements such as lighting and correct portrayal of data attributes. Nevertheless, it is extremely common in visualization literature to recommend that 3D plots be avoided at all costs (**???**).

There is thus good reason to be wary of the use of three dimensions where only two are necessary to convey data (**?**). However, the situation is different now than it was in 1984 when Cleveland & McGill published their seminal work; it has even changed since 2014. Digital graphics has Heer's replication study (**?**) in 2010.

Computer graphics and rendering technology have developed quickly, along with the hardware necessary to support these software developments. As a result, we have much more natural virtual rendering of 3D objectsvirtually, and we can also print graphics in three dimensions , moving the using mass-market, relatively inexpensive 3D printers, moving artificial charts into a more natural, physical setting. As a result, it is reasonable to reconsider the use of 3D charts, not only because of new technological developments, but because these also because 3D printed charts provide the opportunity to make visual data graphics accessible to those with limited or absent vision (**?**).

Here, we provide the process of replication and modernization of testing perceptual judgments to 2D graphs, In this paper, we discuss a study designed to examine Cleveland & McGill's experiments on grouped bar charts using modern graphics in two and three dimensions. This study lays the groundwork for additional empirical studies on the use of 3D graphs projected in 2D environments, and graphics, rendered and 3D printed bar graphs.

Cleveland and McGillprovided a theory and tested for the ordering of perceptual importance for the elements of length, position, and angle. Their first experiment, referenced as the position-length experiment, used five types of bar charts. Two of these were grouped bar charts and the other three were stacked bar charts. Each chart had two bars used for comparison and participants were asked to determine which bar was smaller and give their perceived ratio of the smaller bar to the larger bar . The two grouped bar charts are for the perceptual element of position along a common scale, where one has zero distance between bars and the other has a fixed distance between bars. These grouped bar charts will be referenced as adjacent and separated graph types in this paper , respectively.

Our study replicates the procedure for the comparisons of the two grouped bar charts, but with an objective of detecting differences in accuracy between printed, for visualizing complex data. In order to explore perception of fully 3D graphics, it is prudent to start with the simplest possible 3D graphic: one in which the third dimension is not necessary, so that we can easily compare to two-dimensional representations without loss of information. In the next section we provide details about design and execution of our experiment, including the process of replicating stimuli from **?** in order to create 2Dgraphs, , 3D digital graphs, projected, 3D rendered, and 3D printed bar graphs. The next section presents the results, and we conclude the paper by discussing this experiment in the context of existing work on the perception of 2D and 3D printed graphsgraphical elements.

Table 1: Values used in the experiment to make ratio comparisons, sorted by ratio. The ID label corresponds to the file reference number.

Bar	ID: 01	ID: 06	ID: 05	ID: 04	ID: 02	ID: 03	ID: 09
Smaller	10.00	14.70	14.70	12.10	10.00	12.10	26.10
Larger	56.20	56.20	38.30	26.10	17.80	17.80	31.60
Ratio (%)	17.80	26.20	38.40	46.40	56.20	68.00	82.60

2 Methods

Our study is designed to replicate and expand upon the position-length experiment from Cleveland ~~and~~ & McGill as closely as possible, with the hope of being able to integrate this study with previously reported results in a consistent manner. In this section, we ~~will~~ discuss the replication process and the design of our ~~modified~~ version of this experiment.

2.1 Replicating Cleveland and McGill

The first step of replicating the position-length experiment was to determine the heights of the bars that participants use for comparisons. These values for the bar heights are linear on a log scale and are given by

$$s_i = 10 \cdot 10^{(i-1)/12}, \qquad i = 1, ..., 10$$

Each graph presents two bars from the values given above where the participants are asked to judge the ratio of the smaller bar to the larger bar. The ratio of bars used by Cleveland and McGill were 17.8, 26.1, 38.3, 46.4 (twice), 56.2, 68.1 (twice), and 82.5 (twice). The exact numeric comparisons were not disclosed, but the comparison values used in our study were subjected to the constraints of having the same ratio values and that no value was used more than twice.

~~Each graphis presented so that there are ten bars where only~~

In each graph, there are two sets of five bars each, and two of the bars are marked for identification with a circle and triangle, as in **??**. Cleveland and McGill did not specify the random process for the heights of the eight other bars, so we used a scaled Beta distribution ~~with parameters that limit excessive noise around the bars used for comparisons~~to provide random structure around our fixed values. Code to reproduce the data generation process, data underlying the plots used in this study, the rendered plots and STL files, anonymized user data, and analysis code can found be at `https://github.com/TWiedRW/2023-JDS-3dcharts`.

2.2 Stimuli Construction

The graphs share a common layout across all formats, where two groupings of five bars are identified by "A" and "B", respectively, and circles and triangles are used to identify the bars participants should compare. Example graphs are shown in ~~Figure ??~~**???**. There are some graphical elements that cannot be easily portrayed via 3D printing. For this reason, all graph types do not have axes, grid lines, or floating titles.

The ggplot2 (**?**) package was utilized to create the 2D bar charts. The scale axis was removed, leaving only the bars and a bar grouping identifier. The bars used for comparisons had the

Figure 3: Two dimensional, ~~two-dimensional~~ 3D digital rendering, and 3D-printed charts used in this study.

identifying mark at a height of 5 out of 100 for the 2D plots, and the 3D plots had the identifying marks on top of the bars.

The 3D renderings and 3D printed charts were both created using OpenSCAD (?), which creates STL files from markup describing the object's geometric composition. Charts were composed of a platform, ~~raised labels for~~ with raised text labels centered in front of the A and B groups of bars~~,~~. Bars were created by inserting bar heights into the OpenSCAD template using R, and then raised circle and triangle markers ~~indicating the bars of interest, and the bars themselves; values for the bar heights were inserted into the markup using R (?)~~were positioned on top of the bars to be compared to provide a tactile and visual indicator. In addition, an ID code was engraved into the bottom of the platform to uniquely identify each object; this ~~allows~~ allowed the researchers to ensure that ~~the 3D printedcharts are correctly allocated to stimulus sets~~each kit contained the correct charts throughout the experiment while minimizing the risk of stickers or other identifiers becoming detached from the charts. As printed, the base of each chart was 13cm x 3cm x 1cm, with the highest bar rising 9.5cm above the chart base. Raised letters and shapes were 2mm above the base or bar, respectively. The color of the filament used to print each chart corresponded to the estimation ratio (though this was not obvious to participants), allowing the researchers to quickly determine that a kit of 3D charts contained no ratio duplication. Colors were randomly assigned to ratios so that participants could not gain information about the ratio from any perceived ordering of the colors.

Digital renderings of the generated STL files were created using ~~?~~RGL (?), which integrates into ~~? using the ? extension. Rendered 3D charts were initially angled corresponding to the~~ Shiny (?) using the WebGL (?) extension. The RGL rendering of the STL file was initially angled to correspond to the perspective of default 3D bar charts present in Microsoft Excel, but ~~WebGL's interactivity allows the user to rotate, scale, and otherwise interact with the chart to change the angle. 3D renderings~~ in the WebGL interactive environment, participants could rotate the charts as they pleased to create a useful estimate of the bar ratio. Rendered charts were colored to correspond ~~to the 3D printed chart filament color~~ with the filament color used to print the physical chart for consistency. The default `rgl` lighting was replaced with three lights located in fixed positions around the rendered figure. The lights were positioned so that one was behind the rendered figure, another in front of the figure, and one light below the figure. ~~3D charts were printed with colored filament corresponding to a specific ratio comparison; this allowed researchers to visually assess kits to ensure that they contained five unique ratios. Colors corresponding to each ratio were assigned randomly to ensure that chart color provided no useful information about the ratio value. As printed, the base of the chart was 13cm x 3cm x 1cm,~~

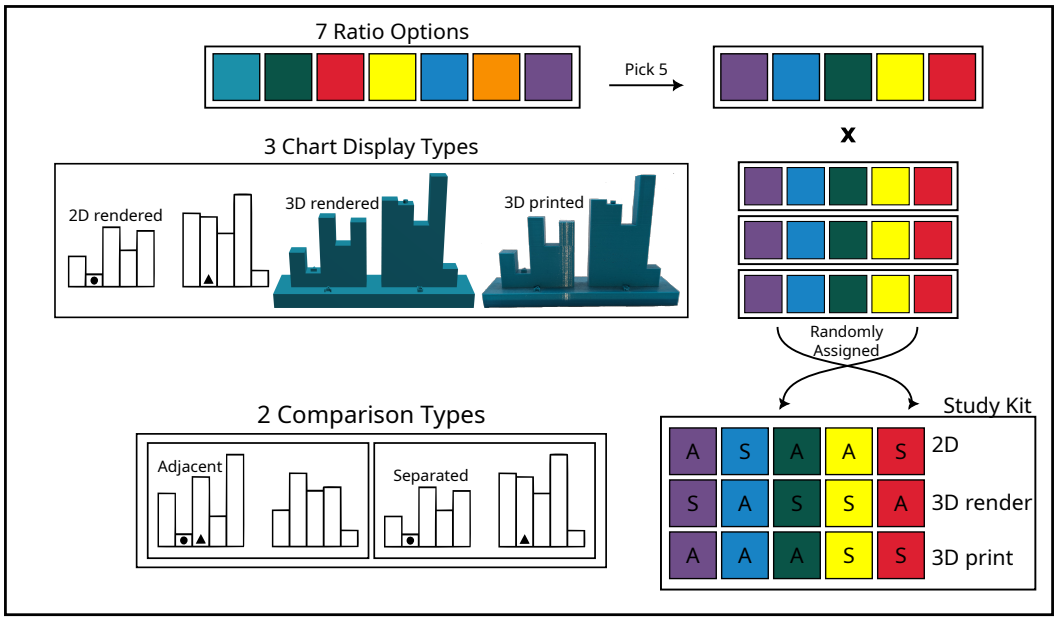


Figure 4: A graphical representation of the study design. Only five of the seven ratios were used in each kit~~where~~ ; at least one of the smallest or largest ratios ~~are randomly~~ were included along with 4 other charts; each selected ratio was displayed in all 3 mediums. For each medium × ratio combination, the comparison type (separated or adjacent) was randomly determined. A total of 21 kits of 3D printed charts were created to include all combinations of the five ratios.

~~with the highest bar rising 9.5cm above the chart base. Raised letters and shapes were 2mm above the base or bar, respectively.~~ This arrangement provided a consistent visual experience and minimized shadows that prevented visual estimation of the bars.

2.3 Experiment Design

~~Participants were provided a kit with five 3D-printed charts, comprising five of the seven unique ratio comparisons; we then used the kit ID to ensure that participants saw computer-rendered charts with ratios corresponding to those in the kit of physical charts~~ A diagram of the experimental design is provided in **??**. The study is set up as a modified randomized incomplete block design, comprised of 3 blocks (presentation mode), 7 bar ratios (5 are selected for each participant). At each combination of ratio and presentation type, the comparison type, adjacent (Type 1) or separated (Type 3), is randomly selected. This process was used to create 21 kits of 5 3D printed charts and 10 virtual charts. There is thus a slight deviation from a completely randomized incomplete block design, in that all possible combinations of ratios, presentation modes, and comparison types are not present in our design. However, this approach provided the best experimental design given the constraints of managing physical stimuli; it took nearly a month of continuous 3D printing to create all of the charts for the kits that we assembled. A dataset containing the kit composition breakdown is available in the github repository for this project. ~~This ensured that the experimental design was balanced across chart type and randomized with respect to the type of comparison (adjacent or separated). We printed 21 kits containing five 3d-printed charts each.~~

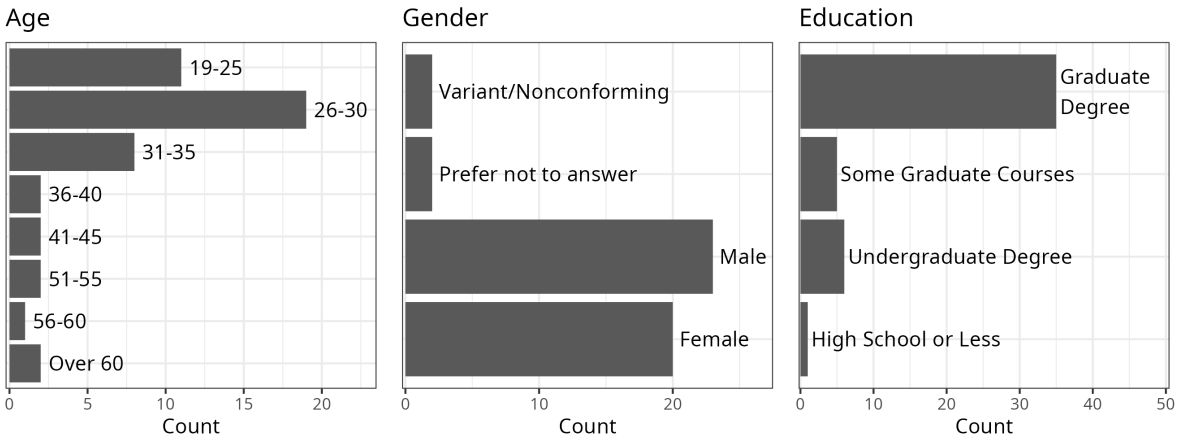


Figure 5: Demographic ~~breakdown~~ characteristics of participants in the study.~~All subjects were recruited from faculty and students in the statistics department at University of Nebraska-Lincoln.~~

2.4 Participant Recruitment

One interesting facet of ~~?~~ ? is the participant recruitment methodology: "For each experiment the subjects fell into two categories: (1) a group of females, mostly housewives, without substantial technical experience; (2) a mixture of males and females with substantial technical training and working in technical jobs. Most of the subjects in the position-length experiment participated in the position-angle experiment; in all cases repeat subjects judged the position-angle graphs first." It ~~would seem~~ seems likely that the authors recruited individuals within their respective departments as well as their wives. In ~~the spirit of replicating the study~~order to replicate this atypical participant recruitment method in the modern era, members of the UNL Statistics department and their spouses, partners, and roommates were asked to participate in our study~~; this~~ . This replicates the spirit of the ~~original study~~ recruitment method without the implicit assumptions that graduate students and professors are ~~(1) largely male, (2)~~ male, heterosexual, and ~~(3) have unemployed partners~~have unemployed spouses.

A total of 48 participants completed the study; demographics are shown in ~~??~~??.

While the results of any sample with recruitment methodology like this are not generalizable to the public, running our initial study on a comparable population to that of the initial study serves a purpose: other replications, such as ?, used online samples of paid participants, who likely have less experience using charts and graphics and making numerical estimations than those in a statistics department (and individuals who cohabitate with them). We have every intention of running this experiment in other populations, but the convenience of online sampling is not compatible with physical objects such as our 3D printed charts, so we will have to recruit in-person participants. Thus, an initial study in a population comparable to ? is a reasonable first step into testing 3D charts.

2.5 Data Collection

A Shiny applet was used for data collection, along with the provided kit of 3D printed charts. Participants provided informed consent through the applet, and then were asked for demographic

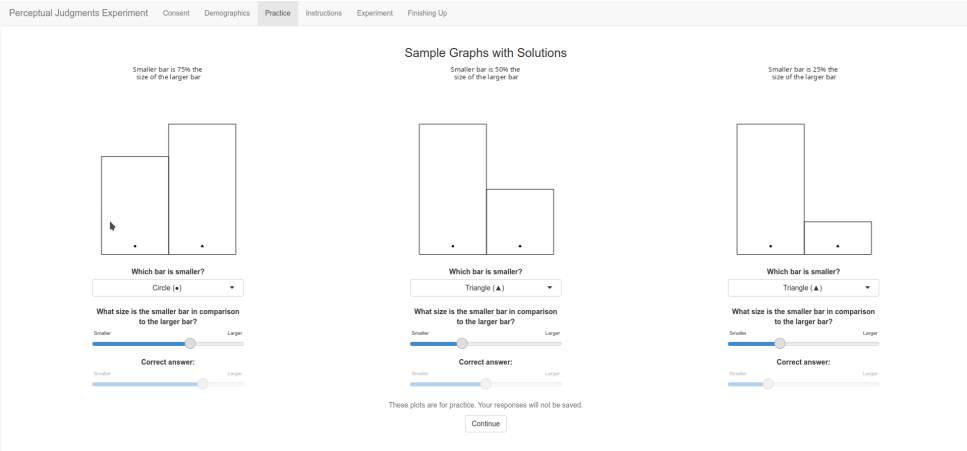


Figure 6: Screenshot of Shiny application practice screen. Three 2D bar charts with different ratios were provided, along with sliders indicating the correct proportion. Participants could practice with the sliders and preview the questions that would be asked as part of the task.

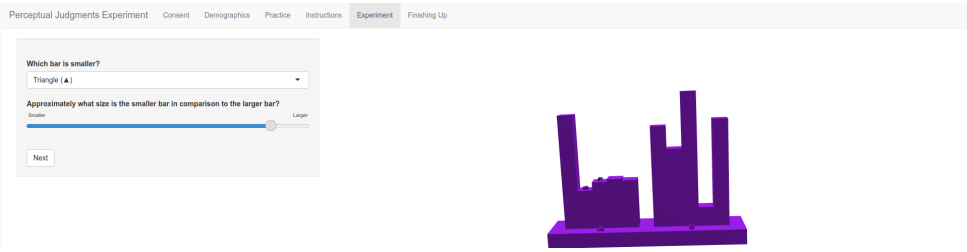


Figure 7: Screenshot of the applet collecting data for a 3D rendered chart task. Participants were asked to select which bar (circle or triangle) was smaller, and then to estimate the ratio of the smaller bar to the larger bar.

information (age, gender, education level). Then, participants were shown a "practice" page which allowed them to experiment with the data collection interface and practice estimating the ratio between the bars, as shown in ~~Figure ??~~ ??. This practice interface did not provide any feedback as to participant correctness, but was merely intended to familiarize the participants with the questions which would be asked as well as the process of estimating the ratio between the two bars.

~~Directly before the experiment started~~After the practice screen, participants were asked to provide the kit ID~~, along with directions indicating that if the instructions indicatedthat participants should use~~ . Participants were also instructed that when indicated, they should choose a 3D chart ~~for a task, the participant should select a chart~~ from the kit ~~, enter~~ and select the ID code of ~~that chart from the bottomof the object, and complete the requested~~ the chart (inscribed on the bottom) from the drop-down menu before completing the estimation task. Participants were also instructed to make quick ~~judgments~~ judgements for each graph and not to measure or estimate ratios using physical objects.

Each graph (or prompt, in the case of 3D printed charts) in the applet had two corresponding questions for participants to answer: first, participants were to identify the smaller bar by shape, and then, participants were to estimate the ratio of the size of the smaller bar to the size of

the larger bar, as shown in ~~Figure ??.~~ ??. It should be noted that we chose to use a slider for entering ratio estimates - in part, this was intended to alleviate rounding effects which occur when participants enter numbers on their own or get numeric feedback from slider entries (??). Another benefit of this modification is that it should reduce participants' cognitive load, allowing them to focus on the ratio between bars rather than the numerical translation process. This differs from ?, but maintains the spirit of the study while using methods which were not available for data entry at the time.

3 Results

All responses $(n = 33)$ that incorrectly identified the smaller bar were removed from the study before analysis~~.~~; this serves as a basic attention check. In addition, one trial was removed because the participant did not enter a valid ID for a 3D printed chart.

3.1 Midmeans of Log Absolute Errors

~~Cleveland and McGill~~ ? used

$$\log_2(|\text{Judged Percent} - \text{True Percent}| + 1/8)$$

to measure accuracy of their participant's responses. In their study, log base 2 seemed appropriate due to "average relative errors changing by factors less than 10." They also added 1/8 to prevent distortions when the errors were close to zero. ? followed the same analysis method, replicating many (but not all) of the results presented in the original paper.

~~Figure ??~~ ?? shows the midmeans of the log absolute errors compared to the true ratio of the bars for each graph type and comparison type. The results ~~tend to indicate~~ suggest that the log absolute errors increase for greater differences between the smaller and larger bars, but are consistent across the graph types.

This is somewhat different than the results in ~~?~~ ? and ?; in both cases the midmean log absolute errors increased until about 55% of the true proportional difference and then decreased. It is possible that this difference is due to the fact that we did not require an explicit numerical estimate of the ratio but instead asked participants to indicate the ratio on a slider (which is essentially a number line). ~~This should reduce the cognitive load required to transition from spatial comparison to numerical comparison and then to compute the proportion, but may also have impacted the results~~It is possible that this change explains the lack of a difference between chart modalities we see, but there are other potential explanations as well. One interesting facet of this data is that there is a consistent reduction in log error when the true proportional difference is near 50%. This may be because of implicit anchoring - we can typically bisect a span relatively accurately, which means that it might be expected that when the true proportion is near 50%, we would both notice that and be able to accurately transfer that information to the slider. This pattern is more noticable in separated (Type 3) comparisons, which also makes sense - as these comparisons are known to be less accurate, a locally minimal error near 50% with increased error relative to adjacent (Type 1) comparisons could easily explain the exaggerated trends we see in ??.

3.2 Generalized Additive Model

In addition to replicating the (primarily graphical) analysis of participant errors, we also took a more statistical approach and fitted a ~~linear mixed effects model~~ generalized additive model (?)

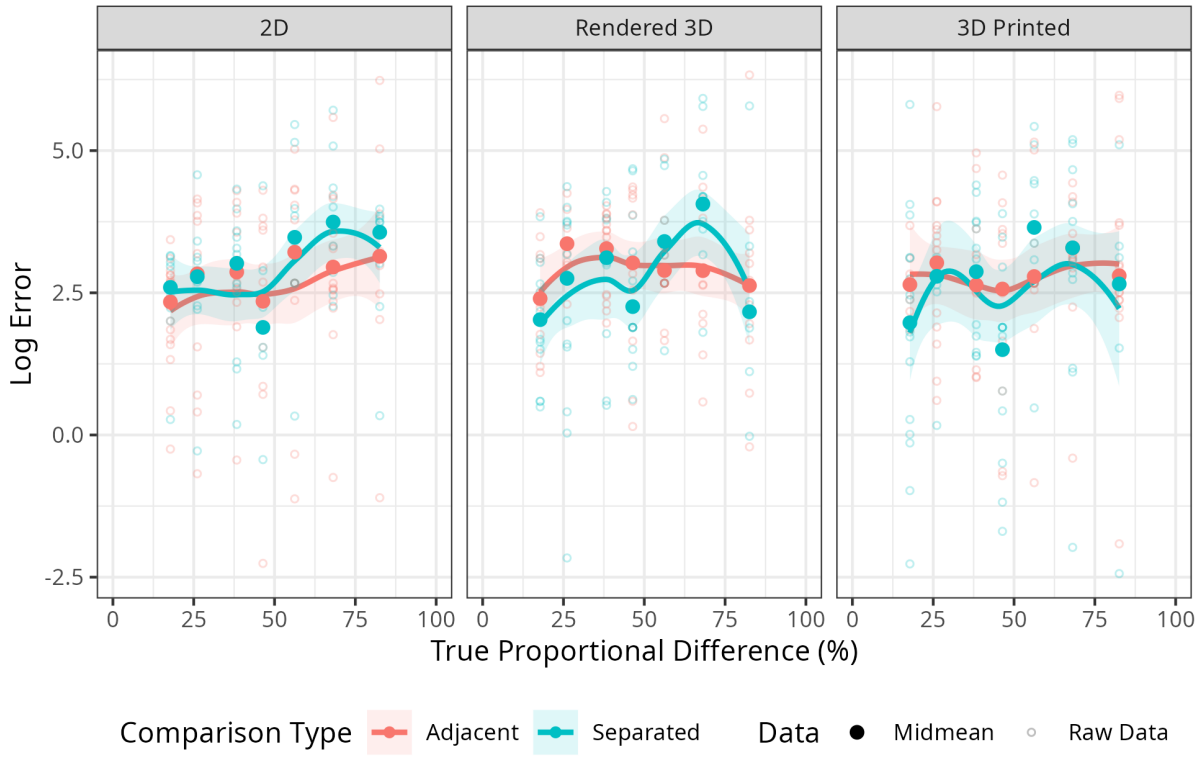


Figure 8: Midmeans and observed values of log absolute errors for the true ratio of bars. ~~Each overlaying line~~ Summary lines are computed ~~with~~ from raw data using a loess smooth.

Table 2: Parametric coefficients in gam model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.6862	0.1780	15.0931	0.0000
plot3dDigital	0.1357	0.1409	0.9626	0.3363
plot3dPrint	-0.0563	0.1415	-0.3980	0.6908
typeAdjacent	0.0968	0.1197	0.8088	0.4191

that accounts for participant variation as well as the effect of comparison type, graph type, and ratio. ~~This allows us to test for significant differences (though given the midmean log absolute error plots, we do not expect to find any) as well as to quantify effect sizes for future studies. The~~

We define the following effects:

- R, the numerical ratio of A/B where A is the length of the smaller bar and B is the length of the larger bar. There are $i = 1, ..., 7$ levels of R in this study, but we will treat R as a numeric variable throughout this model.
- D_j, the display mode, where $j = 1, 2, 3$ correspond to 2D, 3D rendered, and 3D printed charts respectively.
- T_k, the type of comparison, where $k = 1, 2$ correspond to adjacent (Type 1) and separated (Type 3) comparisons, respectively
- P_l, a random effect for participant which describes the overall skill of the participant at visual estimation of ratios.

Generalized additive models depend on a smooth function $s() : \mathbb{R} \to \mathbb{R}$ which is applied to the independent variables; by default, the smooth function is created using thin-plate regression splines. In place of an interaction term when a categorical variable is involved, separate smooths are fit for each level of the categorical variable; we will indicate this by subscripting the smooth function with the categorical variable index, so that $s_1^T(R)$ indicates the smooth over numerical variable R for the 1st level of the categorical variable T, that is, the smooth function over R for adjacent comparisons.

Then the formal statistical model is as follows:

$$y_{ijklm} = \mu + S_i + Rs^D{}_j(R) + Gs_k^T(R)_{(k)j} + D_j + T_k + P_l + \epsilon_{ijklm}$$

where

- ~~$y_{ijklm} = \log_2(|\text{Judged Percent} - \text{True Percent}| + 1/8)$~~ $y = \log_2(|\text{Judged Percent}_{ijkl} - \text{True Percent}_{ij}| + 1/8)$
- ~~$S_i \sim N(0, \sigma_S^2)$ is the effect of the i^{th} subject~~ $s_j^D(R)$ is a thin-plate spline function for R accounting for the display type
- ~~R_j is the~~ $s_k^T(R)$ is a thin-plate spline function for R accounting for the comparison type
- D_j is the fixed effect of the ~~j^{th} ratio~~ display type (e.g. an intercept term)
- ~~$G(R)_{(k)j}$ is the~~ T_k is the fixed effect of the k^{th} ~~graph type nested in the j^{th} ratio~~ comparison type
- ~~T_l is the~~ P_l is the random effect of the l^{th} ~~comparison type~~ participant
- ~~$\epsilon_{ijklm} \sim N(0, \sigma_\epsilon^2)$~~ ϵ is the random error

~~No differences were detected for the true ratio of bars (~~

Table 3: Approximate significance of smooth terms in gam model.

Smooth	edf	Ref.df	F	p-value~~= .7881)~~
s(Ratio):plot2dDigital	0.0005	0.0009	0.0000	0.5000
s(Ratio):plot3dDigital	1.7934	2.2056	1.2807	0.2481
s(Ratio):plot3dPrint	1.0003	1.0005	2.4488	0.1183
s(Ratio):typeSeparated	4.2371	4.7338	5.8074	0.0001
s(Ratio):typeAdjacent	1.0007	1.0014	3.5452	0.0604
s(participant)	27.0449	34.0000	4.7199	0.0000

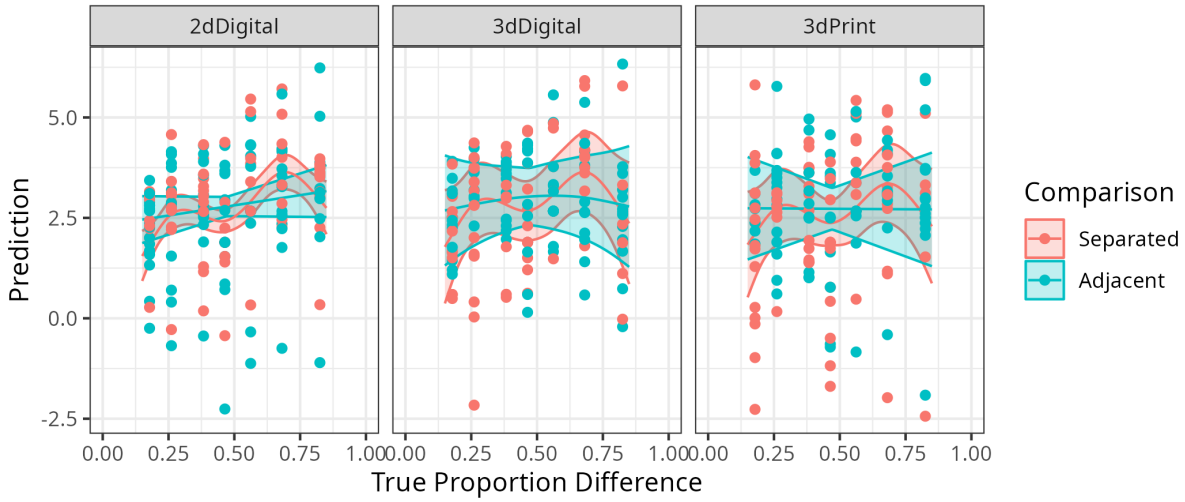


Figure 9: Predictions with standard errors for fixed effects in a generalized additive model with splines for ratio by comparison and ratio by display method. It is clear that the separated comparisons are easier to estimate when small, right around 50%, or large, and harder to estimate between these points. What is interesting is that no such trend is present for adjacent comparisons.

A generalized additive model with random effects for subject was fit using the `mgcv` package (???). Spline terms by comparison type and display type were included for the ratio variable, to account for indications of nonlinearity from ??; cumulative predictions for fixed effects are shown in ??. It is clear that there are differences between the errors for separated comparisons and adjacent comparisons; adjacent comparisons seem to have roughly similar errors for all ratio values, while separated comparisons seem to have lower errors for points near 0, ~~whether the bars were adjacent or separated (p-value = .3375)~~.5, and 1, with higher errors as the distance increases from one of these anchor points; this effect has been explored in ?. Residuals from the model indicate that there is some left skew in the tails, but this is not extreme; code for this analysis can be found in the supplemental materials on github.

While we were initially concerned that this effect was due to the slider initial setting of 0.5, we did examine this hypothesis; only 42 trials (8.81%) had slider values at exactly 0.5, and all but 6 were for ratios on either side of 0.5 (.464, ~~or for the plot nested within the true ratio~~

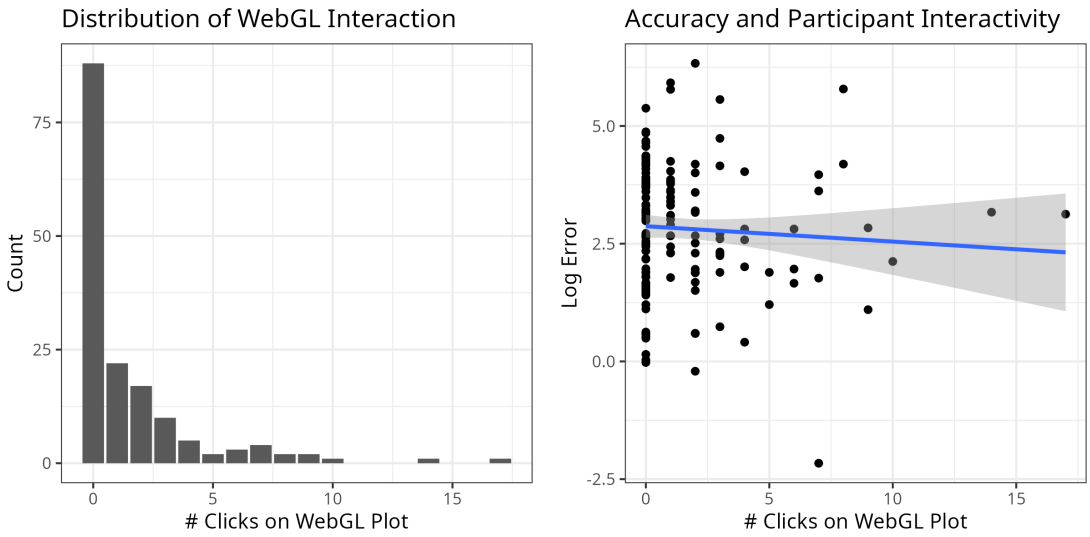


Figure 10: Participant clicks on the WebGL interface (left) in order to interact with the plot. Most participants did not interact with the WebGL interface, perhaps because they did not realize that interacting was an option, but some participants utilized the interactivity heavily. (Right) Participant interactivity was not associated with increased accuracy.

~~(p-value = .6868). While Heer & Bostock~~ .562). This suggests that while a few participants did take the strategy of using the default value for the slider, they did so strategically, not moving the slider only when the ratio was perceived to be sufficiently close to 0.5. In addition, trials where participants utilized this strategy were evenly split between separated and adjacent comparisons, suggesting that this anchoring process is not the reason for the difference in spline fits for the different comparison types. It seems more likely that this effect is at least partially due instead to lower errors around the extremes of the ratio spectrum. A follow-up experiment with more ratios and in particular ratio values closer to suspected anchor points could explore this relationship in more depth - with only 7 different ratios, it is difficult to do more than note the appearance of an effect and speculate about the anchor points, particularly given that we have used spline models with 5 knots to fit the smooth.

? provided a zip file containing data and code for their paper, ~~the link~~ but this link has not been updated and is no longer active, so it is not possible to fit a similar model to their data ~~at this time~~ for comparison purposes; if we could get access to their data, this would allow for investigating whether the differences identified in our model were an artifact of the measurement method (e.g. slider vs. numerical entry) or if this effect was present in previous studies and can now be identified because of more advanced statistical modeling techniques.

3.3 Interactivity with 3D Rendered Charts

When 3D rendered plots were shown, we recorded the number of user interactions (clicks, rotations) with the WebGL plot. **??** shows the distribution of number of clicks over all 3D rendered trials in the experiment, as well as the number of clicks relative to the accuracy measure $\log_2(|\text{Judged Percent} - \text{True Percent}| + 1/8)$. It does not appear that participants who interacted with the WebGL rendering were more accurate than their peers who did not.

4 Discussion and Future Work

Previous work in 3D graphics would suggest that the errors for the 3D graphs would be larger than the errors for the 2D graphs. While we did not find any significant results indicating that 3D graphs are read less accurately, there are two possibilities that might account for this discrepancy.

The first potential explanation is that this study is underpowered - the effect size is small, and our 48 participants were insufficient; the original study included 51 participants, which is slightly larger. However, we should note that the analysis methods used in **?** and ~~**?** do not include significancetesting beyond the graphical display of~~ **?** do not provide numerical tests of statistical significance, instead defaulting to graphical displays which include confidence intervals.

The second possibility is more interesting: we examined 3D charts using rendered 3D graphs and 3D printed charts; both of these options allow for participants to interact with the chart, rotating it, and generally perceiving it as one might perceive any other 3D, real, object. This is a far cry from the 3D perspective charts in the original study, which have a fixed angle and perspective and are thus not equivalent to our 3D charts. Future studies should include an additional fixed 3D perspective bar chart, which will at least enable us to examine whether modern 3D rendering environments allow for more accurate conclusions than fixed 3D perspectives. Future iterations of this study will include "traditional" 3D graphs created by Microsoft Excel (that is, graphs with a fixed 3D perspective rendered in 2D). This option will allow us to examine fixed perspective 3D plots compared to 3D renderings and 2D plots; it will also enable online data collection in addition to the in-person data collection used in this experiment.

Another interesting aspect of our study is that the method used to record participant estimates is different from the method used in the original study as well as Heer & Bostock's replication study. A method similar to our slider input, marking position on a line, was used in **?**, but Spence asked participants to estimate $A/(A+B)$, where we asked participants to estimate A/B; thus, our results are still not directly comparable to previous studies. We expect that the specific ratio estimated would also have an effect on observed participant errors.

The slider method for input of ratio estimates should be easier for participants, as it does not require explicit transformation to the numerical domain. What is clear is that it would be beneficial to assess the impact of the measurement method on participant errors directly, so that the results of these different studies might be explained and interpreted with regard both to the stimuli used and the measurement method employed in the experiment. Future iterations of this experiment will likely address this estimation difference; such modifications in experimental design are relatively straightforward in Shiny and will provide useful insight into the design of future experiments evaluating the perception of statistical graphics.

4.1 Supplemental Material

Stimuli, code, and data for this experiment are provided at `https://github.com/TWiedRW/2023-JDS-3dcharts`.