Accessibility for Color Vision Deficiencies: Challenges and Findings of a Large Scale Study on Paper Figures

Katrin Angerbauer
Nils Rodrigues
Rene Cutura
{firstname}.{lastname}@visus.uni-stuttgart.de
VISUS, University of Stuttgart
Germany

Seyda Öney
Nelusa Pathmanathan
{st144066,st141156}@stud.uni-stuttgart.de
University of Stuttgart
Germany

Cristina Morariu
Daniel Weiskopf
Michael Sedlmair
{firstname}.{lastname}@visus.uni-stuttgart.de
VISUS, University of Stuttgart
Germany

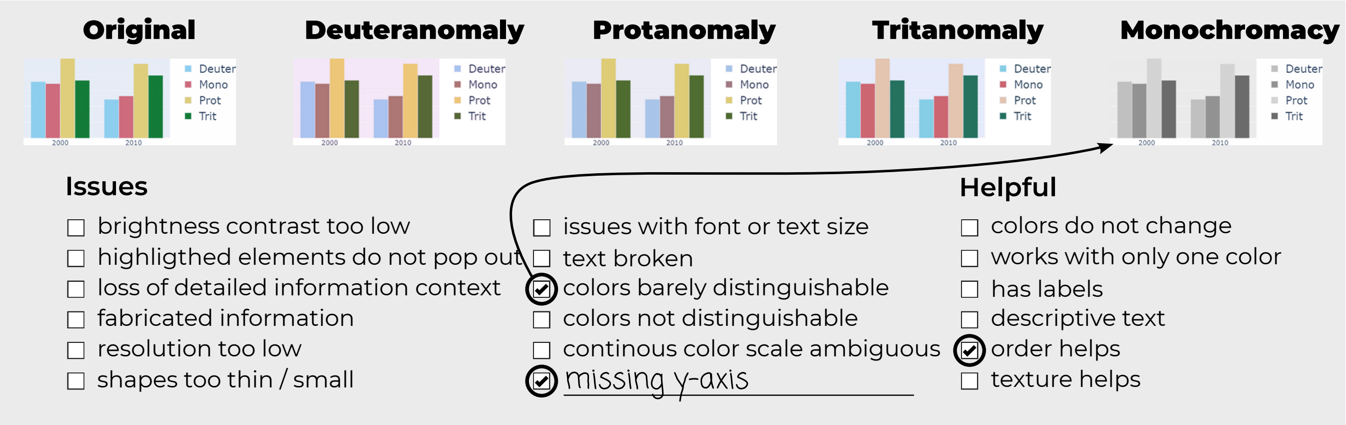


Figure 1: In our study, we looked at images from visualization research papers and simulated them in four different color vision deficiencies (CVDs). We identified issues and helpful aspects regarding accessibility.

ABSTRACT

We present an exploratory study on the accessibility of images in publications when viewed with color vision deficiencies (CVDs). The study is based on 1,710 images sampled from a visualization dataset (VIS30K) over five years. We simulated four CVDs on each image. First, four researchers (one with a CVD) identified existing issues and helpful aspects in a subset of the images. Based on the resulting labels, 200 crowdworkers provided 30,000 ratings on present CVD issues in the simulated images. We analyzed this data for correlations, clusters, trends, and free text comments to gain a first overview of paper figure accessibility. Overall, about 60 % of the images were rated accessible. Furthermore, our study indicates that accessibility issues are subjective and hard to detect. On a meta-level, we reflect on our study experience to point out challenges and opportunities of large-scale accessibility studies for future research directions.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in visualization**; **Empirical studies in accessibility**.

KEYWORDS

accessibility, color vision deficiency, visualization, crowdsourcing

ACM Reference Format:
Katrin Angerbauer, Nils Rodrigues, Rene Cutura, Seyda Öney, Nelusa Pathmanathan, Cristina Morariu, Daniel Weiskopf, and Michael Sedlmair. 2022. Accessibility for Color Vision Deficiencies: Challenges and Findings of a Large Scale Study on Paper Figures. 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, 23 pages. https://doi.org/10.1145/3491102.3502133

1 INTRODUCTION

In recent years, visualization accessibility research is gaining more and more visibility [45], e.g., by literature reviews [42], efforts toward accessible visualizations for people with intellectual and developmental disabilities [91], designing tactile visualizations [15], describing the meta-data of visualizations in multiple modalities [10], as well as making visualizations more accessible within social media, such as memes [22], GIFs [21], and images in twitter posts [23].

Still, accessible visualization design in general is challenging and under-researched [45, 52]. Cornish et al. [12] state that it is often unclear to visualization and graphic designers what accessible

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502133

design should comprise, and that one needs to foster awareness and provide more extensive tool support. Furthermore, accessibility issues in visualizations might be hard to discover for someone who is not used to looking for them. Here, fostering awareness with deficiency simulators often helps to bridge the gap between perspectives [3].

We revisit one of the more extensively researched domains [45], visualization accessibility for color vision deficiencies (CVD), which impacts the ability to see the full color spectrum. In the visualization and human-computer interaction (HCI) community, researchers have already investigated how to make colors in visualizations more accessible [48]. How well is such research turned into practice? What issues or helpful aspects can we identify in figures used in our daily research? What can we learn from those aspects to make our figures more accessible? There is work surveying the accessibility of web pages [2, 50, 58] and research on the accessibility of psychology [20] and biology papers [37]. However, to the best of our knowledge, there exist no large-scale accessibility surveys for figures in the domains of HCI or in the domain of visualization research in particular. Our goal is to help fill this gap with an empirical study for visualization research. Specifically, we assess images in data visualization research papers in terms of accessibility, especially related to CVDs. We identify existing issues, as well as aspects that help accessibility.

To that end, we conducted an exploratory image assessment study [74] on 1,710 images sampled from the VIS30K dataset [9]. We first applied techniques inspired by qualitative research methods, such as open coding [8], to prepare a set of labels regarding accessibility. In the next step, we employed 200 crowdworkers to actually perform the accessibility labeling task. While our annotators rated over half of the pictures as generally accessible, they also discovered at least one issue in the majority of all images. Images simulated in well-known deficiencies were usually more accessible than in rarer ones. Furthermore, accessibility issues seemed rather subtle and depended on individual perception. With this data exploration, we wanted to obtain an initial overview of the CVD accessibility of visualizations in practice, which also provides directions for future studies.

In short, we provide three major contributions. First, a large data study within the context of four CVDs, which to the best of our knowledge has not yet been done before at this scale. Second, we analyze the study results to provide an overview of CVD accessibility for images in visualization papers. Finally, we report on our experiences and challenges during the study to provide pointers for other researchers with similar aspirations. Based on our findings, we identify potential directions for future work.

2 BACKGROUND AND RELATED WORK

Approximately 300 million people all over the world have a CVD, often facing difficulties with respect to accessing visually presented information [11]. For people with full trichromatic color vision, all three types of cones are used for color perception. Having a CVD means that one type of cone does not function properly (anomalous trichromacy) or not at all (dichromacy). In this paper, we consider *protanomaly* (seeing less red), *deuteranomaly* (seeing less green), *tritanomaly* (seeing less blue) and the most severe form of CVD *monochromacy* (seeing no color at all) [11, 65]. In the following, we review previous work on designing and evaluating visualization accessibility with special focus on CVDs.

2.1 Designing for Visualization Accessibility

Color perception [69, 87], differentiation [17, 80], and color design choices [32, 40, 78, 93] are already well researched, but still an important topic of ongoing work. They are also key to CVD accessibility. Therefore, the Web Content Accessibility Guidelines (WCAG) especially stress the need for color contrast both for text and non-text elements [1], as well as other design recommendations like labels or textures [63]. Recent research also investigates the use of heuristics on the basis of the WCAG for image accessibility, concerning aspects like color contrast but also resolution and the presence of labels and other factors [54]. Many color palettes strive to be accessible [62, 88, 89]. Tools like ColorBrewer [28] and others [25] aim to facilitate the selection of accessible colors for visualization design and web publishing [75]. The rainbow color scale is critically evaluated [5], but still in debate for some tasks [49, 68]. It has its accessibility issues and there has been some effort to find alternative more accessible versions [57, 61].

The avoidance of accessibility problems is not limited to the design time. There are also attempts to recolor existing visualizations [38, 39, 59, 67, 70, 76] or to provide help with color identification [16, 51, 85]. The standards and guidelines above inspired us when creating the code set during the first phase of our data study.

2.2 Evaluating Accessibility

Tools that simulate CVDs [19] and other visual impairments, e.g., nearsightedness and farsightedness, glaucoma, and cataracts [24, 44, 73], can raise awareness and help designers evaluate potential accessibility issues. With these simulators, designers can load their images and qualitatively analyze how a person with visual deficiencies might perceive the content. Recent related work pointed out challenges of disability simulations [81]. However, if used with care, visual impairment simulators have proven useful in different visual design-related areas [3, 24] and in raising awareness [18, 86]. Based on these encouraging findings, we decided to use such simulations in our data study as well.

Beyond manual analysis, there also exist tools that automatically check for accessibility issues, for instance, in the context of website design [92]. However, for data visualization, there is still little tool support besides checking the contrast of two colors automatically [14]. While web design is built upon standardized reoccurring structures, the visualization design space is much larger and very diverse [46]. Although there exist automatic approaches to assess visualization design [31, 56], those have not yet been adapted to specifically check accessibility.

Another approach is to run user studies to check for accessibility. Frane et al.'s study [20], for instance, had a similar aim as our data study, i.e., to investigate the accessibility of research papers, in their case 243 illustrations and graphics in psychology journals. They, however, used preselected figures for a controlled study setup. Twenty participants without and five participants with CVD performed a low-level task and judged the visualization's accessibility. Recent efforts by Jambor et al. [37] in the field of biology used a

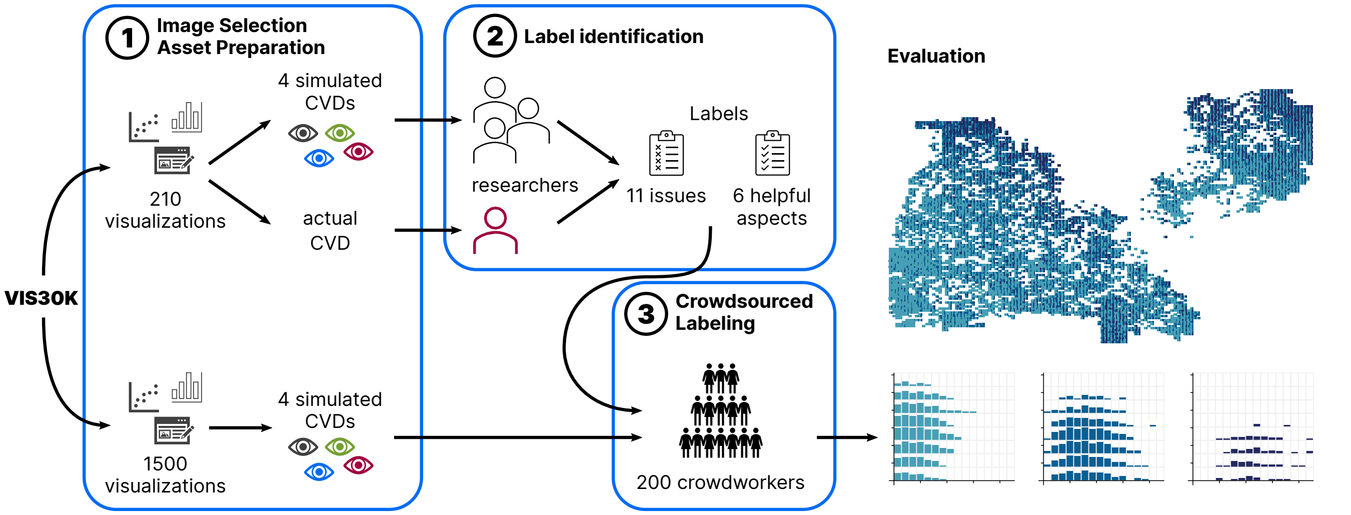


Figure 2: Pipeline of our study with the three main phases: ① Image Selection and Asset Preparation, ② Label Identification, and ③ Crowdsourced Labeling.

systematic review by two researchers to assess image legibility in 580 biology papers, looking also at CVD accessibility among other factors. In contrast to both previous approaches [20, 37], we consider a larger scale of images from the data visualization community and additionally a larger range of CVDs. Furthermore, we combine an exploratory visualization review with a crowdsourcing setup. Crowdsourcing has often been used for labeling tasks to obtain machine-learning data [7]. Over a decade ago, it was also discovered as a useful tool for user studies [43] and is now often used in the HCI community [4, 29]. Prior work used crowdsourcing for the evaluation of the accessibility of sidewalks [26, 72], the readability of Wikipedia articles [60], the creation of image captions [21, 77], and even for the accessibility of crowdwork itself [83]. However, none of these accessibility studies focused on data visualizations, which is the goal of our work.

3 METHODS

The primary goal of our study is to get an overview of, and insights into, the current and previous state of image accessibility in visualization research. We want to answer questions like: How is the perceived accessibility of the images? Which CVD issues are most prevalent? Are issues in general spotted by the majority of viewers, or are issues highly individual?

To answer these questions, we conducted an exploratory image assessment study. We followed the methodological approach of a structured analysis of a large corpus of data, in our case visualization images from scientific publications with different simulated CVDs. Inspired by coding techniques used in the social sciences [8], this method focuses on a large set of data points by a small number of coders. In the visualization community, such approaches have been used before for analyzing large sets of scatter plots [74], keywords [34], and scientific papers [71].

Methodologically, our study is split into three sequential phases (see Figure 2), which we describe in the subsequent sections:

(1) **Image Selection and Asset Preparation:** In a first step, we chose visualization images and color vision deficiencies to consider. We then generated simulated versions of the images and implemented a simple tool to assist the coding process.
(2) **Label Identification:** In this phase, four coders (three of them are co-authors) separately inspected a subset of 210 visualization images to identify issues and helpful aspects with respect to accessibility. The four label sets were iteratively refined, merged, and divided into categories of *issues* and *helpful aspects*.
(3) **Crowdsourced Labeling:** The final set of labels was then applied in a crowdsourced data labeling task on Amazon Mechanical Turk. Overall, we analyzed data from 200 workers that coded 1,500 images.

3.1 Image Selection and Asset Preparation

Image Data. We use visualization figures from IEEE VIS publications for our study. Leveraging the image database by Chen et al. [9], we extracted figures from 2000, 2005, 2010, 2015, and 2019. We started our research before VIS 2020, so 2019 was the most recent year. At the time we conducted the study, there was only a preliminary version available. The differences from the published dataset are file names and a lower resolution. We argue that this does not affect the study, because the file names are not visible to participants and, if necessary, can be mapped to their newer counterparts. We used stationary desktop computers or laptops in our open coding phase and expect most crowdsourced annotators to do the same [30, 36, 90]. While the resolution of the final dataset is 300 dpi, ours had 200. The higher resolution would be of benefit for printed papers. However, our study was performed on digital screens, of which only high-dpi displays could potentially benefit from the higher resolution images. At the time of the study, approx. 78 % [79] of desktop displays had regular resolutions far below

200 dpi. While readers can zoom in on digital papers and increase image size, the underlying resolution of raster graphics does not change. The same applies to the crowdworkers in our study: they were able to maximize the image size and use their web browser to zoom in, but that did not influence the resolution of the underlying raster graphics. However, paper figures should be readable at the same zoom level as the text. We argue that, if the image content is not easily discernible at 200 dpi, the underlying graphics have been rasterized at an insufficient resolution, or are embedded into the paper at a too low scale. Both reasons are detrimental to accessibility, irrespective of our use of 200 vs. 300 dpi in the final published data set.

For the label identification phase, we used 210 pictures from 2019, as we wanted to build up our labels on issues that are still of relevance today. In the crowdsourcing phase, we additionally included the years mentioned above to obtain a broader overview. Accessibility labeling tasks are more time consuming and complex [77] and thus require higher pay. Therefore, although our budget with ca. 10,000 $ for the crowdsourcing phase was not small, it was still not enough to cover all images in the database while considering multiple CVDs. Thus, we had to restrict ourselves to five years with overall 1,500 images. Hence, 300 pictures per year were randomly selected. Pictures labeled in the previous phase were excluded from that selection. Furthermore, we made sure that this sample only included images of visualizations, and not photos, tables, or formulas. In addition, crowdworkers' performance was tested on five images from the label identification phase. We did not include those training results in the final data for analysis.

For each considered year, we randomly selected a balanced number of pictures from each conference track available: Vis/SciVis (for readability reasons from now on only referred to as SciVis), InfoVis, and VAST. InfoVis deals primarily with abstract data,for instance, graphs, and novel interaction techniques. SciVis works with physical data, e.g., from science and engineering and VAST focuses on the entire analytical process. In the earlier years, there were fewer InfoVis and VAST images available (see Isenberg et al. [33]), which led us to the final number of 650 SciVis, 549 InfoVis, and 301 VAST images for the final crowdsourcing phase.

CVD Simulation. To generate the simulated versions of these images, we used Coblis [19], the most commonly used simulator for CVDs [69]. We chose to simulate anomalous trichomacies because they occur more often than total dichromacies [41]—at least in the case of protanomaly and deuteranomaly. Another argument for testing the milder versions was to test out the borders of accessibility: if an issue already occurs in the mild form, it is likely intensified in more severe cases. tritanomaly and monochromacy are rare, but we included them to obtain a broader overview of issues related to CVDs.

3.2 Label Identification Phase

The primary goal of the label identification phase was to generate a codeset that could be used to label a large set of images in the next phase.

Four analysts (three co-authors of this paper) independently analyzed images from VIS 2019. One coder has protanomaly and only assessed the original images. The other three are of normal vision and examined the resulting 1,050 image variants (one original and four simulated versions).

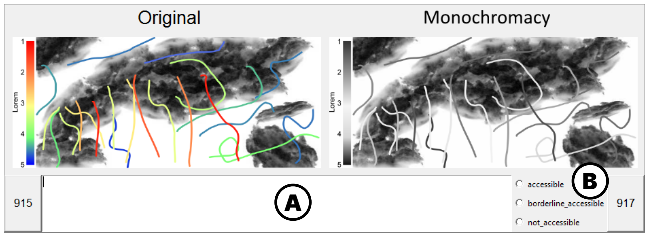


Figure 3: Custom tool for the assessment of the images. Original and CVD-simulated versions of the subject. The free text field Ⓐ provides custom keyboard shortcuts for often used snippets. Radio buttons Ⓑ or corresponding hotkeys store the accessibility rating in an external spreadsheet.

During our label identification process, the following questions guided the assessment of each image:

- What issues are already present in the original image? Is the original image accessible?
- Are there aspects that foster accessible design?
- Look at the simulated image. What issues do you see based on the design?
- In comparison to the original, is the simulated image perceived differently? Why?
- Is the simulated image accessible?

Each of the analysts distilled their own code set after a certain number of images.

We created a simple tool (see Figure 3) to accelerate the workflow. The tool also allowed rating each image either as *accessible*, *borderline accessible* (from now on referred to as *borderline*), or *not accessible*, and supported free text annotations.

We merged the resulting separate label sets into a single one by unifying different terminology that conveyed the same meaning. To further refine our labels, we discussed all images that had at least three different accessibility ratings. Finally, we grouped the labels into *issues* that have a negative impact, and *helpful aspect* that support accessibility. Within these groups, we organized the labels in terms of whether they are (i) directly related to color, or (ii) more generic issues that might impact the readability of images on top of color-related issues. Here, we were inspired by an initial survey regarding general accessibility issues of images with 13 participants (5 with CVD, 6 short-sighted). In this survey, readability issues were mentioned as well as color issues. For example, one participant with monochromacy and short sightedness expressed: *"wording can even merge into background colours and become unreadable"* or one other short-sighted participant that stated *"sometimes insufficient resolution"* as problematic.

While issues with readability, such as font size or resolution, might not relate to color vision, they could impact CVD accessibility nonetheless, e.g., if labels or descriptive text is not readable that should ensure the double encoding of information. Table 1 and Table 2 show and explain the final set of labels.

Table 1: Labels for accessibility issues in images, those in the *generic* category refer to general readability issues, while the others are caused by *color*. We use the listed abbreviations within our figures to avoid repetition and to make more efficient use of space.

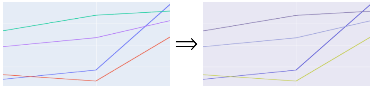
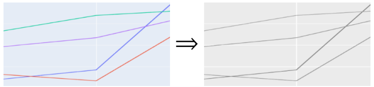
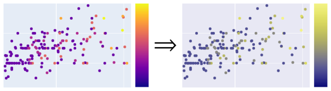
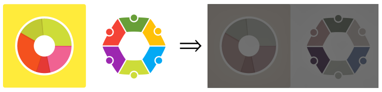
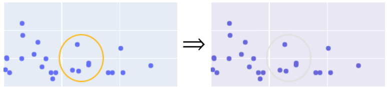
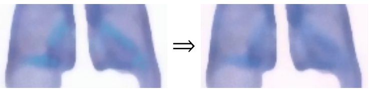
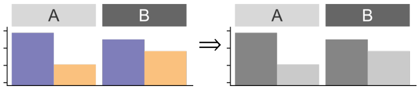
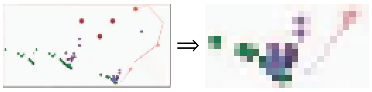
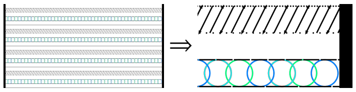
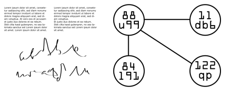
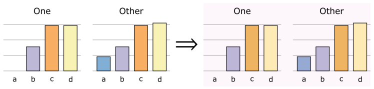
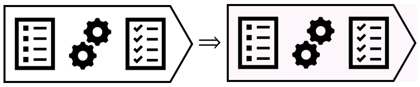
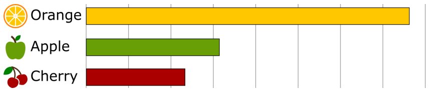
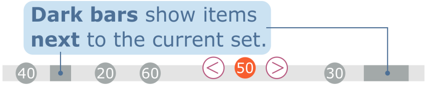
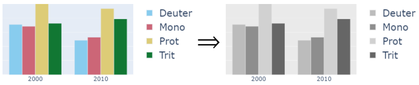
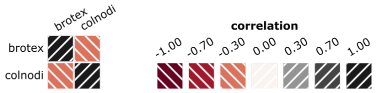
Code	Abbrev.	Explanation	Example (original ⇒ simulation/zoom)
colors barely distinguishable (*color*)	colbadi	Two or more colors that should be different but look similar. It might be difficult, but with enough effort they are still distinguishable.	
colors not distinguishable (*color*)	colnodi	Two or more colors that should be different but are indistinguishable.	
continuous color scale ambiguous (*color*)	scambi	The colors in a continuous scale are ambiguous or indistinguishable. For example, it is hard to discern individual yellow hues of the bar on the right. The original on the left has orange and red hues mixed in.	
brightness contrast too low (*color*)	locont	The brightness contrast between individual elements or between the foreground and background is too low.	
highlighted elements do not pop out (*color*)	nopop	Elements that are marked or highlighted are not immediately noticeable without focusing or requiring effort.	
loss of detailed information content (*color*)	infloss	Information shown in the original image is missing in the simulation. In the example, the turquoise halo near darker image areas disappears.	
fabricated information (*color*)	fabinf	The simulation adds new and imaginary information that was not present in the original or alters existing information, giving it a different meaning. In the example, the boxes encode categorical information (A and B) in bar charts of different measurements (purple and orange). In the simulation, it looks as if the boxes are labels for the bars because their colors are similar to the bars (also a case of *colbadi*).	
resolution too low (*generic*)	lores	Content is hard to see because the image is blurred or pixelated.	
shapes too thin / small (*generic*)	thin	Elements in the images are hard to perceive because they are too thin or too small. This might coincide with *lores*.	
issues with font or text size (*generic*)	texize	Existing text in the visualization is hard to read because of its size or font	
text broken (*generic*)	brotex	Characters within text are mispositioned or misaligned. Might lead to unintelligible gibberish.	2XEl swgeui du ancdb sor t opchanpe t ed

Table 2: Labels for helpful aspects in images, again categorized by *color* and *generic*. Abbreviations are for use within figures.

Code	Abbrev.	Explanation	Example (original ⇒ simulation)
colors do not change (*color*)	colstay	Colors barely change between the original and the simulation. Therefore, the images look almost identical.	
works with only one color (*color*)	singcol	The image is monochromatic and does not rely on multiple colors.	
has labels (*generic*)	labels	Textual labels support the understanding of the image's visual content and meaning.	
descriptive text (*generic*)	desc	Longer text elements that help to understand the image content.	 Dark bars show items next to the current set. 40 20 60 50 30
order helps (*generic*)	order	The order of the elements helps to understand the meaning. In the example on the right, all bars are sorted in the same order as the legend. This assists viewers to match the color of the bars to their labels in the legend.	
texture helps (*generic*)	texture	In addition to color, texture is used to convey the meaning (e.g., hatched, dotted, dashed lines).	

3.3 Crowdsourced Accessibility Labeling Task

In the final phase, we hired crowd workers to label and assess the full set of 1,500 images.

Annotators. In total, we recruited 200 workers (80 female, 120 male) through Amazon Mechanical Turk (MTurk). We considered different age groups, ranging from *18 to 24* to *over 65* years (mode = *18 to 24*). As we wanted to keep our workers diverse, we also welcomed annotators with vision deficiencies. 71 % had no problems with their vision, 22.5 % suffered from myopia, 8 % hyperopia, and 3 % had other issues (astigmatism, cataracts). 1.5 % participants mentioned that they had a color vision deficiency (two participants deuteranomaly, one monochromacy). These workers were instructed only to consider the original images and we evaluated their data as submissions in the deficiency condition they specified.

Furthermore, we used four Ishihara test plates [35] to check whether participants might unknowingly have a CVD.

The test scores for five participants indicate potential color weaknesses, even though they did not specify so in the questionnaire. After careful consideration, we decided to not exclude them from the study, as our goal is exploratory and not confirmatory. Under this methodological lens, the value of including their interesting perspective outweighs the potential bias through their vision anomalies.

Annotator Setup. Figure 4 shows the interface we provided for the coders. Each worker was assigned randomly to one of the four CVD conditions. The image order was randomized and we made sure that each image received annotations from at least five different coders. We also made sure that no image was coded twice by the same coder.

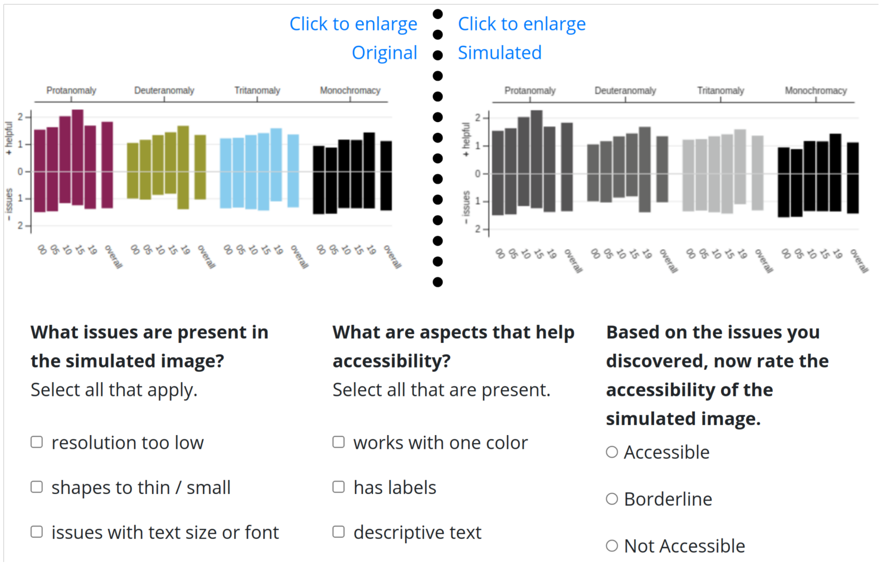
Our labeling task benefits from experience gained over time. Hence, we opted for a setup in which we could make sure that coders would label a minimum number of images. Based on our own experience from the open coding phase, we deemed 30 images to be a good threshold. Therefore, we set up the study with an initial set of randomly chosen 25 images, mostly from 2019, and five test images from the prior phase that needed to be completed as an initial training task.

Annotators were paid $ 12 and took on average of 58 minutes to complete the first set of images. Workers could take breaks whenever they liked.

Based on their performance and interest, we then offered them to continue labeling as many stimuli as they wanted. We paid $ 0.15 for each additional image, as the average trained worker took 50 seconds to complete the task (= $ 10.80 average hourly rate). The annotators should use a resolution of at least 700 x 400. If images needed to be scaled down, we showed a warning to remind workers that they should click to enlarge the pictures to show them in more detail. Additionally, we encouraged annotators to view the enlarged images when they selected codes that could potentially be influenced by screen size. For additional labels or feedback, text fields were provided.

4 RESULTS AND DISCUSSION

To gain insight into our collected data, we used descriptive statistics as well as interactive tool support to conduct our exploratory data analysis. We received over 30,000 answers to our labeling tasks,



What issues are present in the simulated image?
Select all that apply.

□ resolution too low

□ shapes to thin / small

□ issues with text size or font

What are aspects that help accessibility?
Select all that are present.

□ works with one color

□ has labels

□ descriptive text

Based on the issues you discovered, now rate the accessibility of the simulated image.

○ Accessible

○ Borderline

○ Not Accessible

Figure 4: Labeling task interface for the crowd workers. Left is the original image, while right is the CVD simulated version of it. Check-boxes are used for coding, and an open text fields provides the possibility to add free-form feedback.

containing accessibility ratings, labels, and comments by participants. The following subsections present answers to the questions we asked ourselves during the data analysis regarding the way of evaluation as well as the actual discoveries of the analysis.

4.1 How to Evaluate the Data: Challenges during Evaluation

In contrast to other labeling tasks like in Machine Learning, our setup did not target high inter-coder reliability, but rather strove for ecological validity. What one perceives as accessible highly depends on one's abilities and interpretation. Thus, it is highly individual [18]. To honor that individuality, instead of predefining a scale of what makes an image less accessible, we left that up to the user to decide. We merely showed examples to illustrate the labels used in the study.

This approach also brings its challenges, which we will discuss in the following.

4.1.1 What are subjective opinions and what is noise? According to Peer et al. [66], a worker's reputation or approval rate often ensures a certain quality, thus we followed common practice and only recruited workers with a 98 % approval rate and at least 1,000 MTurk task experience. Additionally, we tried to check the data for noise post hoc. However, aside from incomplete answers or random number answers on the Ishihara test plates not indicating a CVD, there were no clear indicators for noise. Nevertheless, we tried to do some quality checks on the data by the following pointers:

- **Analysis of the rating behavior.** We looked at how the participants rated. Here, we identified outliers, like participants that rated all their pictures as accessible or participants that only rated an image as not accessible, without specifying issues.
- **Checking how participants rated the test images.** In the test images, we identified issues as definitely not present and checked whether participants marked at least two of them regardless. If a participant did so in 3 out 5 test images, we marked them as a candidate for a sanity check.
- **Looking at participants comments.** If a participant added random comments to the free text field or comments that indicated they could have misunderstood the task (e.g., P201: *"different color on the image compared to the original"*), we double-checked their data.

If one or more pointers were true, two researchers looked at the data again in more depth to decide whether it was sane or not and then finally decided on the exclusion of the data. That way we excluded in total 386 answers. We decided to keep post hoc exclusion as minimal as possible to avoid selection bias [66].

4.1.2 How to summarize the data? Standard data analysis focuses on general trends we can uncover, often discarding outliers. In our case, outliers might hold interesting findings as well. If the majority of coders agrees on a certain rating or issue, it gets a certain weight. However, if a single participant encounters an issue, it might be also relevant in terms of accessibility, as there might be a mismatch between the needs of an individual and the majority. If the majority is even nondisabled, as was the case for our study, according to Mack et al. [53] you could even be biased by "*ableist beliefs*."

Thus, we try to report different perspectives on the data: We report means to give an overview over general tendencies, but also look at the data through the lens of the majority (more than 50 % of the coders) or minority of coders (at least one coder). In case of accessibility ratings, the majority vote was sometimes not conclusive; in these cases, we counted the image as a *borderline* image.

We report the number of images as percentages compared to either all simulated images (6,000), all images of one CVD (1,500), or images of one year (1,200), or images of one CVD in one year (300), but will state the numbers again when we are switching the base.

To additionally uncover connections or potential clusters in the data, we built our own visual analysis tool[1] using Uniform Manifold Approximation and Projection [55] (UMAP). First, we averaged over the issues and helpful aspects per image. The aggregated data was input to the UMAP algorithm. We projected the data in three different ways: (i) with the values of the average issues only, (ii) the average helpful aspects only, and (iii) with the values of both issues and helpful aspects (see Figure 16 for the projection including both groups). We used the default parameters the UMAP Python library provides. We gridified the resulting scatterplots with Hagrid [13] (default parameters). The tool allows showing thumbnails of the respective images in place, or different coloring to encode information like year, CVD, or accessibility rating (see also Figure 16 and Figure 17.)

4.2 What we found: Discoveries Regarding Visualization Accessibility

4.2.1 How Accessible were the Images in General? We received 82,830 labels in total, from which 48 % were issues and 52 % were helpful aspects.

Overall, on average 60 % of all 6,000 simulated images were rated as accessible by the majority. Borderline were 33 % of the images according to the majority and 7 % not accessible. However, only 28 % of the images stay rated *borderline* by the majority in all four CVDs.

For 97 % of the images, there exists at least one person finding that image accessible. However, 77 % of the images were also *borderline* and 50 % *not accessible* for at least one user.

The rarer the CVD, the worse the accessibility according to the majority. According to Figure 5, tritanomaly and monochromacy have the highest percentage of *not accessible* pictures. Additionally, monochromacy has the lowest number of *accessible* pictures. This result was largely expected given the fact that existing CVD visualization research focuses on the more common red-green CVDs. Interestingly, when taking a closer look at the minority ratings again, the number of pictures rated as *accessible* by the minority does almost not differ among the CVDs, which slightly differs from the majority rating.

Workers agree more on accessible images. Cases where all coders agreed on one accessibility rating were rare but more common for accessible rated pictures, who make out 13.3 % of the in total 13.8 % unanimously rated images of all 6,000 simulated ones. This agreement on accessible images is also slightly reflected when comparing majority and minority votes, the difference of accessible images is not as large compared to the differences between *borderline* and *not accessible* ratings, see the overall ratings in Figure 5.

Almost all images have issues, but also helpful aspects. Of all 6,000 simulated images, only 13 % were without issues but also only 6 % without any helpful aspect. The maximum number of issues an image had on average was five, this was however just the case for three simulated images. Regarding the helpful aspects, the maximum was four and this was the case for two images.

The minority identifies more images with issues and helpful aspects than the majority. On average over all CVDs, at least one coder found issues in 40 % of the 6,000 images, while the majority only found issues in 4 % of the images.

This pattern repeats itself also when looking at helpful aspects, although here the ratio between minority and majority is smaller (61 % of 6,000 pictures vs. 15 %).

4.2.2 What Issues/Helpful Aspects are Most Common? We now explore in how many pictures issues and helpful aspects occurred, based on the overall data and the individual CVDs. Specifically, Figure 6 and Figure 7 show the percentage of pictures an issue or helpful aspect occurred according to the majority or minority. We will now discuss the most interesting observations in more detail.

Regardless of CVD, resolution seemed to be an issue. If one looks at all simulated pictures, the most selected issue by majority (10 % of all 6,000 simulated pictures) and minority (61 %) was *low resolution.* Next up according to the majority selection are *colors not distinguishable* (9 %) and *issues with text/font size* (9 %). For the minority, the second and third most issues overall are *loss of detailed information content* (57 %) and *shapes too thin/small* (51 %).

For red-green CVDs, the most common issues are more non-color related according to the majority. For deuteranomaly, *resolution too low*, *shapes too thin/small*, and *issues with text/font size* ranked the three most common issues for the majority as well as the minority of coders although the ranking order varies (see Figure 6).

For protanomaly, the majority identifies the same issues as for deuteranomaly as the most common ones. Looking at minority ratings, however, a color-related issue *loss of detailed information content* ranks second: 62 % of the pictures are concerned with this issue according to at least one coder. For the majority, this issue only occurs in 3 % of the images.

The minority shows color issues, especially for rarer CVDs. For monochromacy, majority and minority agree on *colors not distinguishable* as the most common issue, but they disagree on the second and third most issues. For the minority, they stay related to color with *loss of detailed information content* and *colors barely distinguishable*, whereas the majority indicates only readability-related issues as problematic, as Figure 6 shows. Regarding tritanomaly, the minority identifies *continuous color scale ambiguous* and *highlighted elements do not pop out* the most occurring issues followed by *low resolution.*

Labels are the most occurring helpful aspect in general. Minority and majority report labels to be the most present helpful aspect with *labels* occurring in 47 % of 6,000 images and 84 %, respectively. In addition, among the top three helpful aspects overall are descriptive text (majority: 16 % and minority 67 %) and texture helps (majority: 11 %, minority: 75 %).

[1] https://renecutura.eu/visacctool/

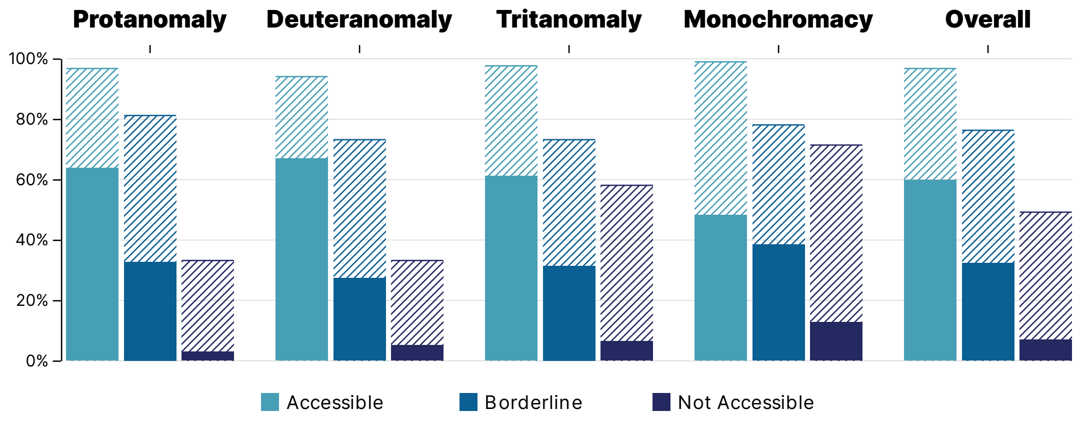


Figure 5: Percentage of pictures rated as accessible, boderline, not accessible by minority and majority. The filled bars reflect the majority, the hatched bars the minority rating.

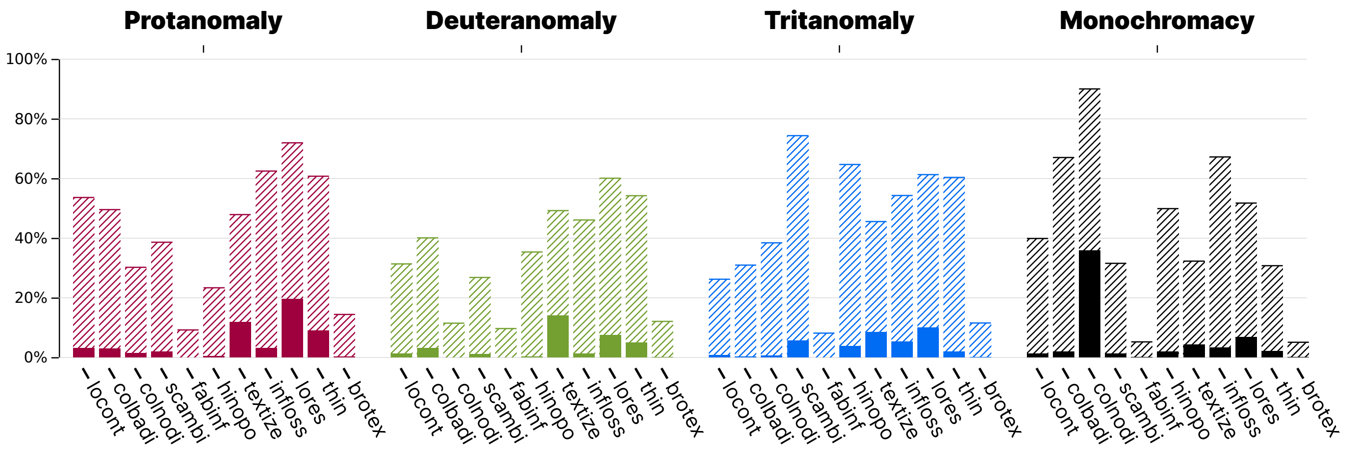


Figure 6: The bars reflect the percentage of pictures where the issue was found by at least one coder (hatched) or the majority (filled).

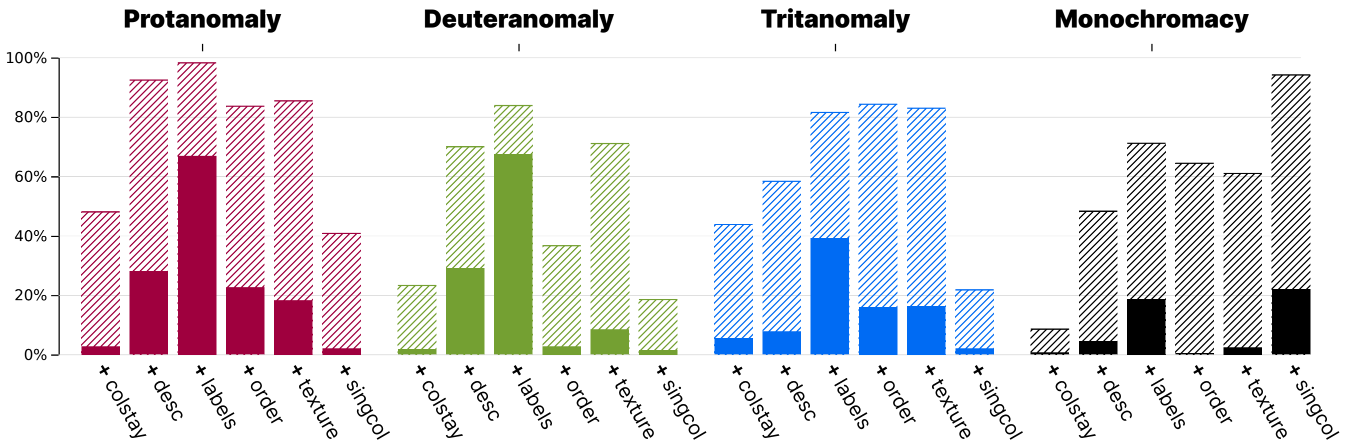


Figure 7: The bars reflect the percentage of pictures where the helpful aspect was found by at least one coder (hatched) or the majority (filled).

The most common helpful aspects do not differ across CVDs, except for monochromacy. Figure 6 denotes the most common helpful aspects for each CVD, but largely reflects the distribution mentioned above. Only monochromacy stands out, as raters naturally seemed to benefit most, if an image *works with one color*.

4.2.3 Does Accessibility Change over Time? The following section provides cautious observations of potential developments over time. Even though our dataset of sample images is not small, it might not be large enough to see robust trends. Thus, the following observations should be tested further in the future.

Overall, image ratings seem to stay rather constant over the years with a slight tendency toward improvement. Looking at the average accessibility votes (based on all votes we got per year), there seems to be a slight trend toward increasing *accessible* and decreasing *not accessible* ratings, see also Figure 8. Only 2005 seems to be an outlier to this trend, with only 55 % of the total ratings being *accessible* and 18 % *not accessible*. Furthermore, the average issue and helpful aspects counts in Figure 8.

Color issues seem to decrease over the years, except for monochromacy. Figure 9 shows how many pictures were reported to have issues. With *colors not distinguishable* and *colors barely distinguishable*, there seems to be a decrease over the years according to the majority and minority ratings. For *colors not distinguishable* it decreases from 12 % (50 % minority) to 3 % (20 % minority) of 6,000 pictures in 2000 and 2019, respectively. Pictures with *colors barely distinguishable* go from 3 % (55 %) to 1 % (18 %).

Interestingly, for both issues, 2005 seems to be an outlier with a higher number of affected pictures: 15 % (51 %) for *colors not distinguishable* and 4 % (62 %) for *colors barely distinguishable*. This might explain why the year 2005 is also an outlier in accessibility ratings. For monochromacy, there is no clear decreasing trend, especially with the minority ratings (see Figure 9 *colnodi* and *colbadi*).

4.2.4 Any Correlations Between Labels and Accessibility? While we know that correlation does not constitute a causal relationship, it could provide a good overview of the labeling quality and might provide hints for possible future studies on causal effects. To calculate the correlation, we first aggregate the data for each simulated image. Here, we map each accessibility rating to a value between from 1 (*not accessible*) to 3 (*accessible*) and average over the number of coders an image had. Similarly, we transfer issues and helpful aspects to values between 0 and 1, by averaging over the coders' observations for that particular label. For more detailed information on the data processing procedure, we refer to the supplemental material. We use the resulting numerical data to calculate the Pearson correlation coefficient summarized in Figure 10. As expected, all available helpful aspects have a positive correlation with accessibility, while all problematic issues have a negative one.

Overall, the issues that correlate most with a bad accessibility rating are *loss of detailed information content* ($\rho = -0.44$), *issues with text size or font*, and *resolution too low*. *Works with one color* and *fabricated information* had almost no correlation with accessibility. This finding seems surprising, given that intuitively these labels should have a large impact. However, *fabricated information* is hard to detect and a visualization that works with a single color only is of real benefit in case of monochromacy. Thus, when we only consider the ratings of stimuli in the monochrome condition, the coefficient ρ rises to 0.64.

Labeling has the strongest positive correlation with accessibility ($\rho = 0.24$). This sounds reasonable, as text is often black or white and will not change much when viewed with a CVD. It can disambiguate where color fails due to changes or not noticeable differences in continuous scales. Robust colors that will stay the same also correlate positively with accessibility, having the second highest correlation coefficient. *Descriptive text* and the *use of texture* follow closely. A consistent *order* of visual elements can help transfer information between them. An example of this are tables where we only write the header once and do not change the column order within the body. However, this method of labeling is only applicable in specific circumstances, leading to a lower correlation coefficient ($\rho = 0.15$).

Issues related to readability often appear together. The Pearson coefficient can also give us insight into connections between the various labels themselves, without taking accessibility into account. The matrix in Figure 10 shows a cluster of relatively high correlation within *resolution too low*, *shapes too thin/small*, and *issues with text size/font*. It seems plausible that a low resolution is insufficient to render thin primitives or text clearly.

Helpful text aspects seem to co-occur. Descriptive text and *labels* seem to often appear together ($\rho \approx 0.61$). They form a cluster of weak negative correlations with color perception issues, *highlighted elements not popping out*, and *loss of detailed information content*. Does this mean that authors who add text and labels to their figures might also be more likely to take the effects of color vision deficiency into account? Could it be that their figures need fewer colors? Or did the annotators potentially not notice some of the issues because they were neutralized? This last cluster can lead to interesting hypotheses, but we would need a more focused investigation to arrive at a valid conclusion.

4.2.5 Can Helpful Aspects Neutralize Issues? Analysis of the correlation coefficient gave rise to the supposition that helpful aspects might be able to neutralize issues. This seems to be plausible according to the principle of least effort: visualization designers would not add helpful aspects if they could not improve the resulting image. However, the correlation matrix in Figure 10 indicates that the issues weigh more heavily toward inaccessible images. To get a better overview of the role that helpful aspects and issues play, we transform the coding results into numerical data. The approach is similar to the one from Section 4.2.4, but aggregates the data even further. We refer to the supplemental material for more detailed information on the transformation procedure. The result is only a simple 3-tuple for each original and simulated image:

$$(\langle\text{% of issues}\rangle, \langle\text{% of helpful aspects}\rangle,\\ \langle\text{majority vote on accessibility}\rangle)$$

The two-dimensional histograms in Figure 11 show the resulting 6,000 data points. To avoid issues with clutter and overplotting, we created separate plots for each accessibility rating. Looking at the visualization of accessible images, we can observe an apparent diagonal cut-off line (red): there are no accessible images beneath it. On the plot for not accessible images, there is a similar line.

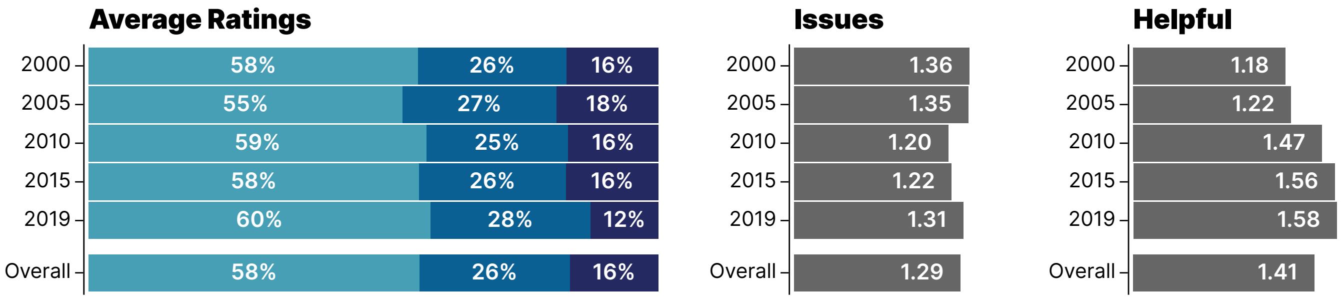


Figure 8: Average ratings over the years, based on the total amount of A B N ratings we received per year, as well as the average amount of issues and helpful aspects over time. *Overall* shows the average over all years.

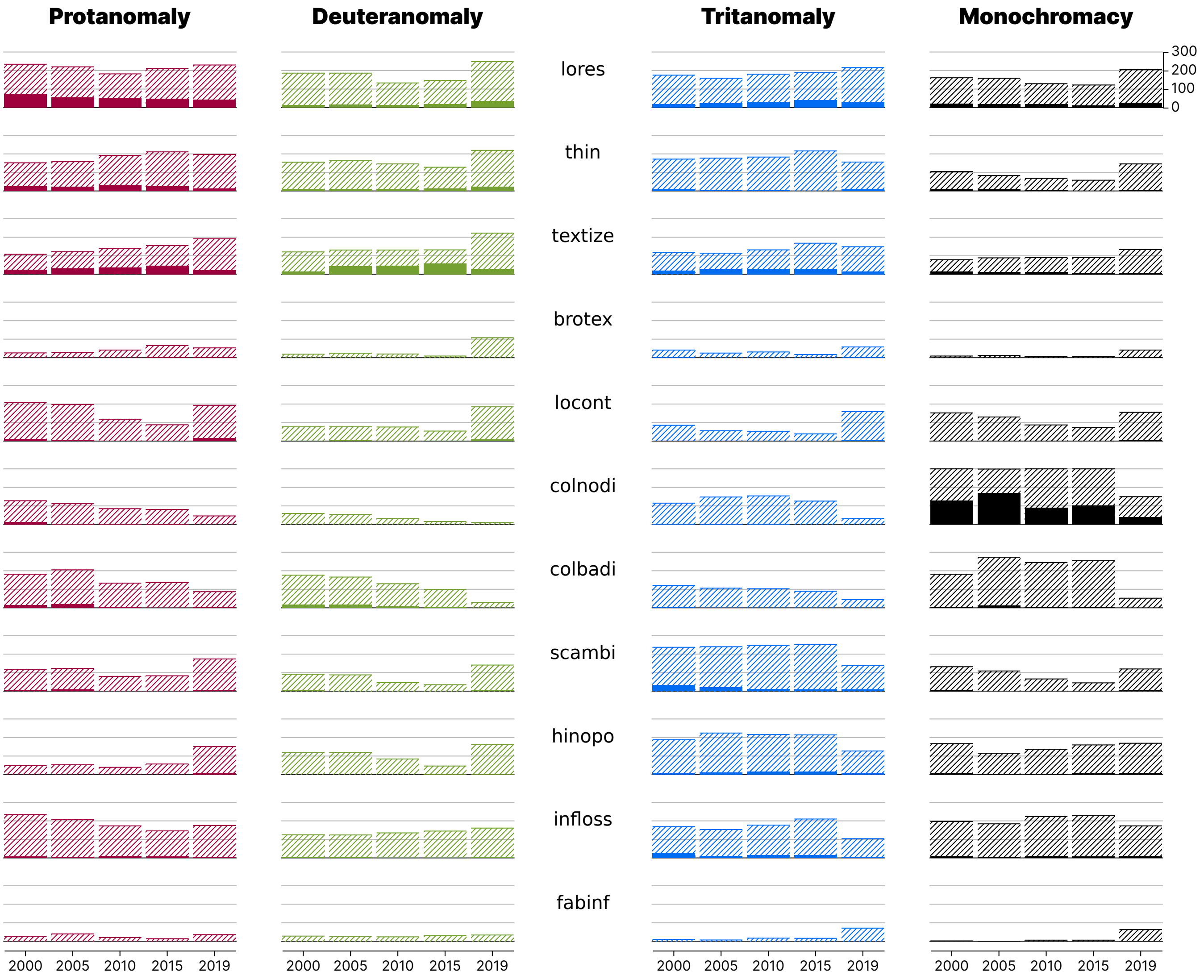


Figure 9: Amount of pictures affected by a certain issue over the years and per CVD P D T M according to minority (hatched) and majority (filled).

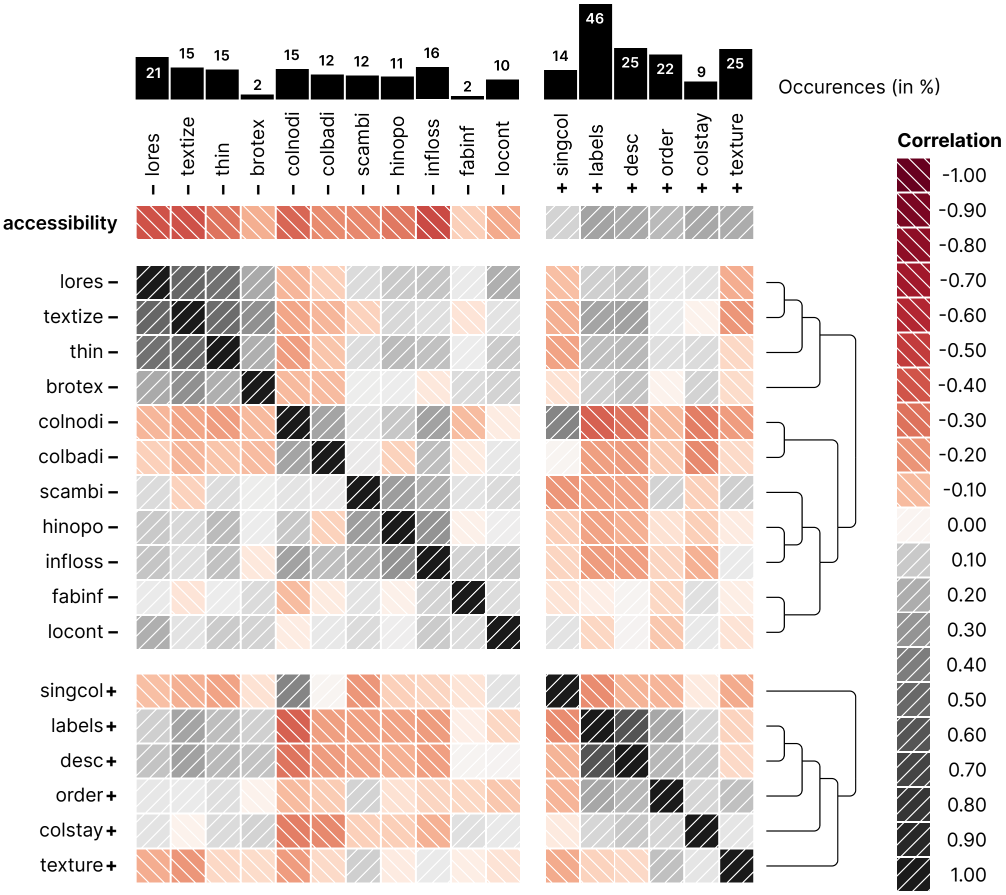


Figure 10: Correlation matrix of codes and accessibility. All issues have a negative correlation with accessibility, whereas helpful aspects have a positive one. The strongest positive correlations are between lores, textize, and thin. The strongest negative correlations are between colbadi, colnodi, scambi, hinopo, and infloss on one side and labels, desc, and colstay on the other.

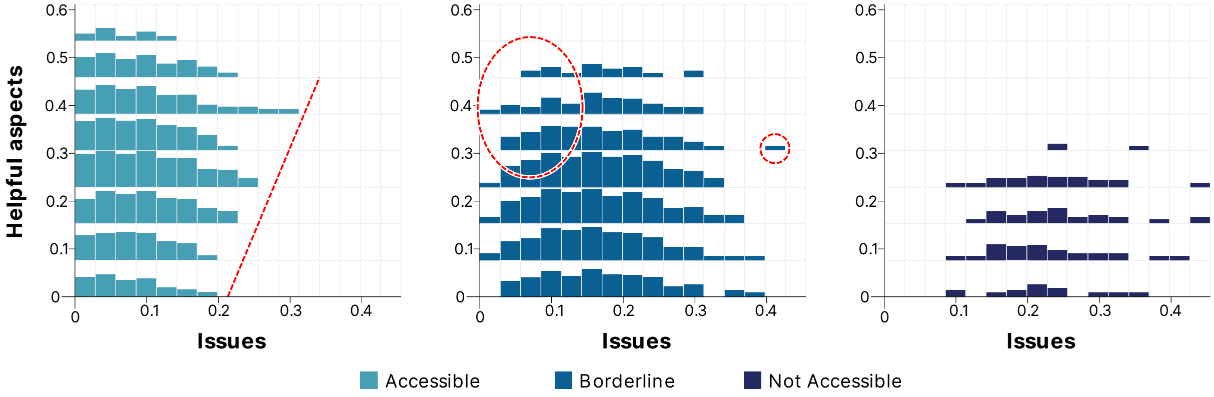


Figure 11: Two-dimensional histograms of image count. The horizontal axis shows the percentage of detected issues. The vertical axis corresponds to helpful aspects. Bar heights scale logarithmically to improve visibility in less populated bins.

This shows that with rising levels of issues, the images require an increase in helpful aspects to remain accessible.

The plot for *borderline* ratings contains images with a low amount of issues and a high amount of helpful aspects (red ellipse). It seems the issues were quite severe and did not allow for accessible images, despite the efforts of the paper authors. On the other hand, there is an outlier close to the maximum number of issues (red circle). It has more helpful aspects than some not accessible images with a similar issue rating. Again, this could indicate a neutralizing effect that might have made the difference between an image being mostly rated *not accessible* and *borderline*.

These aggregated views do not distinguish between the types of issues. Nonetheless, we were able to confirm that helpful aspects could have the potential to mitigate the negative effects of issues to some degree.

4.2.6 What Images are Most Accessible? Using the same data as in Section 4.2.5, we took a closer look at individual images and ranked them. The lower the accessibility, the lower the score. Issues have a negative impact and helpful aspects a positive one (see Section 4.2.4). Please check the supplemental material for specifics of the ranking procedure. At this point, we want to stress that it is not our goal to point out, accuse, or judge any author. We were only interested in the particularities of images with certain ratings.

Less content is ranked more accessible. The top ranks are occupied by images rated as accessible over all CVDs. Note that our sample of the top five images in Figure 12 seems to favor visualizations with a low density of information content. The first two images are schematic diagrams or drawings and the following ones are abstract two-dimensional charts with an increasing number of visual primitives. We suspect that there might be two underlying reasons that would require a more targeted investigation. The less content there is in an image, the less a viewer has to interpret. The higher the amount of monochrome text, the lower the impact of a color vision deficiency.

4.2.7 What Images are Least Accessible?

Low resolution issues revisited. As mentioned before, we do not want to finger point, thus we will not reprint the worst ranked images but provide abstract examples in Figure 13 for better understanding of the issues present. These images consist of a node-link diagram, a spatial representation using a rainbow scale with a complete graphical user interface and window chrome, very small thumbnails of maps and legends, and a map with overlaid line charts. They are severely limited by their *low resolution* (73 % of coders). *Issues with text size or font* and shapes being too thin are the next most reported codes, which could also be related to resolution (see correlation in Figure 10).

4.2.8 How Accessible are Certain Image Types? In Section 4.2.6 and 4.2.7, we looked at particularities of selected images. This analysis inspired us to take a more systematic look at the type of images and their accessibility. At the moment, the VIS image data set does not provide any meta-data on image/chart types. We thus opted to code image types ourselves, for the details of the coding procedure we refer to the supplemental material. The selection of a detailed and complete set of image types is nontrivial—especially for figures with mixed content—and would provide material for an independent publication. Therefore, we only considered seven coarse categories:

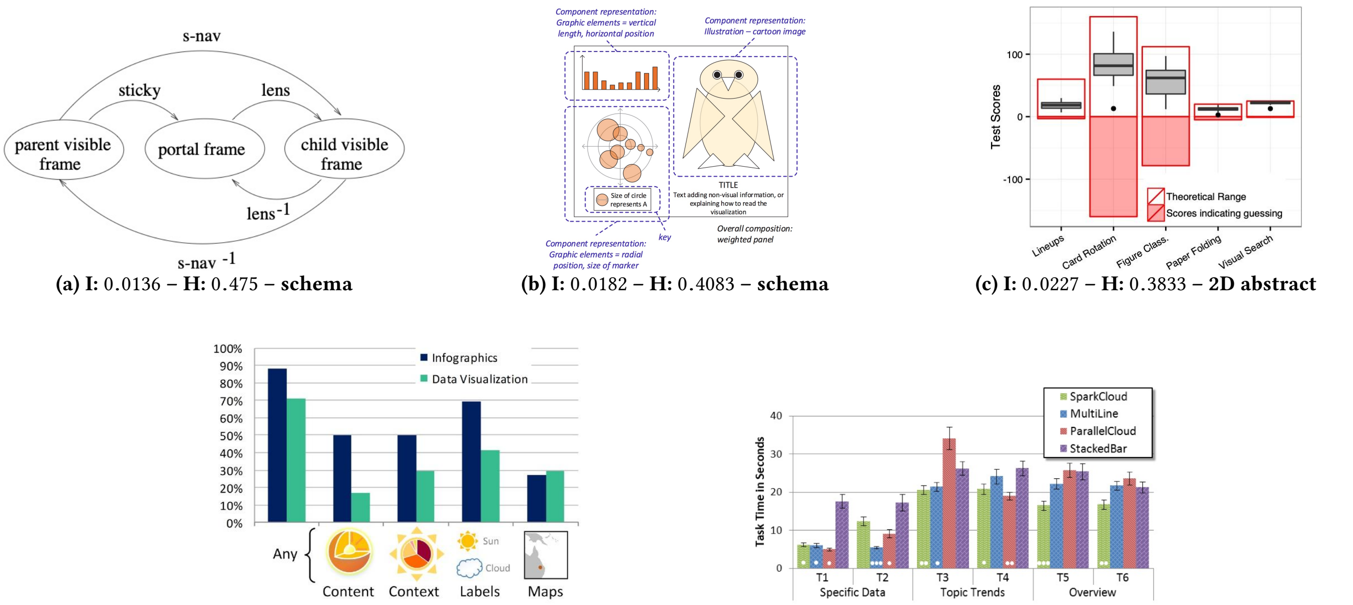
- *2D abstract*: An image belongs to this category if it is, e.g., a bar chart, box plot, or another representation of abstract data (38 % of the images).
- *2D continuous*: Heatmaps or other 2D visualizations with continuous data representations (8 % of the images).
- *Schema*: Figures explaining, e.g., processing steps with text and arrows as example key elements (10 % of the images).
- *3D abstract*: 3D bar charts or other representatives of abstract data in 3D (6 % of the images).
- *3D continuous*: Volume visualization or other forms of 3D visualization with continuous representations (19 % of the images)
- *GUI*: Screenshots of user interfaces (7 % of the images).
- *Undetermined*: Used when we could not agree on the figure type, or if the image did not fit the above categories, or if it matched multiple types (10 % of the images).

The resulting distribution of accessibility ratings across image types is shown in Figure 14.

Simple visualizations like schemas seem more accessible. *Schema*s have the highest value for majority votes of *accessible*. This is in line with the findings from Section 4.2.6. However, figures with *3D abstract* content have the second highest accessibility score, yet did not appear in the top ranked images in Figure 12. The majority ratings for the categories *3D continuous*, *2D continuous*, and *undetermined* are on a high level and similar to each other. Images of *GUI*s are most problematic, as they have approximately the same number of *accessible* and *borderline* ratings and the highest score of *not accessible*. The minority votes do not show large differences between the content types when observing *accessible* and *borderline* images, but *schema*s have the lowest rating of *not accessible*.

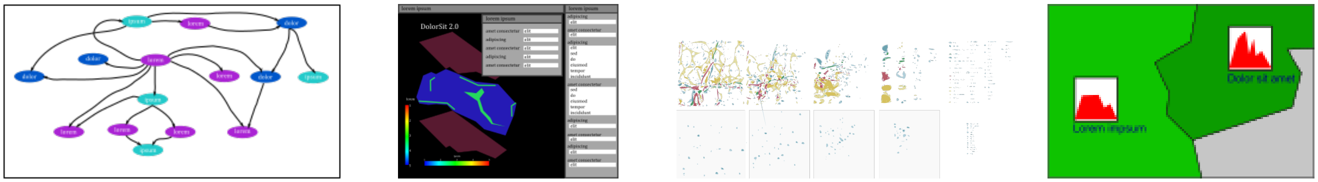
A richer design space might also hold chances for accessibility. We also analyzed the distributions of ranks from Section 4.2.6 by image category. They mostly correspond to the known results. However, contrary to the information in Figure 14, *2D abstract* and *3D continuous* have different distributions in our highscore. While *2D abstract* appears to be a bimodal distribution favoring both high and low ranks, *3D continuous* resembles a normal distribution, see also Figure 15. Could the larger design space with more dimensions and continuous mapping also provide more possibilities for accessible design? These richer configuration possibilities could be an explanation for the normal distribution of *3D continuous*, whereas abstract 2-dimensional visualizations seem to either work well or break down with less middle ground according to the ranking distribution. However, to confirm this phenomenon, more in-depth research needs to be done.

GUIs seem to have more readability issues. We further analyzed whether there were specific issues or helpful aspects of certain visualization types. Here, schematic figures make most use of labels and descriptive text, but textures appear most often in *3D continuous* visualizations. Figures containing *GUI*s seem to suffer the most from issues with low resolution, insufficient font size, and too thin or small visual elements.



(a) **I:** 0.0136 – **H:** 0.475 – **schema**
(b) **I:** 0.0182 – **H:** 0.4083 – **schema**
(c) **I:** 0.0227 – **H:** 0.3833 – **2D abstract**
(d) **I:** 0.0227 – **H:** 0.3333 – **2D abstract**
(e) **I:** 0.0273 – **H:** 0.35 – **2D abstract**

Figure 12: The five best ranked images over all conditions. All have the highest value of accessibility (3) and show either schematic content or abstract 2-dimensional graphs. We used the probability of issues (I) and helpful features (H) being detected to order them from (a) to (e). Reprinted with permission from IEEE: (a) from [64] (Fig. 11), (b) from [6] (Fig. 1), (c) from [84] (Fig. 5), (d) from [6] (Fig. 3), and (e) from [47] (Fig. 6).



(a) **I:** 0.2682 – **H:** 0.1
2D abstract
(b) **I:** 0.2909 – **H:** 0.1583
GUI
(c) **I:** 0.2591 – **H:** 0.1167
2D abstract
(d) **I:** 0.2773 – **H:** 0.125
2D abstract

Figure 13: Abstract representations of the worst ranked images over all conditions. Their mean accessibility score is approximately 1 (*not accessible*). We also used the probability of issues (I) and helpful features (H) being detected to order them (a–d), starting with the worst.

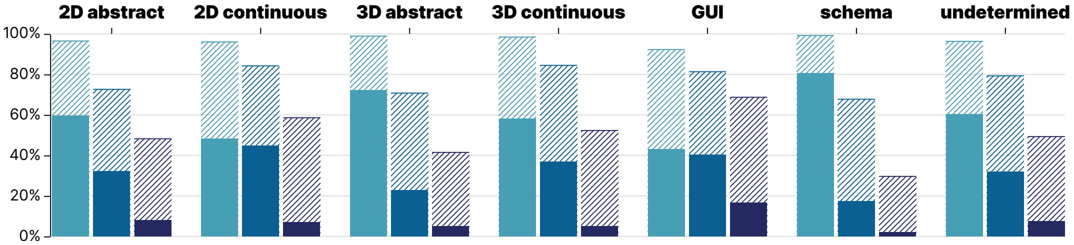


Figure 14: Percentage of pictures of a certain image type rated as accessible, boderline, not accessible by minority and majority. The filled bars reflect the majority, the hatched bars the minority rating.

This might be related to their potential purpose: screenshots might be intended to give a broad overview of software applications, not for reading and interpreting the details of the contained visualizations.

In the interest of keeping the length of this publication within reasonable limits, we do not include all analyses of the issues and

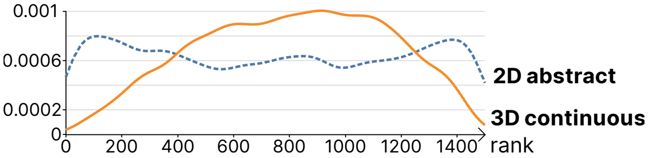


Figure 15: Probability distribution of the image rankings of *3D continuous* (solid line) and *2D abstract* (dashed line).

accessibility of certain image types. Instead, we refer the interested reader to the supplemental material for more details.

4.2.9 Are There Clusters in the Data? To analyze our data for clusters, we created our own tool as described in Section 4.1. Different colors gave us insights into various aspects of the data, see Figure 16 and Figure 17.

There are clusters regarding accessibility as well as CVD.. In region Ⓐ, there seem to be simulated images of all CVDs that are less accessible. Closer inspections lead us to identify screenshots of tools, for example. This is in line with the observations in Section 4.2.8. To see thumbnails of the images, we refer to our online tool.

Cluster Ⓑ is where monochromacy pictures mostly are located, whereas the outer rim shows less accessible pictures compared to Ⓒ. Here, there seem to be clustered pictures that work well with one color.

Some pictures with continuous color scales seem to be clustered in region Ⓓ. Interestingly, on the left of this cluster there are color scales that work for deuteranomaly, but on the right, there are pictures that are more problematic for tritanomaly, which could potentially reflect that tritanomaly has more issues with continuous color scales.

4.2.10 What do participants comment on? Participants could give feedback after the training phase (187 comments), specify problematic color choices (1,218 comments), and provide input for task feedback and other issues/helpful aspects (2,282 comments). In the paper, we will now focus on the task feedback and other issues / helpful aspects and refer to the supplemental material for an evaluation of all comment categories. The submissions in our free-text fields for feedback and other issues concerned 2,261 of 6,000 simulated images.

Three authors performed a content analysis inspired by affinity diagramming [27]. As such, we split the comments into 4,142 statements and assigned them topics, subtopics, and categories. We found seven main topics, which are visualized in Figure 18 with their subtopics, categories and absolute occurrences.

Participants identified the eight additional issues and helpful aspects, besides mentioning already existing issues/helpful aspects. The additional issues/helpful aspects were sometimes specializations or generalizations of our labels like *problems with text color* or *works with less colors*. For the helpful aspects, sometimes the opposite of issue was mentioned, like *colors easily distinguishable*. Additional issues like *relies and colors* or the helpful aspect *simple design* add a more general perspective to our labels that was not there before. The subtopic *bad color choice* generalizes our color labels mentioning, e.g., *"[t]oo bright [colors]", (P38)* or colors that are not aesthetically pleasing: *"Pink shade is awful on the eyes [...]", (P124)*.

In general issues, helpful aspects, and improvement suggestions centered around readability, understandability and color. Improvement suggestions covered a wide range, from *change of color choices* (*"Other contrasting colors should be used in the chart", (P49))* to the call for *interactivity* (*"the ability to zoom in would probably help", (P195)*).

In addition to just pointing out issues and helpful aspects, their effect, or the effect of the simulator was also mentioned. Effects of issues and helpful aspects described again the impact on *readability* and *understandability*, but two annotators even mention *eye strain*. For the simulator effect *color change* comprises neutral comments on color transformations such as *"Pink becomes blue in the simulated image", (P54)*.

Workers also explained their work (T7) and gave feedback on the task itself (T2). Our results are similar to Simons et al. [77], who found that workers often tend to provide explanations behind their thought process or express insecurities and provide feedback on the task itself. Specifically for T7, comments aimed to justify or further specify the accessibility rating, adding a more fine-grained accessibility scale. This includes the justification *requires effort*, where workers judged the accessibility additionally through the effort that was needed to perceive the information: *"Slightly blurred but with some effort accessible (P20)"*. Further, some participants also reported to be *insecure/indecisive* like *"this one is tricky. [I]t's teetering towards not accessible, but I'll stick with borderline", (P6)*. General remarks like *no issues/nothing helpful* were also to justify. An interesting justification was *issues are there by design* where participants made assumptions on the purpose of the visualization: *"The contents of the graphs doesn't seem important here so I don't think that text matters", (P19)*.

5 LESSONS LEARNED, LIMITATIONS, AND FUTURE WORK

In the following, we reflect on our findings, but also share some experiences regarding our study itself, which might be of interest for the community.

5.1 Action Items for Creating Accessible Images (Retold)

Our findings regarding visualization confirm and refine those of Frane et al. [20] and Jambor et al. [37] for the HCI community. In this section, we present aspects to consider while designing CVD-friendly figures. Rather than re-inventing the wheel, they rely on basic design principles which are put into context with our results.

Don't rely on accessible color choices only. In accessibility guidelines for papers for conferences like CHI, advice on creating CVD-friendly figures focuses on color alone[2]. In our image assessment, however, we noticed that issues are often connected to *understandability* and *readability* in addition to color. We hope that the labels

[2] https://sigchi.org/conferences/author-resources/accessibility-guide/

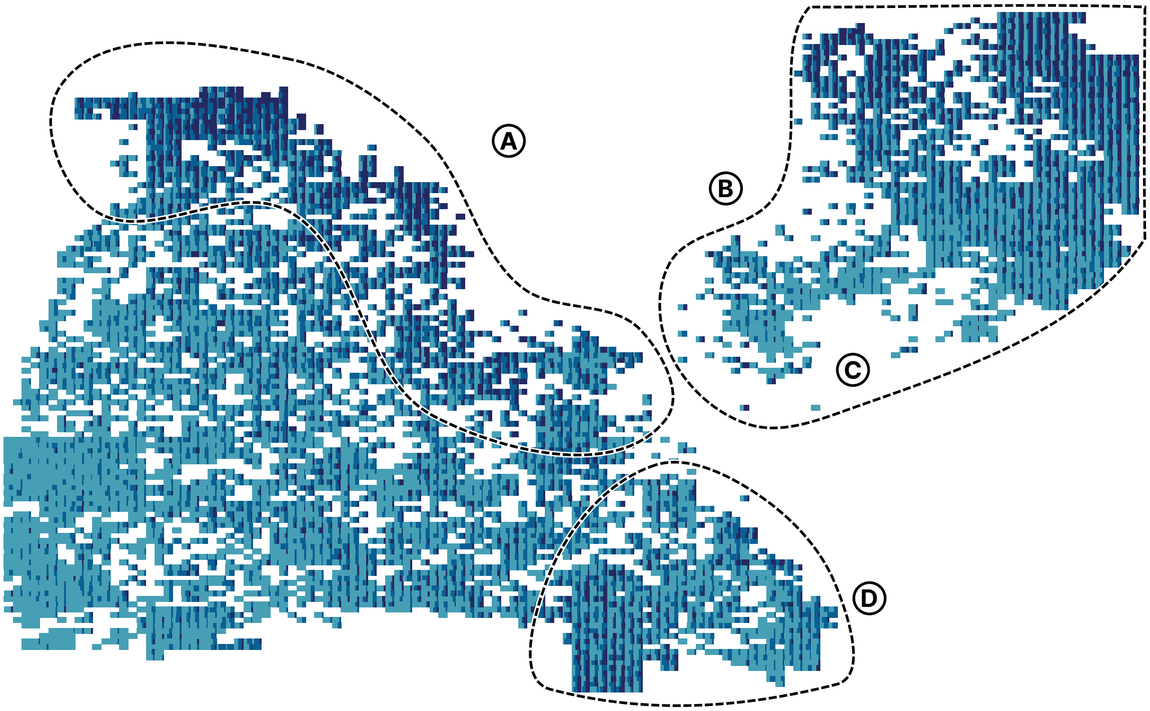


Figure 16: Gridified 2D embedding of the coding results with UMAP. Each cell represents a stimulus as viewed with a specific color vision deficiency. It contains a horizontal stacked bar chart that shows the ratio of the stimulus' accessibility ratings. Accessibility is mapped to color: A B N. Ⓐ to Ⓓ represent interesting image clusters we found.

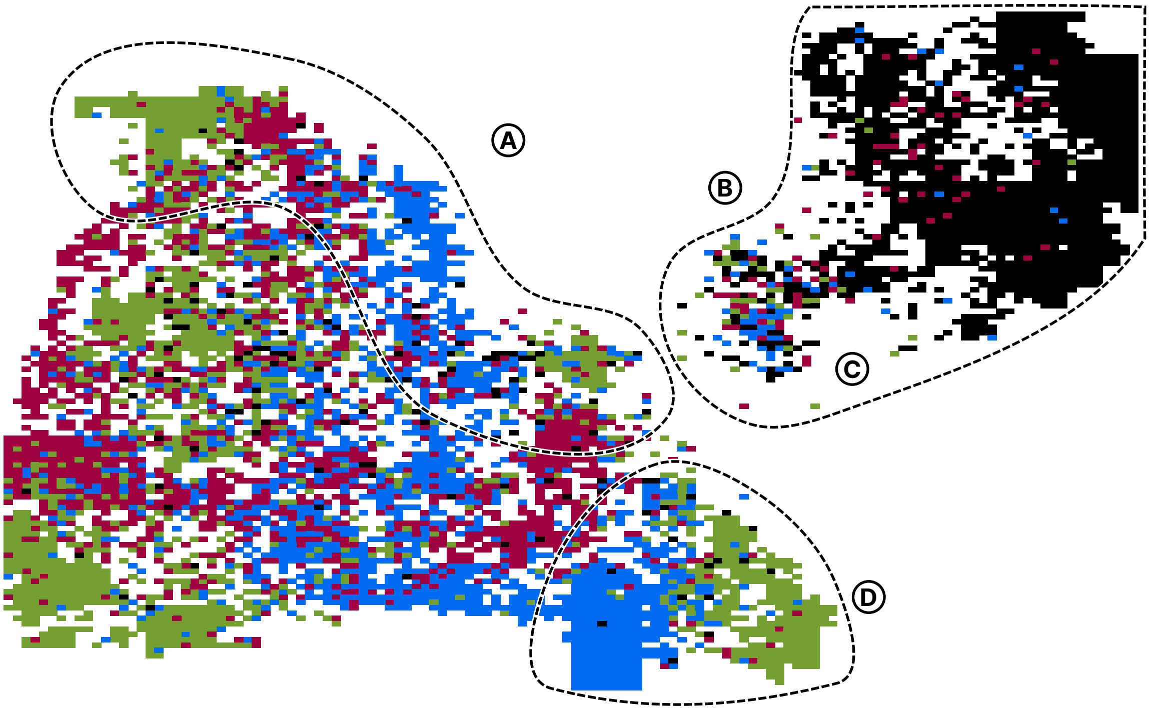
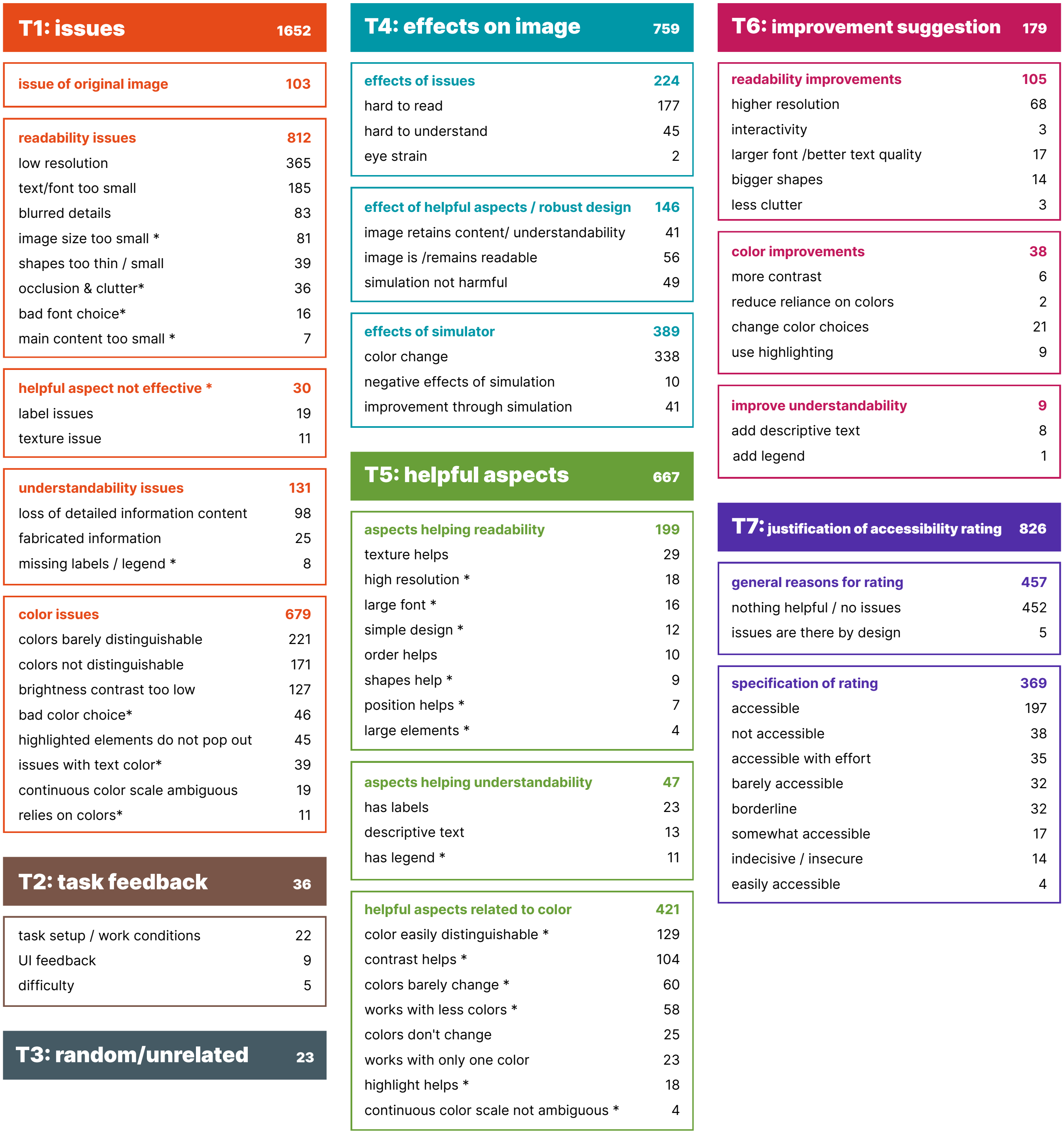


Figure 17: Gridified 2D embedding of the coding results with UMAP. Each cell represents a stimulus as viewed with a specific color vision deficiency and it is colored accordingly P D T M. Ⓐ to Ⓓ again represent the same interesting image clusters.



T1: issues	1652
issue of original image	**103**
readability issues	**812**
low resolution	365
text/font too small	185
blurred details	83
image size too small *	81
shapes too thin / small	39
occlusion & clutter*	36
bad font choice*	16
main content too small *	7
helpful aspect not effective *	**30**
label issues	19
texture issue	11
understandability issues	**131**
loss of detailed information content	98
fabricated information	25
missing labels / legend *	8
color issues	**679**
colors barely distinguishable	221
colors not distinguishable	171
brightness contrast too low	127
bad color choice*	46
highlighted elements do not pop out	45
issues with text color*	39
continuous color scale ambiguous	19
relies on colors*	11

T2: task feedback	36
task setup / work conditions	22
UI feedback	9
difficulty	5

T3: random/unrelated	23

T4: effects on image	759
effects of issues	**224**
hard to read	177
hard to understand	45
eye strain	2
effect of helpful aspects / robust design	**146**
image retains content/ understandability	41
image is /remains readable	56
simulation not harmful	49
effects of simulator	**389**
color change	338
negative effects of simulation	10
improvement through simulation	41

T5: helpful aspects	667
aspects helping readability	**199**
texture helps	29
high resolution *	18
large font *	16
simple design *	12
order helps	10
shapes help *	9
position helps *	7
large elements *	4
aspects helping understandability	**47**
has labels	23
descriptive text	13
has legend *	11
helpful aspects related to color	**421**
color easily distinguishable *	129
contrast helps *	104
colors barely change *	60
works with less colors *	58
colors don't change	25
works with only one color	23
highlight helps *	18
continuous color scale not ambiguous *	4

T6: improvement suggestion	179
readability improvements	**105**
higher resolution	68
interactivity	3
larger font /better text quality	17
bigger shapes	14
less clutter	3
color improvements	**38**
more contrast	6
reduce reliance on colors	2
change color choices	21
use highlighting	9
improve understandability	**9**
add descriptive text	8
add legend	1

T7: justification of accessibility rating	826
general reasons for rating	**457**
nothing helpful / no issues	452
issues are there by design	5
specification of rating	**369**
accessible	197
not accessible	38
accessible with effort	35
barely accessible	32
borderline	32
somewhat accessible	17
indecisive / insecure	14
easily accessible	4

Figure 18: Topics we identified during our content analysis with their subtopics and categories and absolute occurrences in the data. The * signifies new issues or helpful aspects the workers identified.

we identified might serve as a checklist when preparing accessible figures.

Provide helpful aspects to mitigate issues. As our data indicates, helpful aspects could help to make an image more accessible and robust. This is also stressed by participants: *"This has so many labels and text going on that it would be hard to mess up" (P45).* However, Jambor et al. [37] warns not to overdo labeling and descriptive text, as that might be an accessibility pitfall of its own, when introducing clutter (see also the paragraph below).

Keep figures simple. According to our ranking, simply designed figures seemed to do better. Simplicity is also reflected as helpful in some participants comments: *"[...]simplicity [...] that help[s] accessibility" (P148).* More detailed design might need more colors as stated by *P6: "[T]oo much detail needed in the graphs for them to be colorless."* We acknowledge, however, that this might not be trivial to do when trying to convey complex information.

Keep figures readable. In our study, resolution seems to correlate with the accessibility of an image. Furthermore, participants criticized an image as being too small to reflect the main content: *"Main elements of the image are too small for visualization and understanding of the figure" (P49).* Making images small is often one go-to strategy when trying to save scarce space in publications according to our personal experiences. Thus, we support Jambor et al.'s [37] suggestion for more conferences to drop the page limit in favor of a word limit similar to ACM CHI to give authors the freedom to make figures as large as need be. The frequent occurrence of resolution issues might be dataset-specific and dependent on hardware, as well as the way a figure is presented (interactive vs. static). Studies on interactive graphics on the web might yield different results regarding the occurrence of the problem. However, we still argue that one should bare readability problems and compensation strategies thereof in mind.

Pay attention to color nevertheless. While color is not the only aspect affecting accessibility, it is an important factor, as seen most clearly in the monochromacy condition which has the worst accessibility ratings. In particular, continuous color scales and highlighting could be affected by a rarer CVD like tritanomaly, while the color design for deuteranomaly seems already to be fairly robust (see, e.g., Figure 17).

Be aware of diverse vision abilities. Related work [3, 20, 37] urges to use simulators to test for CVD compatibility. We too encourage others to do so, however, also refer to other related work advocating to do so with care [81]. Speaking from the personal experience of the authors, taking part in the label identification phase really changed and sharpened our perspective regarding accessibility issues. Increasing awareness was also noted by *P49* in an email concerning the crowdsourcing phase: *"[The task] made me have more empathy with people who face [CVDs], made me search the internet to better understand the subject and thus be able to perform the task more effectively."*

5.2 Challenges and Opportunities of our Study Experience

We experienced researching CVD accessibility on a larger scale as a nontrivial endeavor. In the following sections, we revisit the challenges we faced during the phases of our study and discuss opportunities.

5.2.1 Study Design.

Deciding on what to evaluate. Even though we designed for a large-scale study, we had to limit ourselves somewhere. In our case, we decided not to explicitly rate the original images in the crowdsourced labeling phase. We suspect that this evaluation might have provided additional insights regarding accessibility ratings and issues. However, some issues like the loss of detail or loss of highlighting and others are only clearly visible in comparison. Furthermore, generic issues like text size could provide hints on the accessibility of the original image. From a more pragmatic perspective, if we had included the original image evaluation as an additional task, it would have resulted in higher costs. While debating this option internally, we rather opted for the consideration of more CVDs as well as more coders per image and a broader time span. Nevertheless, future studies could further investigate the impact of the original image accessibility of course.

Leveraging the perspective of experts and the crowd. Previous related work [20, 37] relied mostly on experts for their accessibility evaluation, producing results with higher internal validity, but potentially lacking ecological validity, due to the nature of expert reviews [82]. While we relied on experts for the initial label identification, we also leveraged abilities of the crowd to gain additional knowledge on how accessibility issues and helpful aspects are perceived by non-experts. Even though workers could provide their own aspects in free-text fields, existing check boxes might have influenced their answers and limited the addition of more aspects to consider. For future work, it would be interesting to directly compare the perspective of the crowd to the perspective of experts to identify potential limitations: What issues are only recognized by experts and not by the crowd and vice versa? If we had not provided check boxes but only free-text fields, would the answers have differed?

Accessibility labeling is not a standard labeling task. Compared to labeling tasks for ML, accessibility labeling adds complexity [77], which was also remarked by workers in the feedback after the training phase, e.g., by P148:*"I found this task more challenging than expected."* As discussed before, this had consequences on the study costs. Furthermore, we opted to iteratively improve our task design based on the feedback we gathered during the pilot and the training phase. However, we only decided to include minor improvements to not bias the results.

5.2.2 Study Execution.

Communication with workers helps to tackle insecurity. During an initial pilot, we struggled to get our tasks done due to the fear of task rejection and the skepticism toward unknown task providers expressed by crowdworkers. This insecurity was also reflected in the feedback after our training task and in the overall comments: *"I* *really hope I don't get rejected. I'm trying my best!" (P64).* To counteract that fear, we actively communicated with workers via e-mail and through the presence within the crowdworking community on Slack[3] and other websites[4]. Around 200 mails and messages were sent to build trust and reduce insecurity. Based on our experience, we believe communication with workers to be essential to a successful collaboration and encourage other researchers to also seek direct communication channels, particularly when offering subjective tasks, although this communication was time-consuming in our case.

Benefiting from the crowdworkers' (intrinsic) motivation. While there was the insecurity mentioned above, we also noted a high motivation and positive feedback on our accessibility tasks, similar to other related work crowdsourcing accessibility tasks [77]. 85 out of 135 feedback statements on the task after training were positive, while the remaining stressed challenges (16), found the task "*easy*" (2), or provided critical improvement suggestions (34). Furthermore, in the general feedback, participants emphasized on gaining knowledge *"[...] I learned a great deal about visualization accessibility [...]" (P52)* or also their intrinsic motivation for the task: *"I was legally blind until two years ago. I judged some of these on the issues I had back then and I hope it helps" (P84)* or *"This was fascinating to me since I'm a long time artist and have dealt with poor images for many years" (P13).*

5.2.3 Data Evaluation.

Taking care of data quality is challenging. As discussed in detail in Section 4.1, by not setting gold standards for labeling, we received more diverse answers, which were challenging to control for noise. Sanity checks were more time-consuming as one had to look deeper into the data to find random answers to remove. However, this study aimed to discover potentially interesting patterns for future investigation, for which we argue the current data quality was sufficient. Our assumptions should be further tested in more controlled setups, to gather more evidence for or against them.

Accessibility aspects are hard to agree on. Participants seemed to agree more on the positive side of accessibility ratings as mentioned in Section 4.2.1. Furthermore, it seemed to be difficult to judge the severity of an issue in relation to the accessibility rating: *"I don't know if [the image] will work due to not seeing the green or red lines and [I] don't know how to rate that" (P2).*

Additionally, it was very rare that all raters agreed on issues and helpful aspects in general. Raters unanimously selected the same helpful aspect for 2 % of all 6,000 simulated images. For issues, this was just the case for 0.2 % of the pictures.

Highest were the unanimous votes for deuteranomaly and protanomaly pictures regarding *has labels*: all workers found labels in 18 % and 13 % of the simulated images of the respective CVD. Thus, the unison ratings for helpful aspects seem slightly higher than those for issues. It could be that helpful aspects like labels are easier to spot (they are either there or not), while issues related to color are inherently subjective. More research is to be done to confirm those observations. Further investigation on the degree of harmful or helpfulness of aspects could bring additional insights as well as also expressed by P158: *"I wish there was some type of slider to indicate how bad a problem is."*

Determining the image type is nontrivial. Our first naive approach to gaining insight into image types was by using the previous conference tracks (Vis/SciVis, InfoVis, and VAST). However, after internal discussion, we abandoned this evaluation because VIS does not distinguish between conference tracks anymore[5] since 2021. Furthermore, the conference track does not necessarily determine all visualization types used, e.g., a SciVis-related paper could still use a bar chart to illustrate their results. Hence, we decided to annotate the image types of the 1,500 images ourselves into coarse categories. Creating a more detailed typology of visualization images is beyond the scope of this paper.

5.2.4 The Minority is Worth a Look.

We tried to get another more individual perspective on the data by looking at both majority and minority judgments. The minority ratings provided an opportunity to look at the data independently of the factor of agreement, which might as well add biases in the case of accessibility research, where one should also take into account individual opinions [53]. Through the minority ratings, we learned that in particular some color issues were often more problematic to individuals than the majority as discussed in Section 4.2.2. Additionally, minority ratings are useful to identify cases of rarer issues like *fabricated information*. However, we stress the need for multiple perspectives to avoid biases or unwanted noise. While minority ratings in our case were useful to identify subtleties in the data, majority or average ratings could be more useful to provide a more general overview.

5.3 Next Steps Toward Visualization Accessibility

In the following, we outline future research directions to gain a better understanding on the complex matter of visualization accessibility by extending our efforts into different directions, as well as potential next steps for more tool support for visualization accessibility.

5.3.1 Toward a better Understanding of Visualization Accessibility.

How do people with CVDs actually experience figure accessibility? A CVD simulation does not fully reflect the abilities of a person with CVDs. Over time, they might develop individual coping strategies to partially compensate the CVD. Nevertheless, we argue that if used with care and with the aforementioned limitations in mind, simulations could still be useful to build bridges and share experiences. We tried to incorporate actual experiences of people with CVDs both in our label identification and in our crowdsourcing phase. However, the data stemming from eight people with actual or potential CVDs was not enough data to detect patterns for people with CVDs specifically. More studies should be run to determine the actual impact of our issues and helpful aspects on the perception of images by people with CVDs.

What is the impact of simulation parameters? Additionally to our main exploration, we started to investigate the effect of simulation

[3] https://turkernation.slack.com/
[4] https://turkerview.com/
[5] http://ieeevis.org/year/2021/info/call-participation/call-for-participation

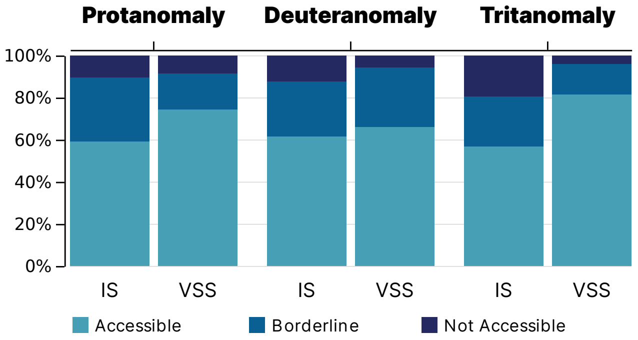


Figure 19: The simulator we use (IS) vs. the VSS from related work [73].

parameters by comparing a subset of our data to another simulator, the Visual System Simulator (VSS) by Schulz et al. [73] with its many adjustable parameters. A first glimpse of the simulation and data showed that monochromacy is not comparable to begin with, because VSS introduces a blur effect. The reason is a reduced number of photo receptors that affect visual acuity [73], which leaves merely blurred shapes as images. For the results of the other deficiencies, see Figure 19. Further research is required to investigate other simulators and parameter settings.

What issues are faced by other disabilities? Visualization accessibility has many facets besides CVD [45, 52, 91]. We purposely started our accessibility review in the CVD domain as it is already well researched and we could draw on many existing findings to generate our labels to ensure a certain validity. However, to get a more complete picture of visualization accessibility in general, it would be necessary to extend our findings to other disabilities as well.

How do visualization context, visualization types, and concrete tasks interact with accessibility? We have specifically chosen a design without a task in mind to foster unconstrained exploration of different visualizations. A subsequent study could be based on certain tasks and context to investigate more specific aspects in a controlled fashion. Our current results could inform the type of stimuli to be investigated. Furthermore, we only scratched the surface with broad visualization type categories. Studying the connections of more specific visualization types to certain issues might be another interesting subsequent study: do medical or scientific visualizations rely more on color and continuous color scales while bar charts struggle with too small labels of the data? Moreover, our study framework could be applied to other domains like visualizations in newspapers or social media to yield interesting results.

What are barriers of accessible visualization design? The fact that minority and majority ratings differ could have another reason besides the difficulty to judge in general. The differences could hint to potential trade-offs while designing visualizations. While going for a good design for the majority, issues of the minority might be overlooked. An example from our own paper writing: choosing aesthetically pleasing colors and designs might discriminate CVD users. Here we tried to opt for the opposite, which resulted in less pleasing but accessible colors. Another trade-off is the simplicity of design vs. the richness of the information. Further investigation should be done in this regard by asking the following questions: what are trade-offs in accessible visualization design and how could we minimize them? For what kind of visualizations do accessible design efforts reach their limits?

5.3.2 Toward More Tool Support for Visualization Accessibility. Accessible visualization design needs more tool support [12]. We envision tools that improve accessibility from the general or the individual perspective.

Developing metrics for accessibility checks. Existing visualization linters already detect basic design issues [31, 56]. With extensions, they could target specific accessibility use cases and automatically identify problematic areas in images. Our labels are not yet usable as metrics, further studies are needed to assess the impact and quantification of the issues and helpful aspects by asking questions like: how do issues or helpful aspects influence accessibility ratings in detail? What is a possible scale for accessibility? The fact that agreement on issues and accessibility ratings was scarce, however, indicates that this is no trivial task, as tools might need to train on and cope with potentially noisy data.

Designing for and with individual abilities. To capture issues that are not implementable by standard metrics, one might include the minority perspective again to gain more insights. We could envision ability-aware visualizations, where as a first step the user's individual abilities or preferences are learned and saved as parameters for visualization creation, which could be either applied proactively or changed in retrospect similar to previous work [18, 67].

6 CONCLUSION

In this paper, we assessed the CVD accessibility of 1,710 paper figures in a large-scale exploratory study. We gained insights into aspects helping or hindering accessibility and further explored clusters, potential trends, and individual images. Overall, 60 % of the images was rated accessible by the majority, however, given the fact that almost every image still had issues, we believe there is still room to improve visualization design with respect to accessibility. Here, we envision efforts toward automatic accessibility evaluation as well as ability-aware visualization as beneficial. On a meta-level, we learned that accessibility evaluation is not a trivial task and data evaluation can be challenging. The subjective nature of accessibility led us to explore the data from different directions, especially the perspective of the minority and majority. We believe that our identified issues and helpful aspects could support authors when designing visualizations. Furthermore, our study experiences and research directions could also help and inform fellow researchers in future endeavors to make visualizations more accessible.

ACKNOWLEDGMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (projects A08 and B01). We thank all our study participants and in particular Sajid Baloch for his valuable input.

REFERENCES

[1] Accessibility Guidelines Working Group. 2018. *Web content accessibility guidelines (WCAG) 2.1.* World Wide Web Consortium (W3C) Web Accessibility Initiative (WAI). https://www.w3.org/TR/WCAG21/. Accessed: 2021-09-04.

[2] Tania Acosta, Patricia Acosta-Vargas, Jose Zambrano-Miranda, and Sergio Lujan-Mora. 2020. Web accessibility evaluation of videos published on YouTube by worldwide top-ranking universities. *IEEE Access* 8 (2020), 110994–111011. https://doi.org/10.1109/access.2020.3002175

[3] Sakire Aytac. 2018. Using color blindness simulator during user interface development for accelerator control room applications. In *Proceedings of the International Conference on Accelerator and Large Experimental Control Systems (ICALEPCS '17).* JACoW, Geneva, Switzerland, 1958–1963. https://doi.org/10.18429/JACoW-ICALEPCS2017-THSH103

[4] Rita Borgo, Luana Micallef, Benjamin Bach, Fintan McGee, and Bongshin Lee. 2018. Information visualization evaluation using crowdsourcing. *Computer Graphics Forum* 37, 3 (jun 2018), 573–595. https://doi.org/10.1111/cgf.13444 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13444

[5] Roxana Bujack, Terece L. Turton, Francesca Samsel, Colin Ware, David H. Rogers, and James Ahrens. 2018. The good, the bad, and the ugly: A theoretical framework for the assessment of continuous colormaps. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (jan 2018), 923–933. https://doi.org/10.1109/tvcg.2017.2743978

[6] Lydia Byrne, Daniel Angus, and Janet Wiles. 2016. Acquired codes of meaning in data visualization and infographics: Beyond perceptual primitives. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (jan 2016), 509–518. https://doi.org/10.1109/tvcg.2015.2467321

[7] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17).* ACM, 2334–2346. https://doi.org/10.1145/3025453.3026044

[8] Kathy Charmaz. 2014. *Constructing grounded theory.* SAGE Publications Ltd.

[9] Jian Chen, Meng Ling, Rui Li, Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Torsten Moller, Robert S. Laramee, Han-Wei Shen, Katharina Wunsche, and Qiru Wang. 2021. VIS30K: A collection of figures and tables from IEEE Visualization Conference publications. *IEEE Transactions on Visualization and Computer Graphics* 27, 9 (sep 2021), 3826–3833. https://doi.org/10.1109/tvcg.2021.3054916

[10] Jinho Choi, Sanghun Jung, Deok Gun Park, Jaegul Choo, and Niklas Elmqvist. 2019. Visualizing for the non-visual: Enabling the visually impaired to use visualization. *Computer Graphics Forum* 38, 3 (jun 2019), 249–260. https://doi.org/10.1111/cgf.13686

[11] Colour Blind Awareness CIC. 2021. Colour blindness. https://www.colourblindawareness.org/colour-blindness/. Accessed: 2020-09-16.

[12] Katie Cornish, Joy Goodman-Deane, Kai Ruggeri, and P. John Clarkson. 2015. Visual accessibility in graphic design: A client–designer communication failure. *Design Studies* 40 (sep 2015), 176–195. https://doi.org/10.1016/j.destud.2015.07.003

[13] Rene Cutura, Cristina Morariu, Zhanglin Cheng, Yunhai Wang, Daniel Weiskopf, and Michael Sedlmair. 2021. Hagrid—Gridify scatterplots with hilbert and gosper curves. In *Proceedings of the International Symposium on Visual Information Communication and Interaction (VINCI).*

[14] Joel Ekman. 2017. *Automatic detection of issues related to colour vision deficient internet users.* Master's thesis. KTH, Media Technology and Interaction Design, MID.

[15] Christin Engel and Gerhard Weber. 2017. Analysis of tactile chart design. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '17).* ACM, 197–200. https://doi.org/10.1145/3056540.3064955

[16] David R. Flatla, Alan R. Andrade, Ross D. Teviotdale, Dylan L. Knowles, and Craig Stewart. 2015. ColourID: Improving colour identification for people with impaired colour vision. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* ACM, 3543–3552. https://doi.org/10.1145/2702123.2702578

[17] David R. Flatla and Carl Gutwin. 2010. Individual models of color differentiation to improve interpretability of information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10).* ACM Press, 2563–2572. https://doi.org/10.1145/1753326.1753715

[18] David R. Flatla and Carl Gutwin. 2012. "So that's what you see!": building understanding with personalized simulations of colour vision deficiency. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility (ASSETS '12).* ACM Press, 167–174. https://doi.org/10.1145/2384916.2384946

[19] Daniel Flück. 2016. Coblis - Color Blindness Simulator. https://www.color-blindness.com/coblis-color-blindness-simulator/. Accessed: 2020-11-28.

[20] Andrew Frane. 2015. A call for considering color vision deficiency when creating graphics for psychology reports. *Journal of General Psychology* 142, 3 (jul 2015), 194–211. https://doi.org/10.1080/00221309.2015.1063475

[21] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. 2020. Making GIFs accessible. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20)* (Virtual Event, Greece) *(ASSETS '20).* ACM, Article 24, 10 pages. https://doi.org/10.1145/3373625.3417027

[22] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019. Making memes accessible. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19).* ACM, New York, NY, USA, 367–376. https://doi.org/10.1145/3308561.3353792

[23] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A browser extension to make Twitter images accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20).* ACM, 1–12. https://doi.org/10.1145/3313831.3376728

[24] Joy Goodman-Deane, Patrick M. Langdon, P. John Clarkson, Nicholas H. M. Caldwell, and Ahmed M. Sarhan. 2007. Equipping designers by simulating the effects of visual and hearing impairments, In Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility - Assets '07. *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '07)*, 241–242. https://doi.org/10.1145/1296843.1296892

[25] Fredrik Hansen, Josef Jan Krivan, and Frode Eika Sandnes. 2019. Still not readable? An interactive tool for recommending color pairs with sufficient contrast based on existing visual designs. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19).* ACM, 636–638. https://doi.org/10.1145/3308561.3354585

[26] Kotaro Hara, Vicki Le, and Jon Froehlich. 2013. Combining crowdsourcing and google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13).* ACM Press, New York, New York, USA, 631. https://doi.org/10.1145/2470654.2470744

[27] Gunnar Harboe and Elaine M. Huang. 2015. Real-world affinity diagramming practices: Bridging the paper-digital gap. In *Proc. 33rd Annual ACM Conf. Human Factors in Computing Systems.* ACM, 95–104. https://doi.org/10.1145/2702123.2702561

[28] Mark Harrower and Cynthia A. Brewer. 2017. ColorBrewer.org: An online tool for selecting colour schemes for maps. In *Landmarks in Mapping.* Vol. 40. Routledge, 184–200. https://doi.org/10.4324/9781351191234-18

[29] 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 (CHI '10)* (Atlanta, Georgia, USA) *(CHI '10).* ACM Press, New York, NY, USA, 203–212. https://doi.org/10.1145/1753326.1753357

[30] Danula Hettiachchi, Senuri Wijenayake, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2020. How context influences cross-device task acceptance in crowd work. In *Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing (HCOMP '20)*, Vol. 8. 53–62. https://ojs.aaai.org/index.php/HCOMP/article/view/7463

[31] Aspen K. Hopkins, Michael Correll, and Arvind Satyanarayan. 2020. VisuaLint: Sketchy in situ annotations of chart construction errors. *Computer Graphics Forum* 39, 3 (jun 2020), 219–228. https://doi.org/10.1111/cgf.13975

[32] Yasuyo G. Ichihara, Masataka Okabe, Koichi Iga, Yosuke Tanaka, Kohei Musha, and Kei Ito. 2008. Color universal design: the selection of four easily distinguishable colors for all color vision types. In *Color Imaging XIII: Processing, Hardcopy, and Applications*, Reiner Eschbach, Gabriel G. Marcu, and Shoji Tominaga (Eds.), Vol. 6807. SPIE, 68070O. https://doi.org/10.1117/12.765420

[33] Petra Isenberg, Florian Heimerl, Steffen Koch, Tobias Isenberg, Panpan Xu, Charles D. Stolper, Michael Sedlmair, Jian Chen, Torsten Möller, and John Stasko. 2017. Vispubdata.org: A metadata collection about IEEE Visualization (VIS) publications. *IEEE Transactions on Visualization and Computer Graphics* 23, 9 (sep 2017), 2199–2206. https://doi.org/10.1109/tvcg.2016.2615308

[34] Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. 2017. Visualization as seen through its research paper keywords. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (jan 2017), 771–780. https://doi.org/10.1109/tvcg.2016.2598827

[35] Shinobu Ishihara. 1972. *Tests for Colour-Blindness.* Kanehara Shuppan Co., Ltd., Tokyo, Japan. http://www.dfisica.ubi.pt/~hgil/p.v.2/Ishihara/Ishihara.24.Plate.TEST.Book.pdf

[36] Jason T. Jacques and Per Ola Kristensson. 2017. Design strategies for efficient access to mobile device users via amazon mechanical turk. In *Proceedings of the First ACM Workshop on Mobile Crowdsensing Systems and Applications.* ACM. https://doi.org/10.1145/3139243.3139247

[37] Helena Jambor, Alberto Antonietti, Bradly Alicea, Tracy L. Audisio, Susann Auer, Vivek Bhardwaj, Steven J. Burgess, Iuliia Ferling, Małgorzata Anna Gazda, Luke H. Hoeppner, Vinodh Ilangovan, Hung Lo, Mischa Olson, Salem Yousef Mohamed, Sarvenaz Sarabipour, Aalok Varma, Kaivalya Walavalkar, Erin M. Wissink, and Tracey L. Weissgerber. 2021. Creating clear and informative image-based figures for scientific publications. *PLOS Biology* 19, 3 (mar 2021), e3001161. https://doi.org/10.1371/journal.pbio.3001161

[38] Luke Jefferson and Richard Harvey. 2006. Accommodating color blind computer users. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '06)*, Vol. 2006. ACM Press, 40–47. https://doi.org/10.1145/1168987.1168996

[39] Luke Jefferson and Richard Harvey. 2007. An interface to support color blind computer users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, 1535–1538. https://doi.org/10.1145/1240624.1240855

[40] Bernhard Jenny and Nathaniel Vaughn Kelso. 2007. Color design for the color vision impaired. *Cartographic Perspectives* 58 (sep 2007), 61–67. https://doi.org/10.14714/cp58.270

[41] Michael Kalloniatis and Charles Luu. 2005. The perception of color. In *Webvision: The organization of the retina and visual system [Internet]*, Helga Kolb, Eduardo Fernandez, and Ralph Nelson (Eds.). University of Utah Health Sciences Center, Salt Lake City (UT), USA. https://www.ncbi.nlm.nih.gov/books/NBK11538/

[42] Nam Wook Kim, Shakila Cherise Joyner, Amalia Riegelhuth, and Yea-Seul Kim. 2021. Accessible visualization: Design space, opportunities, and challenges. *Computer Graphics Forum* 40, 3 (jun 2021), 173–188. https://doi.org/10.1111/cgf.14298

[43] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM Press, 453–456. https://doi.org/10.1145/1357054.1357127

[44] Katharina Krösl, Carmine Elvezio, Laura R. Luidolt, Matthias Hürbe, Sonja Karst, Steven Feiner, and Michael Wimmer. 2020. CatARact: Simulating cataracts in augmented reality. In *Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR '20)*. IEEE. https://doi.org/10.1109/ismar50242.2020.00098

[45] Bongshin Lee, Eun Kyoung Choe, Petra Isenberg, Kim Marriott, and John Stasko. 2020. Reaching broader audiences with data visualization. *IEEE Computer Graphics and Applications* 40, 2 (mar 2020), 82–90. https://doi.org/10.1109/mcg.2020.2968244

[46] Bongshin Lee, Kate Isaacs, Danielle Albers Szafir, G Elisabeta Marai, Cagatay Turkay, Melanie Tory, Sheelagh Carpendale, and Alex Endert. 2019. Broadening intellectual diversity in visualization research papers. *IEEE Computer Graphics and Applications* 39, 4 (July 2019), 78–85. https://doi.org/10.1109/mcg.2019.2914844

[47] Bongshin Lee, Nathalie Henry Riche, Amy K. Karlson, and Sheelagh Carpendale. 2010. SparkClouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (nov 2010), 1182–1189. https://doi.org/10.1109/tvcg.2010.194

[48] Wanda Li and David R. Flatla. 2019. 30 years later: Has CVD research changed the world?. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. ACM, 584–590. https://doi.org/10.1145/3308561.3354612

[49] Yang Liu and Jeffrey Heer. 2018. Somewhere over the rainbow: An empirical assessment of quantitative colormaps. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, Vol. 2018-April. ACM. https://doi.org/10.1145/3173574.3174172

[50] Rui Lopes, Daniel Gomes, and Luís Carriço. 2010. Web not for all: a large scale study of web accessibility. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A '10)*. ACM Press, 1–4. https://doi.org/10.1145/1805986.1806001

[51] Kecheng Lu, Mi Feng, Xin Chen, Michael Sedlmair, Oliver Deussen, Dani Lischinski, Zhanglin Cheng, and Yunhai Wang. 2021. Palettailor: Discriminable colorization for categorical data. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (feb 2021), 475–484. https://doi.org/10.1109/tvcg.2020.3030406

[52] Alan Lundgard, Crystal Lee, and Arvind Satyanarayan. 2019. Sociotechnical Considerations for Accessible Visualization Design. In *Proceedings of the 2019 IEEE Visualization Conference (VIS '19)*. IEEE, 16–20. https://doi.org/10.1109/visual.2019.8933762

[53] Kelly Mack, Emma McDonnell, Dhruv Jain, Lucy Lu Wang, Jon E. Froehlich, and Leah Findlater. 2021. What do we mean by "accessibility research"? A literature survey of accessibility papers in CHI and ASSETS from 1994 to 2019. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 371, 18 pages. https://doi.org/10.1145/3411764.3445412

[54] Rubén Alcaraz Martínez, Mireia Ribera Turró, and Toni Granollers Saltiveri. 2021. Methodology for heuristic evaluation of the accessibility of statistical charts for people with low vision and color vision deficiency. *Universal Access in the Information Society* (may 2021). https://doi.org/10.1007/s10209-021-00816-0

[55] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *ArXiv e-prints* (feb 2018). arXiv:1802.03426 https://arxiv.org/abs/1802.03426

[56] Andrew McNutt, Gordon Kindlmann, and Michael Correll. 2020. Surfacing visualization mirages. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376420

[57] Anton Mikhailov. 2019. Turbo, an improved rainbow colormap for visualization. https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html. Accessed: 2021-09-08.

[58] Aliaksei Miniukovich, Michele Scaltritti, Simone Sulpizio, and Antonella De Angeli. 2019. Guideline-based evaluation of web readability. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300738

[59] Shigeki Nakauchi and Tatsuya Onouchi. 2008. Detection and modification of confusing color combinations for red-green dichromats to achieve a color universal design. *Color Research and Application* 33, 3 (2008), 203–211. https://doi.org/10.1002/col.20404

[60] Scott Novotney and Chris Callison-Burch. 2010. Shared task: crowdsourced accessibility elicitation of eikipedia articles. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)* (Los Angeles, California) *(CSLDAMT '10)*. Association for Computational Linguistics, USA, 41–44. https://www.aclweb.org/anthology/W10-0706/

[61] Jamie R. Nuñez, Christopher R. Anderton, and Ryan S. Renslow. 2018. Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data. *PLoS ONE* 13, 7 (aug 2018), e0199239. https://doi.org/10.1371/journal.pone.0199239

[62] Masataka Okabe and Kei Ito. 2008. Color universal design (CUD): How to make figures and presentations that are friendly to colorblind people. *J* Fly: Data Depository for Drosophila Researchers* (2008). https://jfly.uni-koeln.de/color/

[63] Manuel M. Oliveira. 2013. Towards More Accessible Visualizations for Color-Vision-Deficient Individuals. *Computing in Science & Engineering* 15, 5 (sep 2013), 80–87. https://doi.org/10.1109/mcse.2013.113

[64] Chris Olston and Allison Woodruff. 2000. Getting portals to behave. In *Proceedings of the IEEE Symposium on Information Visualization 2000 (INFOVIS '00)*, Jock D. Mackinlay, Steven F. Roth, and Daniel A. Keim (Eds.). IEEE, Salt Lake City, Utah, USA, 15–25. https://doi.org/10.1109/infvis.2000.885087

[65] Mohana Kuppuswamy Parthasarathy and Vasudevan Lakshminarayanan. 2019. Color vision and color spaces. *Optics and Photonics News* 30, 1 (jan 2019), 44–51. https://doi.org/10.1364/opn.30.1.000044

[66] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* 46, 4 (dec 2014), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

[67] Jorge Poco, Angela Mayhua, and Jeffrey Heer. 2018. Extracting and Retargeting Color Mappings from Bitmap Images of Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (jan 2018), 637–646. https://doi.org/10.1109/tvcg.2017.2744320

[68] Khairi Reda and Danielle Albers Szafir. 2021. Rainbows revisited: modeling effective colormap design for graphical inference. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (feb 2021), 1032–1042. https://doi.org/10.1109/tvcg.2020.3030439

[69] Theresa-Marie Rhyne. 2020. Color basics for digital media and visualization. In *Proceedings of the ACM SIGGRAPH 2020 Courses (SIGGRAPH '20)*. ACM, 1–78. https://doi.org/10.1145/3388769.3407478

[70] Madalena Ribeiro and Abel J. P. Gomes. 2019. Recoloring algorithms for colorblind people: A survey. *Comput. Surveys* 52, 4 (sep 2019), 1–37. https://doi.org/10.1145/3329118

[71] Dominik Sacha, Leishi Zhang, Michael Sedlmair, John A. Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, and Daniel A. Keim. 2017. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (jan 2017), 241–250. https://doi.org/10.1109/tvcg.2016.2598495

[72] Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, and Jon Froehlich. 2019. Project Sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–14. https://doi.org/10.1145/3290605.3300292

[73] Christoph Schulz, Nils Rodrigues, Marco Amann, Daniel Baumgartner, Arman Mielke, Christian Baumann, Michael Sedlmair, and Daniel Weiskopf. 2019. A framework for pervasive visual deficiency simulation. In *Proceedings of 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR '19)*. IEEE, 1–6. https://doi.org/10.1109/vr44988.2019.9044164

[74] Michael Sedlmair, Andrada Tatu, Tamara Munzner, and Melanie Tory. 2012. A taxonomy of visual cluster separation factors. *Computer Graphics Forum* 31, 3pt4 (jun 2012), 1335–1344. https://doi.org/10.1111/j.1467-8659.2012.03125.x

[75] Irena Serna-Marjanovic, Anel Tanovic, and Ajla Cerimagic. 2020. Accessibility Standards and Their Implementation in Custom Data-Driven Maps. In *Proceedings of the 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO '20)*. IEEE, 1674–1679. https://doi.org/10.23919/mipro48935.2020.9245417

[76] Junko Shirogane, Yuko Iwase, Hajime Iwata, Miho Saito, and Yoshiaki Fukazawa. 2017. A method for converting colors for color-impaired people, considering saturation and contrast ratio. In *Communication papers of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS '17) (Annals of Computer Science and Information Systems, Vol. 13)*. IEEE, Prague„ 357–366. https://doi.org/10.15439/2017f151

[77] Rachel N. Simons, Danna Gurari, and Kenneth R. Fleischmann. 2020. "I hope this is helpful": Understanding crowdworkers' challenges and motivations for an image description task. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (oct 2020), 1–26. https://doi.org/10.1145/3415176

[78] Stephen Smart, Keke Wu, and Danielle Albers Szafir. 2020. Color crafting: Automating the construction of designer quality color ramps. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (jan 2020), 1215–1225. https://doi.org/10.1109/tvcg.2019.2934284

[79] Statcounter Global Stats. [n.d.]. Desktop screen resolution stats worldwide. https://gs.statcounter.com/screen-resolution-stats/desktop/worldwide/#monthly-202010-202103-bar. Accessed: 2021-08-31.

[80] Danielle Albers Szafir. 2018. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (jan 2018), 392–401. https://doi.org/10.1109/tvcg.2017.2744359

[81] Garreth W. Tigwell. 2021. Nuanced perspectives toward disability simulations from digital designers, blind, low vision, and color blind people. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 378, 15 pages. https://doi.org/10.1145/3411764.3445620

[82] Melanie Tory and Torsten Moller. 2005. Evaluating visualizations: do expert reviews work? *IEEE Computer Graphics and Applications* 25, 5 (sep 2005), 8–11. https://doi.org/10.1109/mcg.2005.102

[83] Stephen Uzor, Jason T. Jacques, John J. Dudley, and Per Ola Kristensson. 2021. Investigating the accessibility of crowdwork tasks on Mechanical Turk. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 381, 14 pages. https://doi.org/10.1145/3411764.3445291

[84] Susan VanderPlas and Heike Hofmann. 2016. Spatial reasoning and data displays. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (jan 2016), 459–468. https://doi.org/10.1109/tvcg.2015.2469125

[85] Yunhai Wang, Xin Chen, Tong Ge, Chen Bao, Michael Sedlmair, Chi-Wing Fu, Oliver Deussen, and Baoquan Chen. 2019. Optimizing color assignment for perception of class separability in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 820–829. https://doi.org/10.1109/tvcg.2018.2864912

[86] Zhiquan Wang, Huimin Liu, Yucong Pan, and Christos Mousas. 2020. Color blindness bartender: An embodied VR game experience. In *Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW '20)*. IEEE, IEEE, 519–520. https://doi.org/10.1109/vrw50115.2020.00111

[87] Colin Ware. 2004. *Information Visualization: Perception for Design.* Morgan Kaufmann, Boston, Massachusetts, USA.

[88] Michael L. Waskom. 2021. Choosing color palettes - seaborn 0.11.2 documentation. https://seaborn.pydata.org/tutorial/color_palettes.html. Accessed: 2021-09-04.

[89] Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software* 6, 60 (apr 2021), 3021. https://doi.org/10.21105/joss.03021

[90] Alex C. Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. 2019. The perpetual work life of crowdworkers. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (nov 2019), 1–28. https://doi.org/10.1145/3359126

[91] Keke Wu, Emma Petersen, Tahmina Ahmad, David Burlinson, Shea Tanis, and Danielle Albers Szafir. 2021. Understanding data accessibility for people with intellectual and developmental disabilities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 606, 16 pages. https://doi.org/10.1145/3411764.3445743

[92] Marco Zehe. 2019. Auditing for accessibility problems with firefox developer tools. https://hacks.mozilla.org/2019/10/auditing-for-accessibility-problems-with-firefox-developer-tools/. Accessed: 2021-09-08.

[93] Liang Zhou and Charles D. Hansen. 2016. A Survey of Colormaps in Visualization. *IEEE Transactions on Visualization and Computer Graphics* 22, 8 (aug 2016), 2051–2069. https://doi.org/10.1109/tvcg.2015.2489649