UTRECHT UNIVERSITY

Department of Information and Computing Science

Information Science Bachelor Thesis

Utrecht University

Validating Visual Separation Metric Kappa: A User Study Questionnaire

First Examiner:
Alexandru Telea

Second Examiner:
Michael Behrish

Candidate:
Carlijne Govers
2521970

July 7, 2023

Abstract

A new application of Cohens' κ on results of a pseudo-labeling algorithm on projections shows potential for evaluating projection techniques' effectiveness by measuring visual separation of labels. This study assesses the human perception of visual separation and examines whether there is a relationship with kappa. Conducting a user study survey including material of various datasets and projections, with many participants of various demographic backgrounds, demonstrates a comparison of visual separation of given scores and kappa. A Pearson Correlation tests shows there is a significant positive correlation between the two variables. Projections of lower kappa values show they are not representative for measuring human perception of visual separation.

Keywords: visual separation, projections, kappa

Table of Contents

1. Introduction 1
   1.1. Motive for the Study 1
   1.2. Research Objective 4
   1.3. Thesis Structure 5
2. Background 6
   2.1. Data Visualization 6
      2.1.1. Human Visual Perception 6
      2.1.2. Design Principles 7
   2.2. Projections 8
   2.3. Quality Metrics 9
      2.3.1. Kappa 9
3. Methods 10
   3.1. Material 10
      3.1.1. Datasets and Projections 10
      3.1.2. Metrics 12
      3.1.3. Scatterplot Images 13
   3.2. Survey Setup 15
      3.2.1. Survey Reasoning 15
      3.2.2. Participants 15
      3.2.3. Survey Contents and Flow 17
      3.2.4. Survey Randomization 18
      3.2.5. Survey Details and Looks 19
   3.3. Data Collection, Reformatting, and Analysis 20
4. Results 21
   4.1. Statistical Analysis 21
      4.1.1. Descriptive Statistical Analysis 21
      4.1.2. Inferential Statistical Analysis 24
   4.2. Qualitative Analysis 29
5. Discussion 30
6. Conclusion 32

References 33

Appendices 35

Appendix A. Survey Contents 35

Appendix B. Dataset Format 55

1. Introduction

1.1. Motive for the Study

Usage of data has taken an important role in many domains of current society. For instance, in the field of astronomy, Goodman (2012) highlights the importance of exploratory data visualization. Furthermore, data mining applications have become widespread. Thuraisingham (2000) provides examples, such as pharmaceutical companies using prescription analysis to target customers effectively, and credit bureaus making loan decisions based on observations of individuals with similar purchasing behaviors, income levels, and credit histories.

Data contains a number of data points that represent observations. Each data point consists of various dimensions that are also known as properties, features, variables, or attributes. Observations with their dimensions describe a certain phenomenon. In the case of high-dimensional data, the number of dimensions is close to or larger than the number of observations. High-dimensional data has applications in diverse domains, including biomedical research, web analytics, education, medicine, business intelligence, and social media. These applications encompass various data formats, such as text, digital images, speech signals, and videos (Ayesha et al., 2020).

Analyzing data is necessary to find patterns, relationships, outliers, or other observations that form information. Often, the underlying structure of the data is visible through classification, which involves labeled data, or clustering which groups similar data points together. Similarly to low-dimensional data, high-dimensional data can be analyzed through data visualizations for the use of either exploration, confirmation, and presentation (Rau et al., 2017). Examples of advanced visualization techniques that display multivariate data structures are mosaicplots, parallel coordinate plots, and trellis displays (Theus, 2008). However, high-dimensional data visualization techniques are limited in their ability to display data with a large amount of dimensions.

In addition to challenges within visualization, problems arise in high-dimensionality with respect to accurate classification and pattern recognition. Moreover, model learning is difficult due to high computational complexity (Ayesha et al., 2020). For a machine learning model to be effective, sufficient data points are required for each dimension. The curse of dimensionality implies that for increasing dimensionality, the amount of data points necessary for good performance of machine learning algorithms increases exponentially.

Rather than adding more data for each extra dimension, it is possible to perform dimension reduction through either feature extraction or feature selection. The dimension reduction techniques in machine learning used to reduce dimensions in high-dimensional data are also known as projection algorithms. Projections map nD data onto representational lower dimensional data of 2D or 3D while retaining patterns from the original data as well as possible. Their purpose ranges amongst others, from exploration of high dimensional data and model understanding, to creating better classification

methods.

The performance of a projection technique on a dataset determines the feasibility of executing such tasks. Visual separation is a benchmark used for analyzing high-dimensional data through projections. For a projection to exhibit ‘good’ visual separation, several desirable criteria include the formation of well-clustered points that are closely situated within the group, and devoid of overlapping with other groups.

Crucially, if data separation observed in the dataset aligns with the obtained visual separation of the projection, indicating a correspondence between the two, the separation can be employed in evaluating the classification of the dataset. The following two extremely simplified examples are artificially created, thus not based on a real dataset or projection and function solely for explanation purposes. Imagine we know that the used artificially created dataset has good data separation in clusters. Labels are used to define the data separation and visual separation. Then a projection technique creates a scatterplot on the data. The colors represent the labels. An example of good visual separation is shown in Figure 1. The labels are well separated from each other, thus the used projection technique captures the data separation of clusters well. Oppositely, an example of bad visual separation in Figure 2 does not capture the data separation well as labels are clearly mixed.

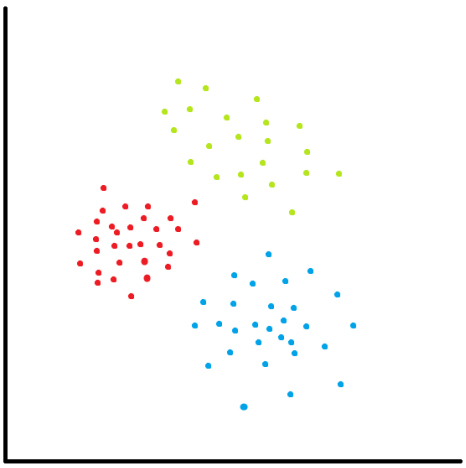

Figure 1. Example of good visual separation.

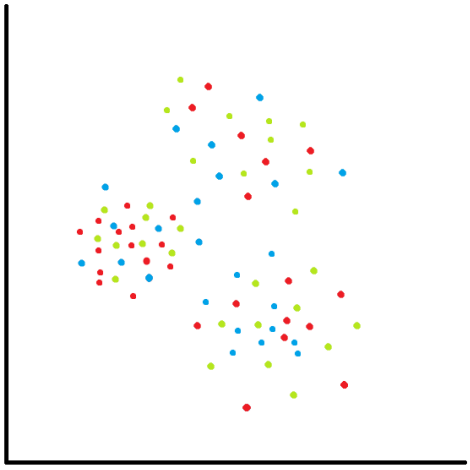

Figure 2. Example of bad visual separation.

This assessment, in turn, helps determine the suitability of the dataset-projection combination. Thus its usability for machine learning models becomes clear, following the principle of "garbarge in, garbage out" as well as gaining better model convergence. The process of transforming the data by reducing input dimensions, often as part of the machine learning pipeline, precedes the application of e.g. regression or classification models. The purpose of dimension reduction can also extend to simply gaining a better understanding of the multivariate relationships within the data, as mentioned earlier.

Metric kappa is used to assess performance of classifiers, and is applied on a certain classifier algorithm in recent research to measure the visual separation of labels. Projections visually map a 2D scatterplot of a dataset-projection combination. For this

combination a corresponding kappa value is measured. In theory, kappa represents how well the visual separation in the projection plot is. However, it is unknown whether this application of kappa is in line with human perception of visual separation. By extension, it is unknown whether it is accurate and usable in practice for its intended goal. The metric can be validated by conducting a user study in which participants are asked to score projections based on how well they think labels are separated. Those scores can be correlated with the metric generated by the algorithm. This approach allows comparison between the metric and human perception, while making use of representative datasets, projections and participants, etc. Furthermore, conducting a user study leads to more interesting insights on the metric for which other questions arise.

Conducting a user study allows us to judge the kappa's usability for its intended purpose. A positive correlation as a result of the user study would imply that kappa is a quick measure of visual separability of projections, and a tool for measuring a projections' effectiveness in the context of the labeled dataset.

1.2. Research Objective

Resulting from the lack of user experiments for evaluating the relationship between metric kappa and human perception, the following research question is presented.

Research question
Is there a positive relationship between metric kappa and human perception of visual separation?

Based on the research question the following hypotheses are defined.

Hypothesis
H0: There is no positive correlation between metric kappa and human perception of visual separation.
H1: There is a positive correlation between metric kappa and human perception of visual separation.

Sub-questions
-What is the influence of dataset or projection characteristics on the correlation between kappa and human perception score, and whether these characteristic values are suitable for human perception of visual separation.
-What is the influence of participant characteristics on the correlation between kappa and the human perception score, e.g. does a difference occur between scoring of participants with or without prior knowledge of data visualization or data analytics.

1.3. Thesis Structure

Effectively answering the stated research questions requires a structured approach. Section 2 provides a comprehensive review of existing literature, theories, and concepts that are relevant to the research topic, the experimental setup, and material. Section 3 describes the used research methodology and approach employed in the user study design, data collection methods, sample selection, and data analysis techniques used to answer the research question. Section 4 presents the findings from the data analysis including quantitative and qualitative data analysis. Section 5 interprets and analyses the results in light of the research question and hypothesis. Findings are compared with related literature and both implications and limitations are mentioned. Finally, Section 6 concludes the above, answers the research question, and recommends future practical applications and research directions.

2. Background

2.1. Data Visualization

Visualizing data enhances its interpretability for humans, which is particularly valuable in domains that deal with large volumes of data (Musa et al., 2016). The preceding procedure focuses on designing, developing, and implementing computer-generated graphical depictions of the data. Effective visualizations enable exploration, analysis, and various visualization tasks that aid in comprehending information, identifying patterns, and forming informed opinions. Consequently, such visualizations facilitate effective decision-making.

Among the existing visualization techniques available, numerous options emerge as suitable approaches for visualizing high-dimensional data. Grinstein et al. (2001) identifies several notable high-dimensional visualization techniques, including 2D and 3D scatterplots, matrix of scatterplots, heat maps, iconographic displays, multidimensional scaling, Sammon plots, and Grand Tours. Among these techniques, the scatterplot stands out as the most commonly utilized method for data visualization. A scatterplot represents data points in a 2D or 3D dimensional space. When bringing down nD dimensions to 2D or 3D, projections results in a scatterplot suitable for visual inspection.

2.1.1. Human Visual Perception

Visual inspection of data visualizations is susceptible to human perception. The process of visual perception for humans is described in a three-stage model by (Ware, 2019) and summarized as follows by Koponen & Hildén (2019). In short, initially the neurons in both the retina and visual cortex collaboratively seek out characteristics in the visual field such as orientation of edges, motion and colors. Subsequently the brain detects simple patterns that may be described by the Gestalt laws as visual features. Lastly, these simple patterns merge to form more intricate visual entities. Aside from storing these visual working memory, comparison with preceding memories lead to recognition and interpretation of visual variables.

Following from the initial stage come the so called three feature channels which are shape features, color features and motion that may emphasize data patterns. According to Koponen & Hildén (2019) visual processing tasks are most manageable if they are focused on one of these distinct features at a time. A non-exhaustive list of the distinguishable features based on various authors is described and may be found below:

- Shape: size, elongation, round vs. sharp, orientation, fill, closure
- Other shape features: texture, sharpness
- Position: grouping, quantity, density
- Color: hue, lightness, intensity, added surround color, opacity
- Motion and change: speed, direction, vibration

In the second stage, shapes formed from using multiple of the previously mentioned features recognized in the preceding stage, may be observed as groups. They are defined in the Gestalt laws which may be considered outdated in nowadays research. The seven most accepted rules and one general rule are listed below based on Koponen & Hildén (2019):

- Figure-ground articulation
- Law of proximity
- Law of common fate
- Law of similarity
- Law of continuity
- Law of closure
- Law of good Gestalt
- Past experience experience
- Connectedness and connecting regions

2.1.2. Design Principles

Designing well perceivable data visualizations is done according to the design principles that follow from human visual perception. The most important element in visualization design is consistency. Consistency is important because it allows for the crucial ability to compare within and between visualizations (Koponen & Hildén, 2019).

Visual variables encompass a diverse range of visual techniques employed to convey additional information within the graphical elements that represent data points. Depending on which data type is used, the visual variable is suitable so that we read the data precisely. The variables possess an inherent order of applicability depending on this data type as seen in Table 1. Across all data types, position is the most prominent and distinguishable encoding method.

Table 1. Accuracy of visual variables for data types, adapted from Koponen & Hildén (2019).

	Most accurate	**Less accurate**	**Least accurate**	**Not usable**
Ratio & interval	Position Length	Angle Slope Area	Volume Color density Color saturation	Color hue Texture Shape Connection
Ordinal	Position Color density	Color saturation Color hue Texture Connection Length	Angle Slope Area Volume	Shape
Nominal	Position Shape Color hue	Texture Connection Color density Color saturation	Length Angle Slope Area Volume	-

Color as one of the visual variables is a very powerful visual cue. The perception of color exists of three elements, hue, lightness and saturation, for which the human eye can observe up to a million shades of color. In cases of color vision anomalies this perception might differ. Most common is red-green colorblindness. Blue-yellow colorblindness is much rarer, as well as the inability to distinguish color at all which is extremely uncommon. Estimation is that 8% of men and 0.5% of women suffer one of the deficiencies.

While creating color palettes it is not safe to purely base color differences on hue as different colorblindness brings different vision. This causes for one palette to be clearly visible for one, but unclear for the other. Moreover, the most important of the three color elements for our perception is lightness. When aiming to create contrast for color salience and make an element stand out from its background, simply adjusting the hue and saturation is not sufficient. The color lightness should differ well with the lightness of the background. Depending on the use of your color, the palette might use natural colors, considering the concept of cold-blue and warm-red perception. A color key should be used when colors are ambiguous. There is also a distinction between qualitative and quantitative color scales. Qualitative scales are used for nominal data (categorical) in which it is important that each color most distinct as possible from each other. This can be done well up until an amount of 12 well-distinct colors, after which quality of distinction declines (Arnkil, 2021). Quantitative scales are used for numerical data (ratio and interval) as well as ordinal data (categorical), often using a graduated color scale as encoding.

2.2. Projections

As mentioned earlier, projections reduce dimensions with minimized information loss, to enable easier data analysis. There are many varying projection techniques, e.g. Principle Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Multidimensional Scaling (MDS) to name some well-known techniques. Projection techniques employ different mathematical algorithms and criteria to determine the optimal representation of the high-dimensional data in a lower-dimensional space. There is a trade-off between preserving original data characteristics and reducing dimensionality. How well the existing different projection techniques perform depends on both the data and projection technique. A review of different dimension reduction techniques is done in works such as Carreira-Perpiñán (1997) or more recently by Ray et al. (2021). Furthermore an overview and comparison of such dimensionality reduction techniques for high-dimensional data is given in works such as Van Der Maaten et al. (2009) or a more recent one by Ayesha et al. (2020). A most recent and extensive quantitative evaluation of projections is done by Espadoto et al. (2019). The study shows projections seem similar, based on the quality metrics used.

2.3. Quality Metrics

Literature research by Bertini et al. (2010) has created a systematic overview of quality metrics specifically within high-dimensional data visualization. A distinction between multiple characteristics of the quality metric is made. These are the visualization technique, what is measured, where it is measured, the purpose, and the interaction. The visual quality metrics have the goal of showing the user whether a visualization is usable, based on the metrics' specifications. Knowing what different visual metrics indicate also allows for comparison between the metrics.

Very commonly used metrics for assessing projection quality are the trustworthiness, continuity, normalized stress, Shepard correlation and distance preservation metrics. The quality metric neighborhood hit is the most commonly used metric for assessing projections of labeled data. It measures how well labeled points in 2D seem to be grouped with relation to the labeled points in nD. While this is estimates visual separation better than other metrics, it does not seem to measure what users perceive in visual separation.

Metrics useful for measuring visual perception are e.g. class consistency and distance consistency. However, characteristics of their way of functioning lead to difficulties in distinguishing complex shapes. There are many other visual metrics, however they do not seem to compare with a generalizable experimental setup of various projections and datasets.

2.3.1. Kappa

Kappa is a classification metric with as purpose, as mentioned in Section 1.1, to see if a projection is usable on the dataset. Formally kappa is known as Cohen's κ. It assesses the level of agreement or inter-rater reliability between two or more raters when categorizing or classifying data. The measure the extent to which the agreement observed between raters exceeds what would be expected by chance alone. Kappa is widely used in various fields, such as psychology, medicine, and social sciences, to evaluate the consistency or agreement in categorical data coding or classification tasks.

In the context of projections however, kappa has been used in a very recent study, as metric for a pseudo labeling algorithm to assess visual separation in a projection space. Naming some of the preceding research. e.g. validating semi-supervised deep learning based on label propagation in a 2D embedded space (Benato et al., 2021), or linking data separation, visual separation, and classifier performance using pseudo-labeling by contrastive learning (Benato et al., 2023), that have used kappa for classification measurement. Kappa has a value in the range [-1, 1], where $\kappa \leq$ means no possibility, and $\kappa = 1$ means full possibility, of agreement.

Current evaluation of this application of kappa on the algorithm exists of quantitative and qualitative analysis, showing outperformance compared to common projection quality metrics, making its application very promising.

3. Methods

This section provides an introduction to the methods employed in this study. It starts by elaborating on the material used to generate a number of projection plots, using different datasets and projection techniques. Important characteristics of the plots are outlined. Subsequently, the methodology for the distribution of the survey among participants and a description of the population sample, as well as creating the survey including its contents, flow, questions to ask scores, and randomization of questions among participants and the material is described. Lastly, the used correlation analysis between the collected data scores of participants with kappa is explained.

3.1. Material

3.1.1. Datasets and Projections

Datasets and projections used for generating various kappa's and scatterplot images are chosen in the prior research on kappa, based on label-containing datasets and projections that are widely used within the research field of machine learning and data reduction, and thus very generalizable to other research. Each projection technique's parameters are set to the default of the original author. An overview of the used datasets and projections is listed in Table 2 and Table 3 respectively. Other characteristics of the datasets and projections are out of scope for this research, but are very relevant to the generalizability of kappa's usage.

Table 2. Characteristics of the 18 used datasets. Number of labels after applying projection and number of datapoints are shown.

dataset	number of labels	number of datapoints
bank	2	2059
cifar10	10	3250
cnae9	9	1080
coil20	20	1440
epileptic	5	5750
fashion_mnist	10	3000
fmd	10	997
har	6	735
hatespeech	3	3221
hiva	2	3076
imdb	2	3250
orl	41	400
secom	2	1567
seismic	2	646
sentiment	2	2748
sms	2	835
spambase	2	4601
svhn	10	732

Table 3. List of 39 used projections.

projection	
DM	LSP
UMAP	LTSA
FA	MC
F-ICA	MDS
GPLVM	M-LLE
G-RP	N-MDS
H-LLE	NMF
IDMAP	PBC
I-PCA	PCA
ISO	PLSP
K-PCA-P	P-PCA
K-PCA-R	S-PCA
K-PCA-S	SPE
LAMP	S-RP
LE	T-SNE
L-ISO	T-SVD
LLC	UMAP
LLE	
L-TSA	
L-MDS	
L-MVU	
LPP	

As seen in Figure 2, 18 different datasets and 39 varying projections are used. Each dataset has at least 34 of these 39 projections applied to them resulting in 678 out of the maximum of 702 combinations.

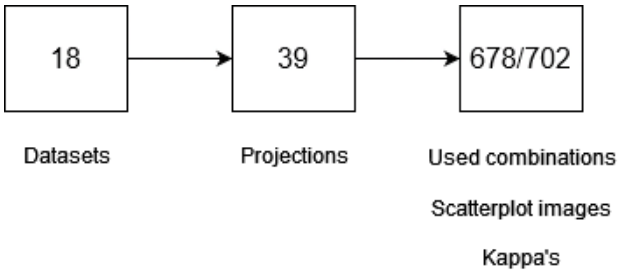

Figure 2. Process of generating images and kappa values.

For each dataset-projection combination, there is a new dataset containing datapoint coordinates and respective datapoint labels to generate a scatterplot image with. Dependent variables that are measured for each dataset-projection combination are the amount of labels, amount of data points, and manually labeled extreme cases of datapoint positioning of stacked points, stacked lines or hidden colors. These last three groups are manually categorized as outliers through manual review of the visualization. The category stacked points visually hides datapoints that overlap. The category stacked lines demonstrates datapoints ordered in lines and thus also overlapping slightly. These two categories possibly result in the last category, hidden colors. The legend shows all present colors, however, not all colors are visible. Examples of respective outlier images are shown in Figure 3, Figure 4, and Figure 5 below. The insight for these concepts are dependent on the participants' conceptual understanding. A lack of conceptual understanding can cause confusion for the participant, or result in deviating answers. The images categorized as outlier are retained in the survey to analyze whether these scores perform differently.

Figure 3. Example of category stacked points, from projection "imdb_M-LLE_0,390154".

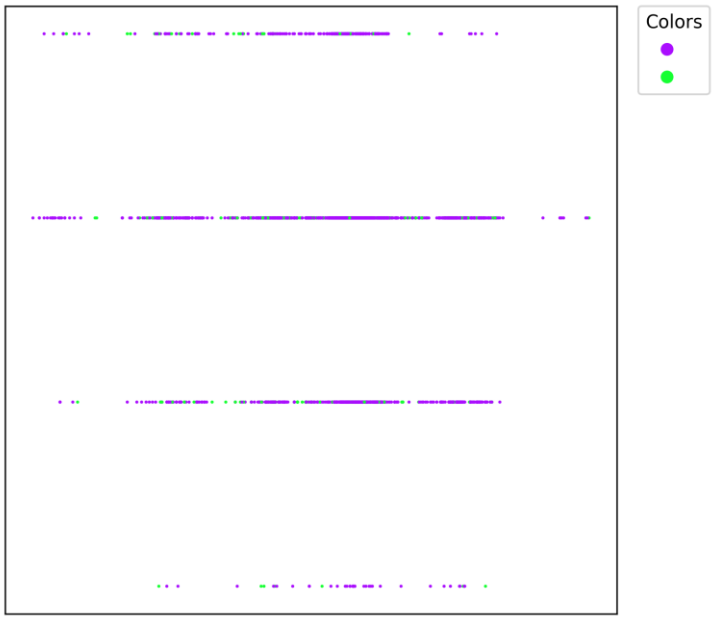

Figure 4. Example of category stacked lines, from projection “bank_S-RP_0,417644”.

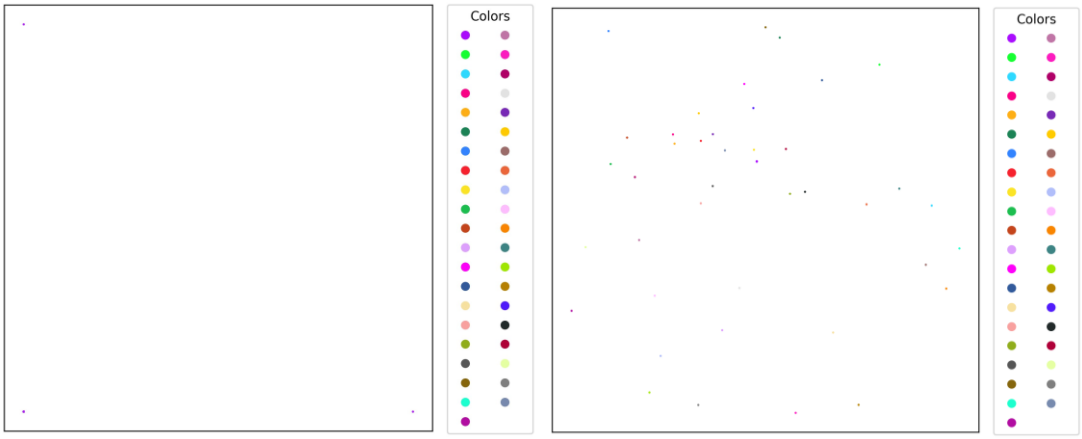

Figure 5. Examples of category hidden colors, from projections “orl_M-LLE_0,120833” (left) and “orl_LLC_0,685825” (right).

3.1.2. Metrics

Simultaneous to coordinate and label datasets, each combination results in calculated metrics being four scalar metrics and kappa. Usage of the other metrics than kappa is out of scope for this user study. The kappa scores in the used data vary from 0,120833 to 1. The distribution, as indicated in Figure 6, shows a division skewed to the right. To ensure equal testing of the whole value range the kappa is divided with an interval of 0.1, resulting in 9 groups referred to as round kappa. This equal interval level is deemed to have sufficient level of detail to see differences in the kappa range and allow for further participant image assignment based on this distribution which is discussed in section 3.2.4 Survey Randomization.

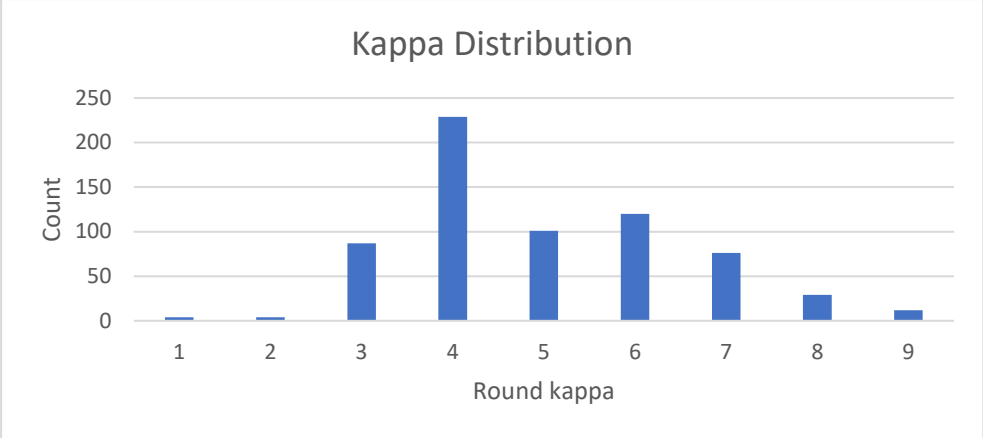

Figure 6. Bar chart showing the count of kappa values in the data.

3.1.3. Scatterplot Images

Important is to ensure that the scatterplot images generated from the coordinates and labels generated, are valid, qualitative and usable testing material. Keeping in mind basic design principles of data visualization all images are created consistently over controllable factors.

First of all, labels will be represented by colors since they are reliable visual cues for categorical data, see Section 2.1.2. Following this, visual separation for this study is defined by the separation of colored labeled groups from other colored labeled groups. The used color palette covers all labels from 2 to 41 with a static order, meaning every label number is assigned a color. All projection images use this same palette and order to mitigate cognitive load of processing a different order of label colors for each image. Label numbers with their corresponding colors and hex code are shown in Table 4. The color palette utilized in this study is based on the “alphabet” palette derived from “Polychrome 36” (Coombes et al., 2019). The original palette is generated by a tool that picks a well contrasting selection of colors suitable for qualitative palettes, even when dealing with a large number of categories. The alphabet includes colors up to a number 26. The remaining 14 colors are manually picked from Dark24, oldsky.colors and sky.colors (Coombes et al., 2019), so that all colors still contrast as well as possible in the full palette, looking at lightness and saturation. The order is decided manually with the knowledge that distinguishing colors can be done well with up to 12 colors. Each scatterplot image includes a legend showing the present label colors in that image. This is to provide participants expectations of what colors to look for as in some cases not all colors are fully visible or some are even completely hidden.

Table 4. Color palette presenting each label with corresponding color sample.

label	hex code	color	20	#B10DA1	
0	#AA0DFE		21	#C075A6	
1	#16FF32		22	#FC1CBF	
2	#2ED9FF		23	#B00068	
3	#FA0087		24	#E2E2E2	
4	#FEAF16		25	#782AB6	
5	#1C8356		26	#FFCB00	
6	#3283FE		27	#9B6C6A	
7	#F6222E		28	#EB663B	
8	#FBE426		29	#B1BEFA	
9	#1CBE4F		30	#FEBDFF	
10	#C4451C		31	#F98500	
11	#DEA0FD		32	#3D8484	
12	#FE00FA		33	#9FE600	
13	#325A9B		34	#B68100	
14	#F7E1A0		35	#511CFB	
15	#F8A19F		36	#222A2A	
16	#90AD1C		37	#AF0038	
17	#565656		38	#E5FFA3	
18	#85660D		39	#7F7F7F	
19	#1CFFCE		40	#778AAE	

Secondly, sizes of markers, legend, figure and image are balanced so that markers are large enough to be visible on screen and small enough to present the data in an interpretable manner. Another important factor for accessibility is image resolution that allows for participants to zoom in. Image quality details are found in Table 5 and are controlled using the Matplotlib library in Python. The images are saved as a PNG. Images are automatically rescaled by the survey platform to fit the question block and participants screen-size.

Table 5. Listed image quality characteristics of scatterplot images.

PNG size (px)	figure size (inches)	marker size	dots-per-inch (dpi)
1667 x 1446	6 x 6	0.5	300

3.2. Survey Setup

3.2.1. Survey Reasoning

This user study aims at gaining responses from many participants. As compensation for the generally low response-rate for surveys, they are easily distributed among platforms to reach a high amount of potential participants. Furthermore, the task is kept simple and is intended to be executed fast. This is because the task indicates an intuitive response, and is executed in a short total time span to retain cognitive attention span and prevent fatigue. Lastly, the 678 to be scored images are distributed among participants so that the distribution is uniform among participants and kappa. Randomization logics that are embedded in the survey platform Qualtrics are a suitable tool to control the question distribution flow.

3.2.2. Participants

The survey is sent to potential participants of various backgrounds as it is believed all humans are representative for perception of visual separation, regardless of experience in the data domain. Potential participants who were sent the survey consist of colleagues, friends, family, student colleagues and professors. This type of sampling is a clear case of convenience sampling. Participants were requested to distribute the survey in their social circle, which is the type snowball sampling. Distribution of the survey is done through various platforms such as Teams, Discord, WhatsApp, E-mail and LinkedIn. Tracing back the origin of responses is not possible due to the surveys' anonymity.

At the moment of data collection the sample consisted of a total of 108 survey responses. A full overview of the distribution of their attributes is found in Table 6. Most notable is that participants' ages ranged from 18 to 60 years or older of which 57% within the range of 21-30. Regarding education level ~2% of the participants finished intermediate vocational education and 34% of the participants finished secondary school. It is probable they are currently doing their Bachelors, this is explained by the amount of young people as can be seen in Figure 6 and considering that the survey is spread in student communities.

Table 6. Descriptive statistics of respondents.

		N (total = 108)
Gender	Female	37 (34%)
	Male	64 (59%)
	Non-binary / third gender	2 (2%)
	Prefer not to say	5 (5%)
Age (years)	18-20	16 (15%)
	21-30	62 (57%)
	31-40	9 (8%)
	41-50	4 (4%)
	51-60	12 (11%)
	> 60	5 (5%)
Education	Secondary school	37 (34%)
	Intermediate vocational education	2 (2%)
	Bachelor	44 (41%)
	Master	21 (19%)
	Doctor, PhD	4 (4%)
Experience (years)	< 1	13 (12%)
	1	7 (6%)
	2	15 (14%)
	3	5 (5%)
	4	3 (3%)
	5 >=	5 (5%)
	No	60 (55%)
Device	Laptop or desktop	51 (47%)
	Mobile phone	57 (53%)

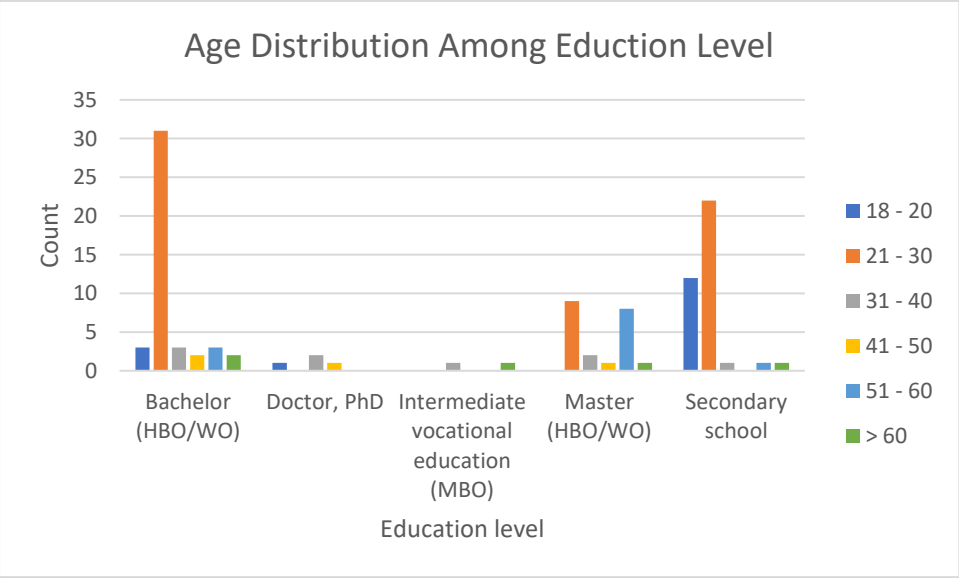

Figure 6. Bar chart showing count of ages per education level.

3.2.3. Survey Contents and Flow

Before participants execute the main task, there are some other elements that are shown to them as seen in Figure 7. The contents of the survey can be found in Appendix A. The introduction focuses on the context of research, shortly explaining the goal while not mentioning kappa, and requires agreement to the informed consent to continue. The explanation of the actual task includes the meaning of visual separation and how to score the scatterplots. Then a control exercise is given to create an equal minimum global understanding among participants. The purpose is to lessen the influence of potential experience induced bias in which a difference in conceptual understanding and thus interpretation of certain visualizations causes different scorings. One disclaimer is given to ensure all participants know that there are some plots in which data points may overlap. Subsequently, a hidden background process decides what questions are presented to the participant for the main task. Next, the main task questions are presented. The wrap up contains some demographical questions including experience, as well as unrequired questions on their physical conditions and open text questions on the survey.

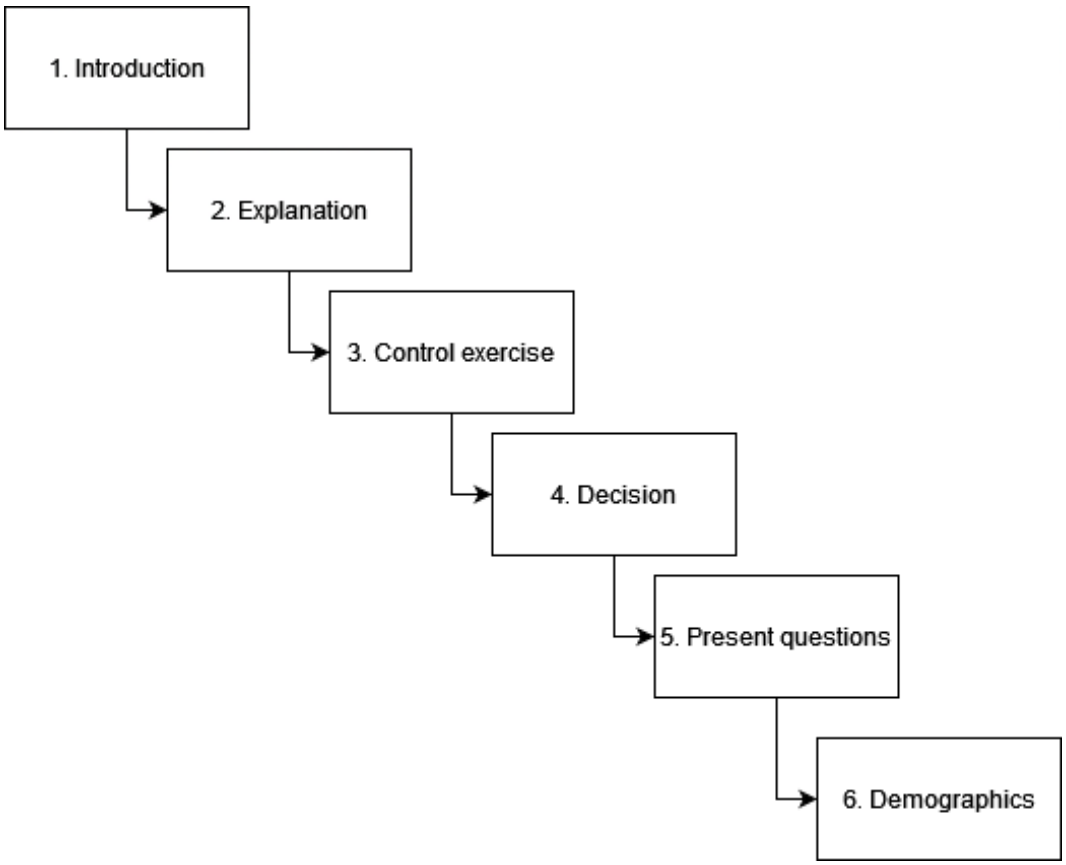

Figure 7. Schematic overview of the survey flow.

3.2.4. Survey Randomization

The distribution of questions is done through Qualtrics logics. The main process is executed in the decision part which is hidden for participants. The randomization is done according to the grouped kappa values so that questions belonging to a kappa ground are uniformly distributed among participants, therefore creating an uniform kappa distribution. A schematic overview of this process is shown in Figure 8. A question counter ensure equal display of questions within a kappa group.

The question order is also randomized in the present question part. This implies that a participant sees all 27 questions, of which 3 questions for each rond kappa, in random order to prevent learning and order bias. The effect of learning bias is not fully mitigatable. However, it is negligible in the overall analysis due to randomized order of questions and amount of questions distributed among participants.

Showing questions from the whole kappa range ensures prevention of perception bias, in the sense that participants observe the whole spectrum of 'bad' and 'good' images to be able to slightly learn and compare, rather than having no accurate comparison material. The control exercises have also participated in giving an indication of present material, but are made from toy datasets and thus not fully representative of the actual scatterplot images. In general, the learning process is faster with a varying order of images, therefore ensuring more representative responses

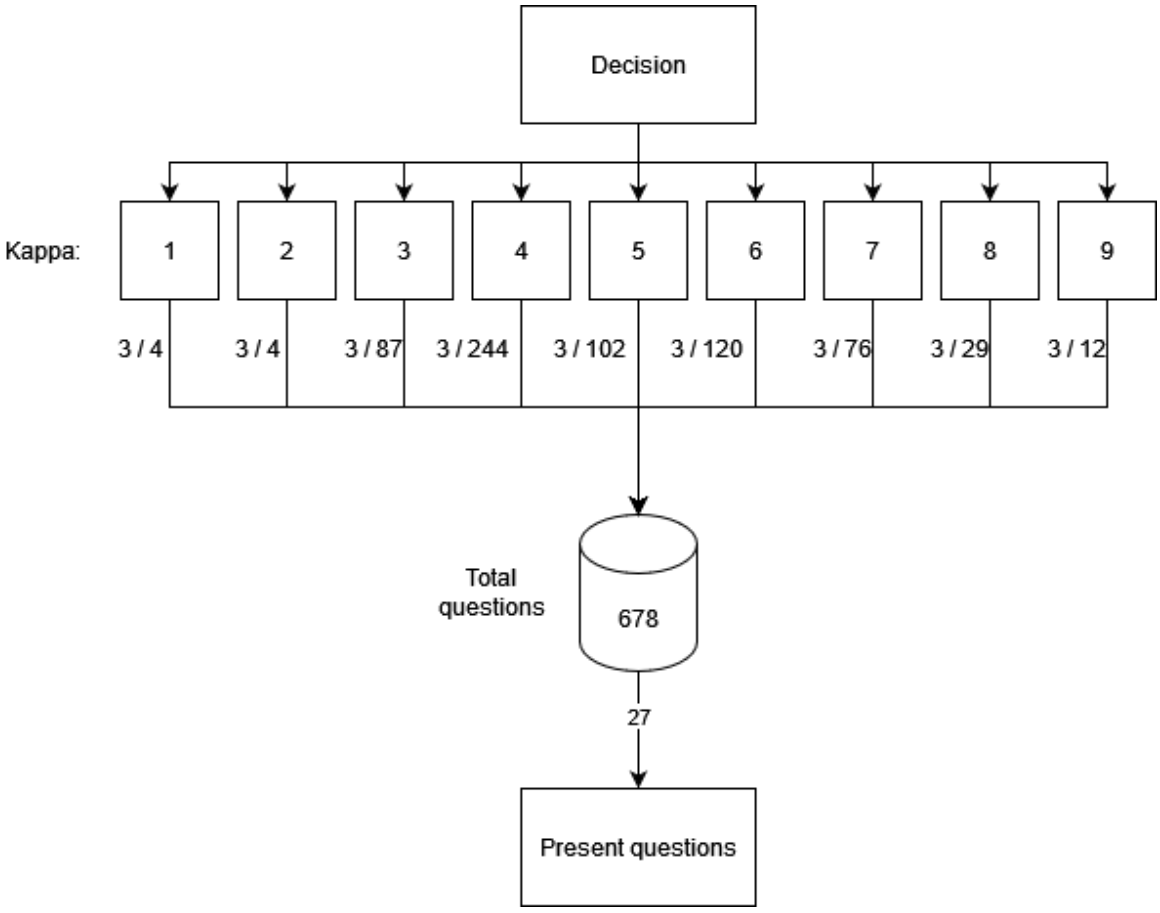

Figure 8. Schematic overview of the randomization flow.

3.2.5. Survey Details and Looks

The explanation and main task are kept simple, yet accurate, to ensure that the survey is accessible for large variety of participants. Furthermore, the simplicity ensures that participants will understand and therefore have a least error prone and valid answered survey. An example of a main task question on a mobile phone is found in Figure 9.

Certain design choices were made for the survey. The main choices are:

- Including a progression bar, not being able to turn back and showing 1 question on each page to prevent direct comparing/non intuitive answers as well as updating the progress bar.
- Including a horizontal slider bars instead of radio buttons because these show horizontally on a large device as well as a mobile device (limitation of Qualtrics).
- Including a Likert 5 scale because more options was deemed too hard to distinguish meaning for each Likert point. The scale ranges from -2 to 2 with labels “extremely bad, bad, neutral, good, extremely good” to indicate a negative and positive opinion but no definition of what bad or good is because this is up to the participants to decide.
- Allowing participants to finish later.
- The UU theme to associate with a professional setting.

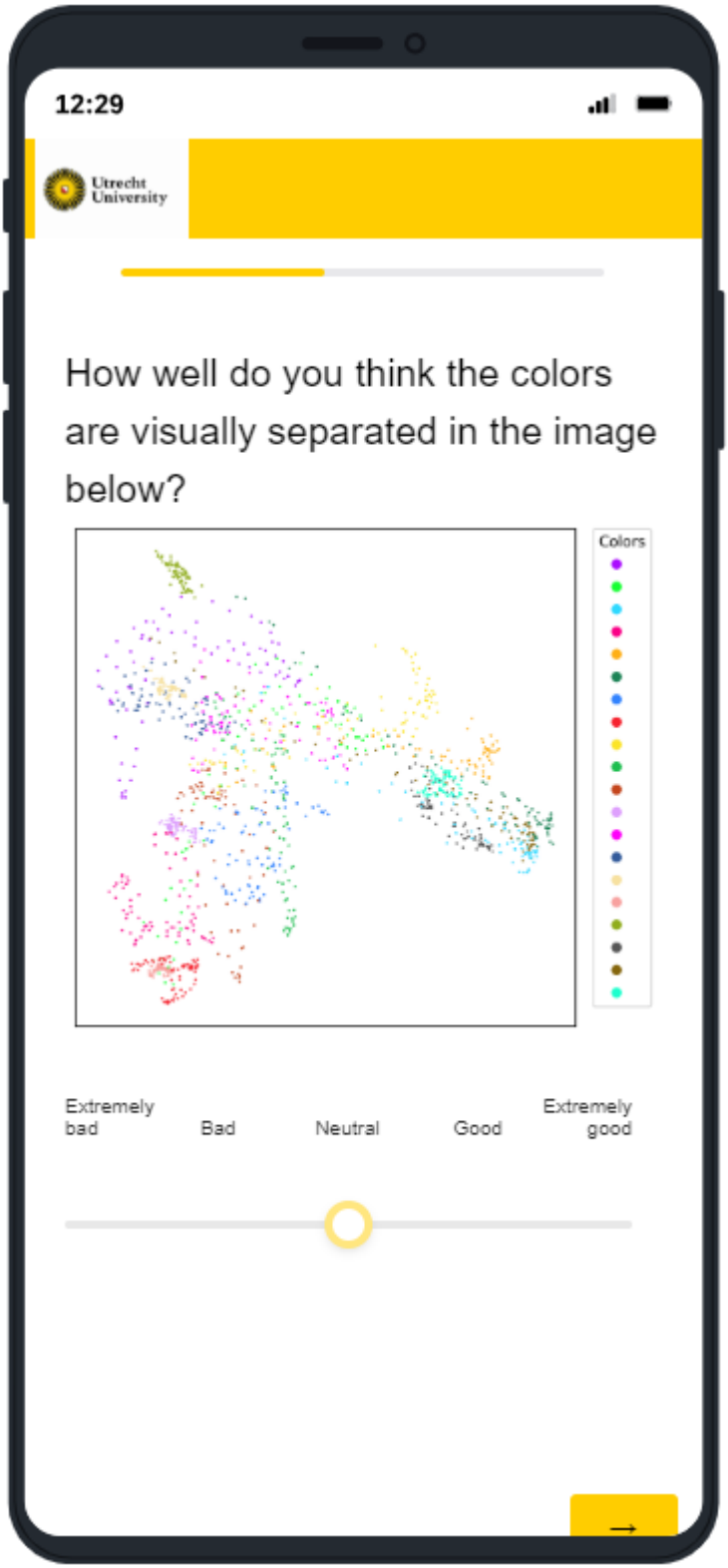

Figure 9. Survey question page on a mobile phone.

3.3. Data Collection, Reformatting, and Analysis

Initial collection of the responses is done via the Qualtrics platform. Multiple downloadable formats are available. Before analysis, uncompleted responses are removed. No outliers are found thus no other entries are removed. The response data is merged with data of the images resulting in a format in which each answer is an entry. This format allows for correlating the score with kappa. Analysis is done in Excel. Descriptive analysis is mostly done using pivot tables.

To address the research question, kappa and human score are correlated with a Pearson Correlation. Each images' kappa is measured against the mean of the human scores for that image, examining the presence of a linear relationship. A Pearson is preferred above Spearman due to the expectation of having a proportional relationship. A Spearman Correlation provides a nonparametric measure of rank correlation purely indicating dependence between rankings of variables. On the other hand, Pearson provides an indication of the strength and direction of the relationship between the variables.

4. Results

A full structure of dataset with relevant variables of which most are derived from the method is found in Appendix B. Questions have the name in the format “dataset_projection”. In addition to the answered scores, an average score is calculated for each unique question. This average score is scaled on a range 0 to 1 in another column, which allows comparison with kappa.

4.1. Statistical Analysis

4.1.1. Descriptive Statistical Analysis

The sample description is found in Section 3.2.2.

The material used existed of 678 unique questions, each with a dataset, projection, and kappa, of which 16 questions from kappa 4 and 5 are not presented due to uncompleted responses. A total of 2913 questions are answered. The corresponding question distribution is found in Figure 10 and subsequently the resulting kappa distribution is shown in Figure 11.

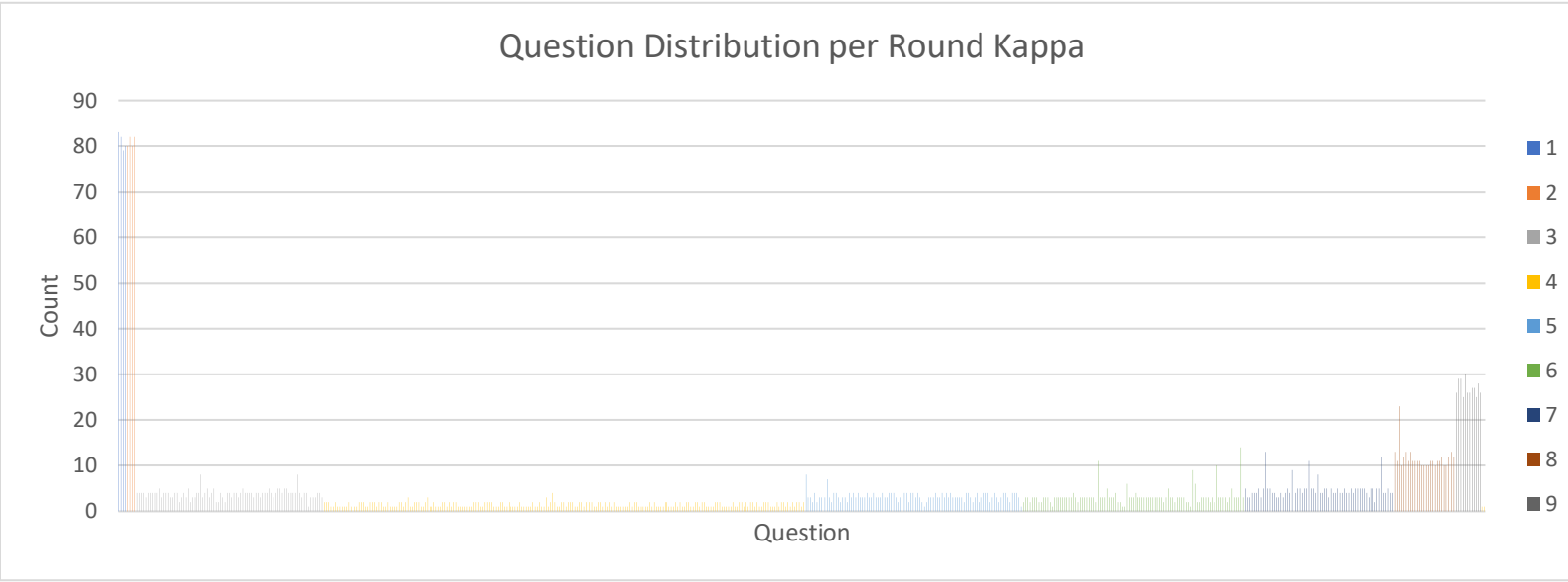

Figure 10. Bar chart showing the count of question appearance in the survey by its round kappa value.

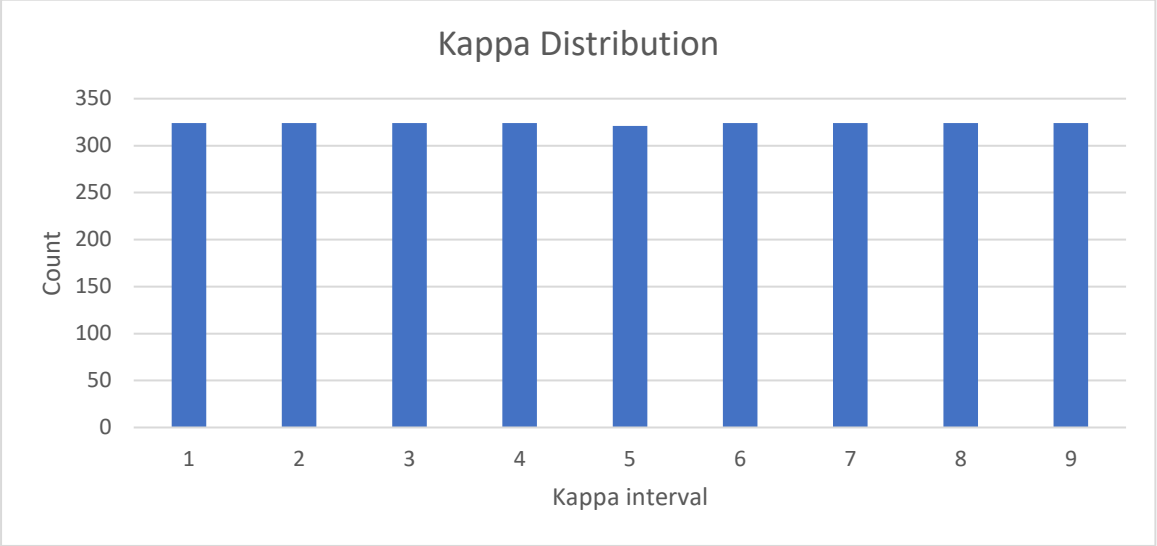

Figure 11. Bar chart showing equal distribution among kappa appearance in the survey.

The distribution of different datasets and projections demonstrated a relatively even spread, as depicted in Figure 12 for datasets and Figure 13 for projections respectively.

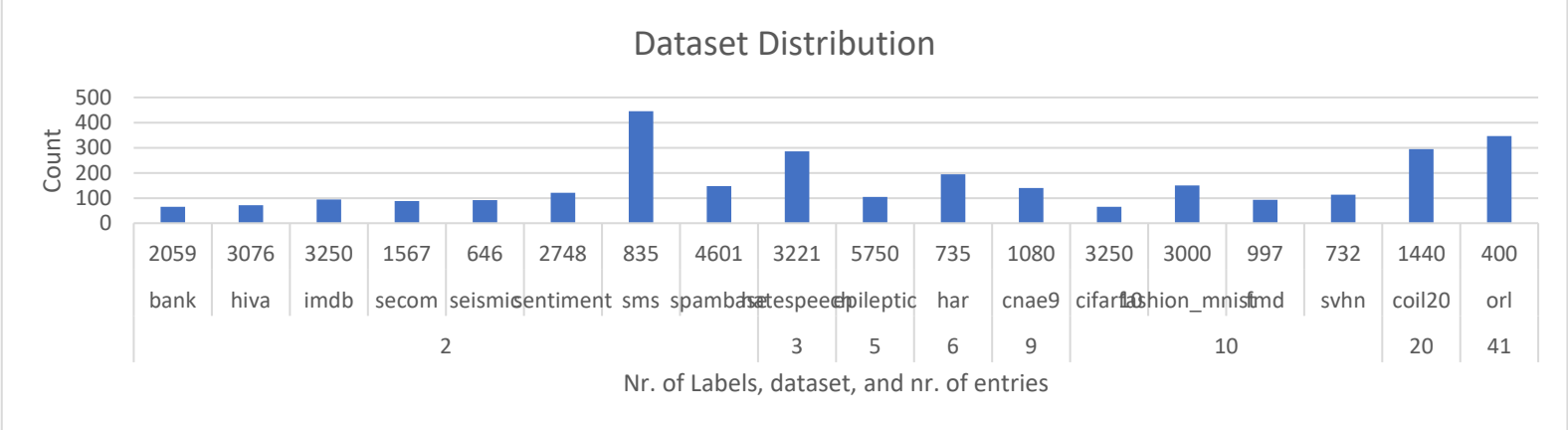

Figure 12. Bar chart presenting the distribution of the used datasets, with the corresponding number of labels and entries.

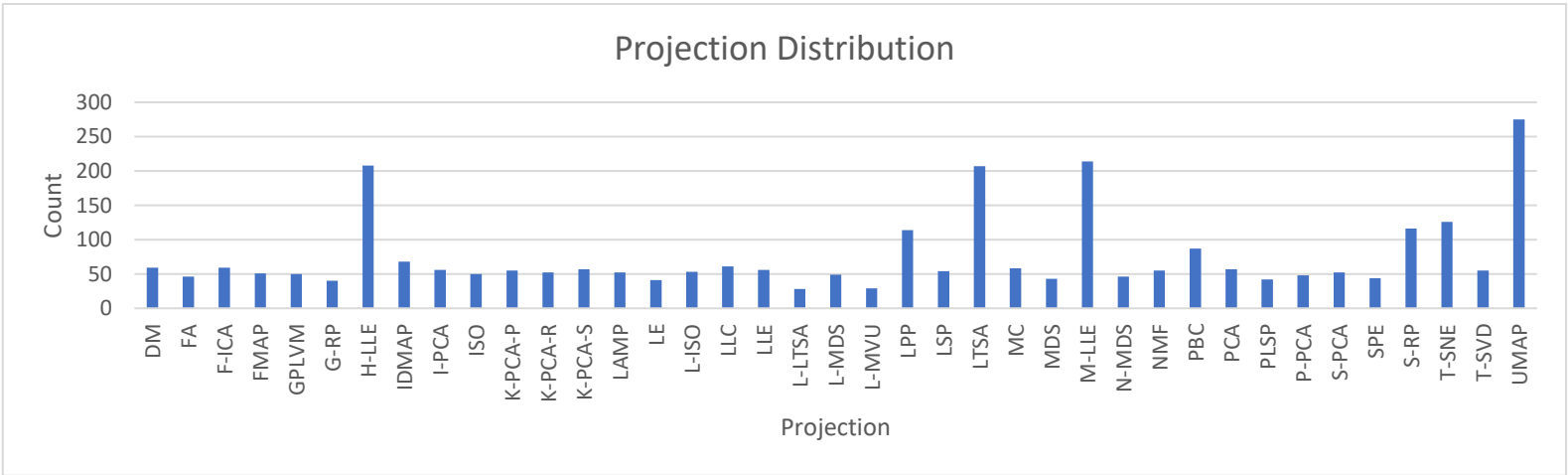

Figure 13. Bar chart presenting the distribution of used projections.

The total count of presented outlier categories showed a skew towards lower kappa values as seen in Figure 14. This is in line with the question distribution, meaning the 4 and 4 images in round kappa 1 and 2 both proportionally contained a larger of cases of outlier category compared to the other round kappa's.

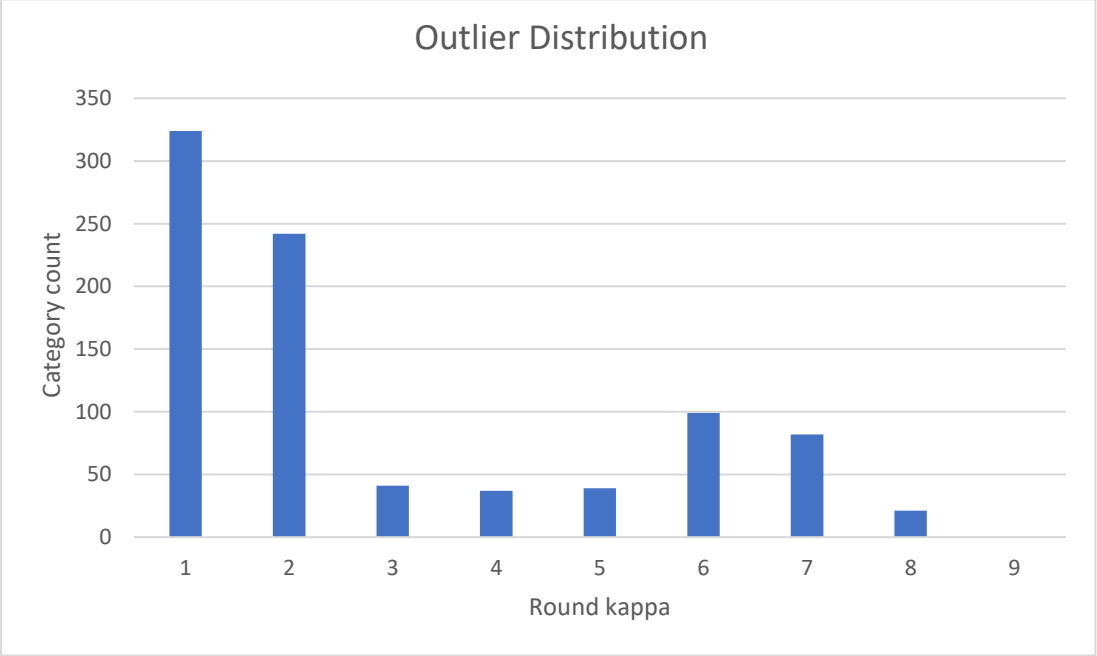

Figure 14. Bar chart presenting count of category cases shown in the survey per round kappa.

The scores in round kappa 1 and 2 showed differences in values compared to other round kappa's, considering whether the outlier category applies or not. Figure 15 displays this difference. The outlier categories, represented by "Yes", score higher than kappa in round kappa 1 and 2. For other round kappa's the opposite was true, disregarding of the category outlier value "Yes" or "No", the score was lower than kappa. In addition, the overall average of scores in round kappa 1 and 2 is higher, as will become clearer later.

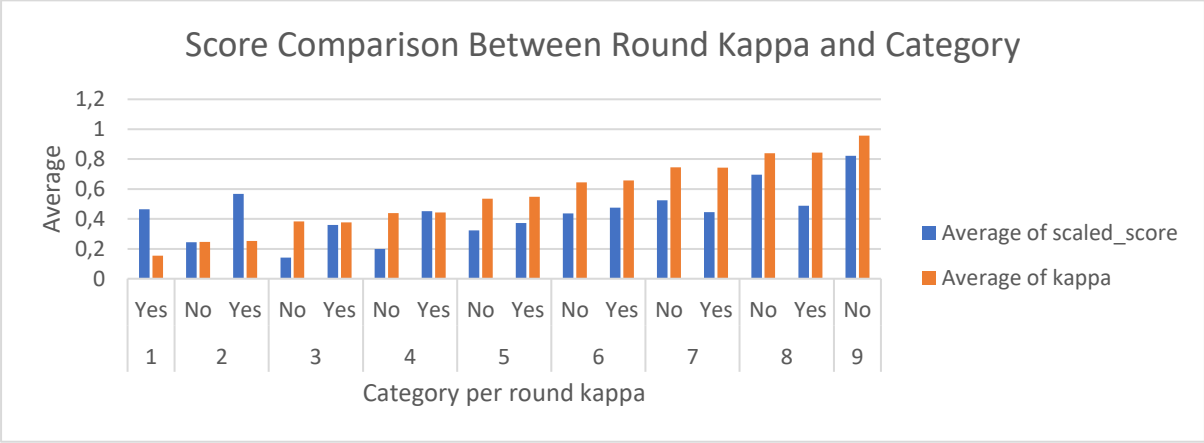

Figure 15. Bar chart showing the score relative to the kappa based on category per round kappa.

4.1.2. Inferential Statistical Analysis

First of all, possible influences of demographical variables on scoring with relation to kappa were analyzed. Noteworthy was that differences in education for MBO (n=2) were not significant to make observations. Experience in data analytics or data visualization did not show considerable difference in trends in the average picture, however the total scores given were more negative the more experience with exception respondents with experience for 4 years, see Figure 16. Gender, age and device did not show any noteworthy differences between scoring habits.

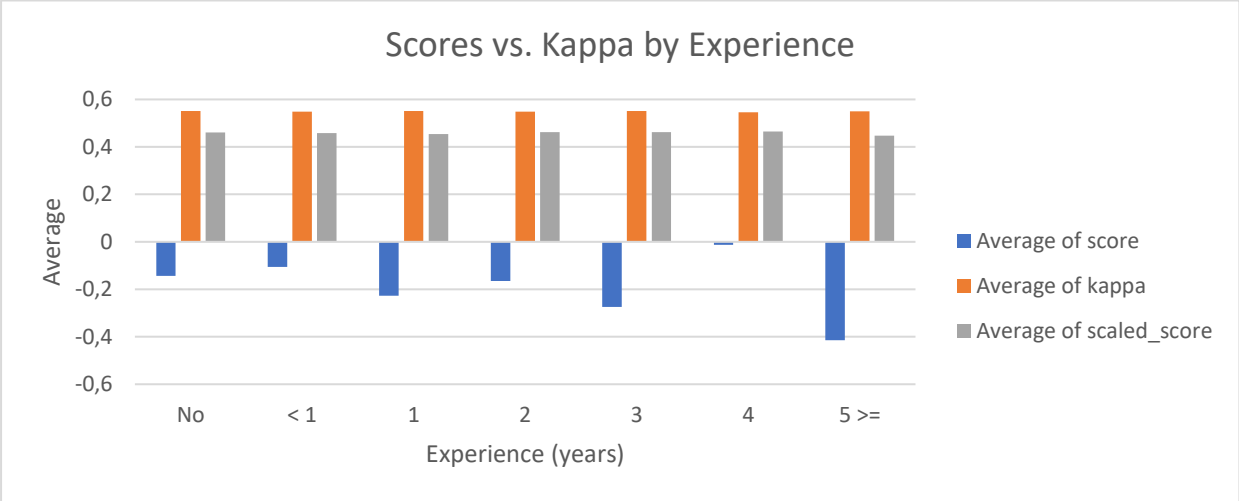

Figure 16. Bar chart showing difference between scores and kappa value based on experience.

Secondly, possible influences of data variables on scoring with relation to kappa were analyzed. Figure 17 shows the score vs kappa per dataset, for which the number of datapoints and number of labels are also shown. Dataset svhn and cifar10 show a very low average of scores.

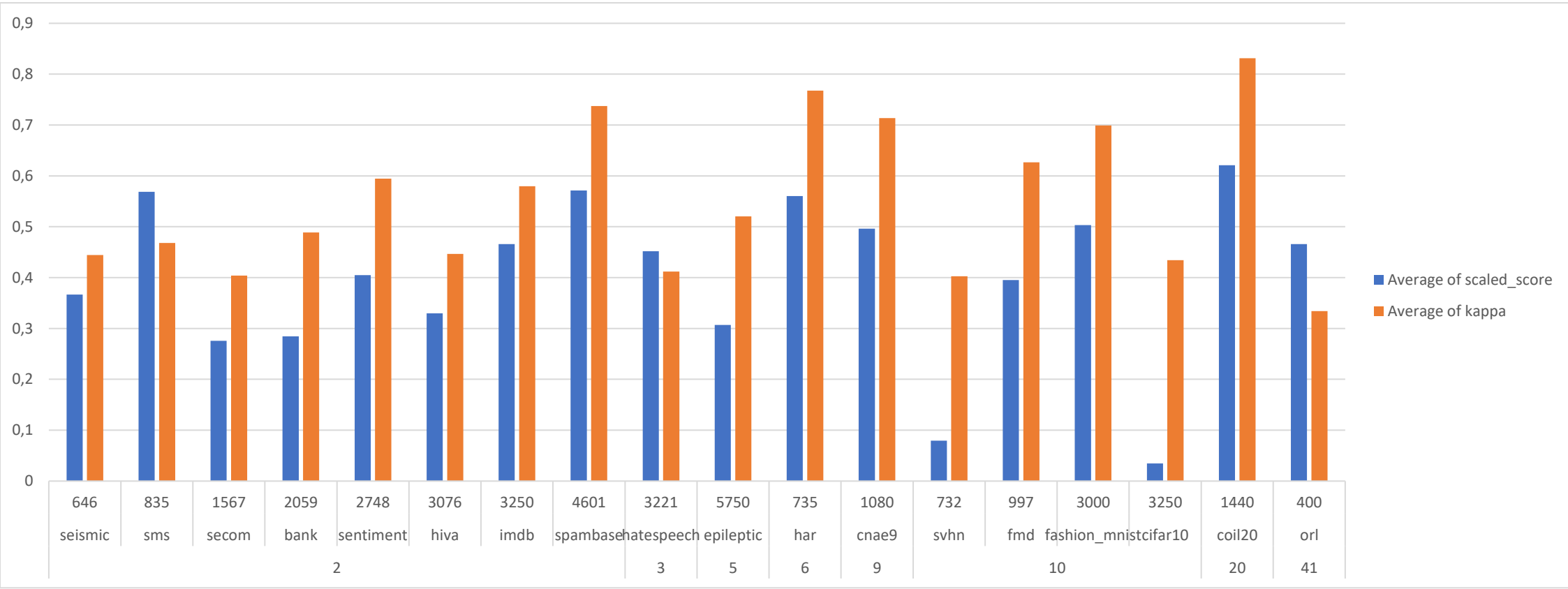

Figure 17. Bar chart showing each dataset with corresponding number of labels (below) and number of datapoints (above) showing difference in scores and kappa values.

To answer the research question, a correlation between kappa and given scores was analyzed. Observations from Figure 18 show a rising line in averaged scores per question in relation with a higher value of round kappa. Likewise, Figure 19 illustrates a trend in which the distribution of given scores shifts towards higher values as the rounded kappa increase. An exception for this trend were scores in round kappa 1 and 2.

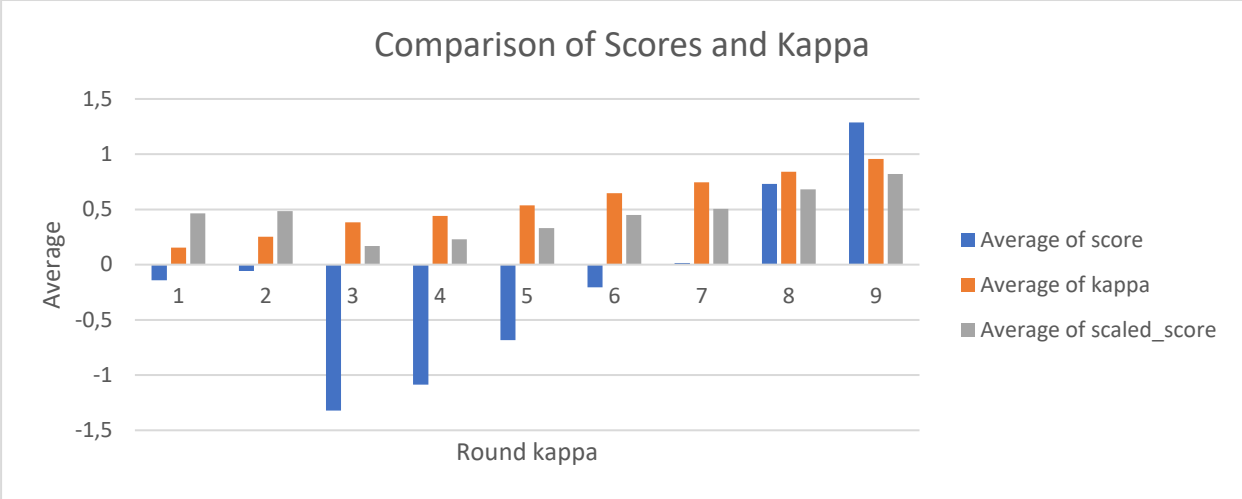

Figure 18. Bar chart showing difference between scores and kappa within each round kappa value.

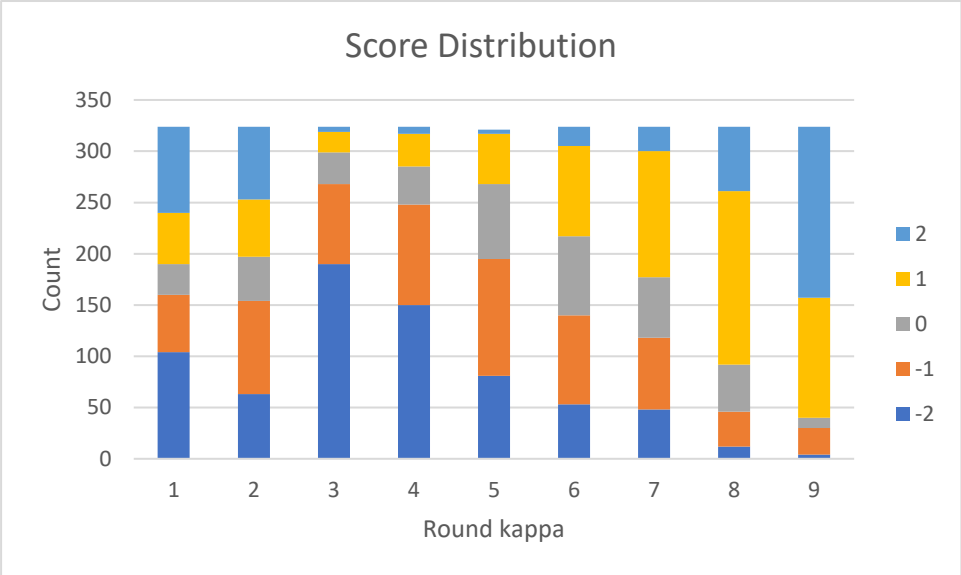

Figure 19. Bar chart showing the count of different answers given within a round kappa.

The comparison of individual scores by kappa is found in Figure 20 and its spread in Figure 21. Note that these figures are based on the actual kappa value with the range of 0 to 1. The figures show that for a higher kappa, more higher scores are given. The appearance of score value 2 exhibits a greater variability compared to the other scores.

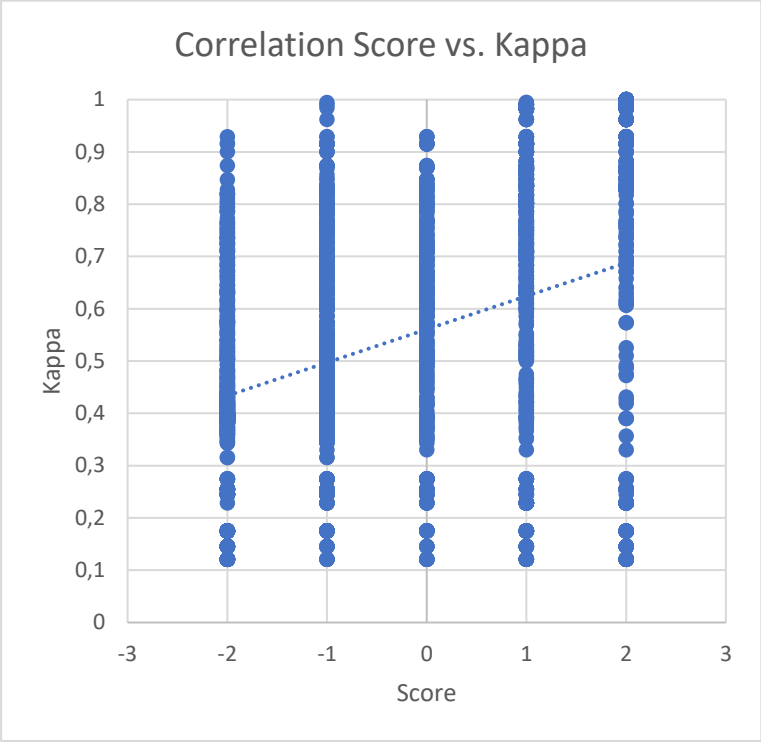

Figure 20. Scatterplot showing an upward trendline between score and kappa.

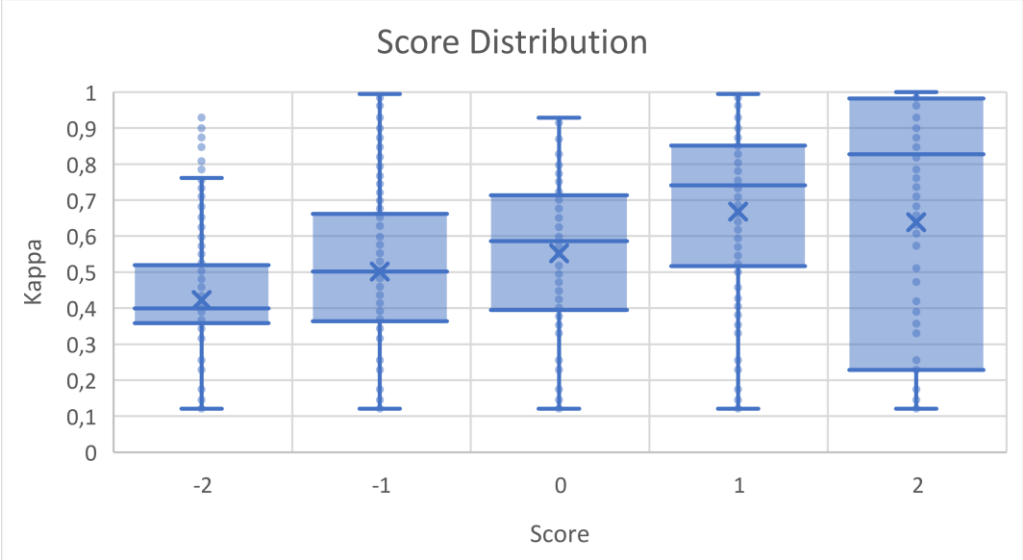

Figure 21. Boxplots visualizing the distribution of the given scores among the kappa value.

The Pearson correlation coefficient for kappa and score was .352, which was significant (p < .001 for a one-tailed test), based on 2913 complete observations, see Table 7.

Table 7. Pearson correlation between kappa and score.

		kappa	score
kappa	Pearson Correlation	1	.352**
	Sig. (1-tailed)		<,001
	N	2913	2913
score	Pearson Correlation	.352**	1
	Sig. (1-tailed)	<,001	
	N	2913	2913

**. Correlation is significant at the 0.01 level (1-tailed).

If kappa 1 and 2 were taken out of the analysis, the Pearson correlation coefficient for kappa and score was .636, which was significant (p < .001 for a one-tailed test), based on 2265 complete observations, see Table 8.

Table 8. Pearson correlation between kappa and score, excluding round kappa 1 and 2.

		kappa	score
kappa	Pearson Correlation	1	.636**
	Sig. (1-tailed)		<,001
	N	2265	2265
score	Pearson Correlation	.636**	1
	Sig. (1-tailed)	<,001	
	N	2265	2265

**. Correlation is significant at the 0.01 level (1-tailed).

4.2. Qualitative Analysis

From open question answers it seemed a lot of participants had difficulties when filling in the questions, disregarding the included explanations. Listed below are the main summarized difficulties stated by participants that lead to confusion or uncertainties.

- Overlapping points: lead to hidden colors that were visible in legend but not in plot, or datapoint positioning as if there is no visual material. Some participants did not grasp this concept while other did (halfway through).
- Out of proportion colors in the visualization: it was harder to decide on an answer if one color was abundantly present.
- High amount of colors in the visualization: due to more cognitive load some found it harder to decide on an answer.
- Distinguishing colors: especially light ones and certain combinations, also due to white background.
- Small dots: hard to observe. Zooming in or inability to distinguish dots / colors.
- Distinguishing answers: some found it hard to distinguish what the exact meaning of each parameters is. Specifically, some participants found it hard to decide between mid-section. Someone found 3 options sufficient. Another participant was conscious about learning through the process because of comparability to other images.
- Control exercise: caused confusion to some, and thus the feeling of uncertainty and pressure during the actual questions.
- Slider: was stuck on Neutral, someone thought it was not allowed.
- Conceptual understanding visual separation: some found it hard to think in terms of colors rather than positioning of clusters.

Some participants mentioned physical constraints listed below, that might have influenced the answers. The amount is negligible in the total result.

- Dyslexia (no applicable, only to explanation interpretation) (n = 1)
- Monochromacy (n= 1)
- Red green colorblindness (n = 3)
- ADHD, taking Methylphenidate (n = 1)
- Tired eyes (n= 1): a lot of staring, however it is not mentioned that 27 questions was too much

Other comments involved the following mentions listed below.

- Some images looked strange, some participants seem to have lacking conceptual understanding of data visualization or did understand the task.
- There are 2 mentions of using a screen mode (dark and night).

Concluding, some pitfalls were observed by participants which state ground for the discussion. Nevertheless the survey has been considered good by participants.

5. Discussion

To interpret the findings from Section 4. we first look at the population sample. An underrepresentation of MBO students and overrepresentation of young Bachelor (HBO/WO) might have potential influence on the results. Additionally, the survey has been updated during distribution process due to forgetting the MBO option, thus the exact amount of education levels might deviate.

There are negligible cases of self-reported color-blindness (n=3). However, Koponen & Hildén (2019) mention that many are not self-aware of reduced or deviating color vision. This is unpreventable and negligible considering the estimated number of colorblind cases.

Looking at the validity of tested question distribution we see an overrepresentation of the questions in round kappa 1, 2, and slightly in 8 and 9, due to lack in varying material. It is highly probably this lack in material is due to either the behavior of projections on the datasets, or to the behavior of kappa on the projections.

The outlier categories are manually defined and images are categorized manually. The findings show potential difference due to categories in round kappa 1 and 2 only. This means we cannot generalize the statement of unrepresentative images to the outlier categories. We state that round kappa 1 and 2 are not representative to score human perception of visual separation due to outliers in the data.

The amount of questions in round kappa 1 and 2 are high proportionally considered outliers. This, in combination with the randomization that compensates for lack in evenly distributed material over kappa, causes overrepresentation in kappa 1 and 2 to have significant influence on the correlation coefficient.

Additionally, a side effect of leaving outlier categories in the survey is possible confusion. The same risk holds for including example questions. Participants clearly stated their confusion and feeling of uncertainty. A different kind of explanation and example question from the real data might be more suitable as not all participants conceptually understood visual separation, nor did they read or understand the comment about overlapping datapoints. Additionally, control exercised are not guaranteed to work. Example questions answer can be a good indication for conceptual understanding of participants. The conceptual understanding of participants could also be tested if data is available of given scores before and after control questions overtime. Concerning the explanation, including a A/B-test element where one group gets an explanation, and the other does not, can prove influences.

Other factors that potentially influence the correlation are number of data points and number of labels. There seems to be no significant influence from findings. However, datasets svhn and cifar10 show drastically lower scores. It is unclear why these datasets seem to deviate from others.

Possible influence of factors that are uncontrollable and unmeasurable, can be controlled, measured, and analyzed in future research. Preferred is for each participant to use a similar setup for the survey rather than each individuals own customized setup including differences

in used device, screen size, brightness, potentially dark or night mode, etc. Furthermore the color palette is hard to distinguish. It is currently applied statically on the labels to provide consistency over the images and participants. An option is to dynamically apply a (customizable) color palette so that colors with good contrast are placed close as possible, since colors next to each other influence the perception as they are not absolute (Koponen & Hildén, 2019). Adding other visual cues than color might allow for easier distinction between label because up to 12 colors are distinctive as mentioned in Section 2.1.2., even though this brings risk of increasing cognitive load.

Qualitative analysis based on participants' comments gave insight in clear confusion about certain visualizations not showing all colors or showing little amount of data points. A potential solution for being able to distinguish the outlier categories with stacked points is using smaller marker size, the zoom function and transparent datapoints. More stacked datapoints means less transparency. It might be possible to distinguish different stacked colors with this functionality.

6. Conclusion

We have looked for a relationship between kappa and human perception of visual separation through a user study. Recalling the concluding findings from Section 4.1.2. shortly, kappa and score had a statistically significant positive linear relationship ($r=.352$, $p < .001$) for which the strength is approximately moderate ($.3 < |r| < .5$), meaning that the variables tend to increase together (i.e. higher scores are associated with a higher kappa). Excluding round kappa 1 and 2, kappa and score have a statistically significant positive linear relationship ($r=.636$, $p < .001$) for which the strength is approximately strong ($|r| > .5$). With this significant finding, disregarding leaving in round kappa 1 and 2, we may reject the null hypothesis and accept the alternative hypothesis, saying there is a linear correlation between metric kappa and human perception of visual separation. This allows for using kappa in a practical workflow. If a projection shows good visual separation according to kappa and other metrics, it is suitable for visual exploration for users and may be used for further assessment.

This leads to ideas for future research. A user study with more controlled experimental setup can assess the practical usability of kappa in a workflow. Interestingly it is unknown what setup works best for assessing visual separation. An experimental setup allows for controlling, measuring and thus testing customizable features such as color randomization, control over plot sizing, hovering for selection and zooming etc. to provide customizable for best personal perception. Moreover, further controlled material can be introduced looking at characteristics of the datasets and projections. This possibly includes the positioning of datapoints and its groups. Their influence on the perception of visual separation is uncertain.

Finally, it is worthwhile to conduct additional analysis on the correlation between scores and other metrics. By comparing the scores with other measured metric values, we can observe any differences in correlation compared to the correlation of kappa with visual separation. These comparisons may provide insights that align with previous theories on quality metrics or are consistent with the behavior of kappa. Moreover, it is important to note that alternative types of statistical analysis beyond linear correlation can also be applied.

References

Arnkil, H. (2021). *Colours in the Visual World*. Aalto ARTS Books.

Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, *59*, 44–58. https://doi.org/10.1016/j.inffus.2020.01.005

Benato, B. C., Falcão, A. X., & Telea, A.-C. (2023). *Linking data separation, visual separation, and classifier performance using pseudo-labeling by contrastive learning*. http://arxiv.org/abs/2302.02663

Benato, B. C., Gomes, J. F., Telea, A. C., & Falcão, A. X. (2021). Semi-supervised Deep Learning Based on Label Propagation in a 2D Embedded Space. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 25th Iberoamerican Congress, CIARP 2021, Porto, Portugal, May 10–13, 2021, Revised Selected Papers*, *12702*, 371–381. https://link.springer.com/bookseries/7412

Bertini, E., Tatu, A., & Keim, D. (2010). *Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization*. http://scholar.google.com/

Carreira-Perpiñán, M. A. (1997). *A Review of Dimension Reduction Techniques **.

Coombes, K. R., Brock, G., Abrams, Z. B., & Abruzzo, L. V. (2019). Polychrome: Creating and assessing qualitative palettes with many colors. *Journal of Statistical Software*, *90*. https://doi.org/10.18637/jss.v090.c01

Espadoto, M., Martins, R. M., Kerren, A., Hirata, N. S. T., & Telea, A. C. (2019). Towards a Quantitative Survey of Dimension Reduction Techniques. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, *2019*(27:3), 2153–2173.

Goodman, A. A. (2012). Principles of high-dimensional data visualization in astronomy. *Astronomische Nachrichten*, *333*(5–6), 505–514. https://doi.org/10.1002/ASNA.201211705

Koponen, J., & Hildén, J. (2019). *Data Visualization Handbook* (P. Alapeteri & L. Koivunen-Niemi, Eds.; 1st ed.). Aalto ARTS Books.

Musa, S. M., Akujuobi, C., Sadiku, M. N. O., Shadare, A. E., Akujuobi, C. M., & Perry, R. G. (2016). DATA VISUALIZATION. *International Journal of Engineering Research And Advanced Technology*. https://www.researchgate.net/publication/311597028

Rau, R., Bohk-Ewald, C., Muszyńska, M. M., & Vaupel, J. W. (2017). *Visualizing Mortality Dynamics in the Lexis Diagram* (Vol. 44). https://doi.org/https://doi.org/10.1007/978-3-319-64820-0_1

Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review*, *54*(5), 3473–3515. https://doi.org/10.1007/s10462-020-09928-0

Theus, M. (2008). *Handbook of Data Visualization*. https://doi.org/https://doi.org/10.1007/978-3-540-33037-0_7

Thuraisingham, B. (2000). *A Primer for Understanding and Applying Data Mining D ata mining can be a powerful tool for extracting useful information from tons of data WHAT IS DATA MINING?*

Van Der Maaten, L., Postma, E., & Van Den Herik, J. (2009). *Dimensionality Reduction: A Comparative Review Dimensionality Reduction: A Comparative Review*. http://www.uvt.nl/ticc

Ware, C. (2019). *Information Visualization: Perception for Design* (4th ed.).

Appendices

Appendix A. Survey Contents

Below the survey contents are shown. They are shortened due to excessive amounts of answers and display logics. Images are left out.

Survey Flow

Standard: Introduction (2 Questions)
Standard: Explanation (1 Question)
Standard: Control exercise (10 Questions)
Block: Decision (9 Questions)
Block: Present questions (678 Questions)
Standard: Demographics (9 Questions)

Page Break

Start of Block: Introduction

Introduction Dear participant,

Thank you for your interest in participating in this survey on assessing the effectiveness of a visual separation metric with human perception. This study is conducted by Carlijne Govers, a Bachelor Information Science student from Utrecht University, for the Bachelor Thesis. In short, this study intends to collect scores from participants on tasks that are related to rating the visual separation of labels in scatterplots. No prior knowledge of data visualization is required, as everything will be explained.

The survey consists of an explanation of the topic and task, followed by 27 images to score, and concluding questions. The survey is expected to take approximately 10 minutes to complete. It is recommended to take the survey on a large-screen device, but a phone will suffice.

If you have questions, comments or concerns about this research project, please contact this email address: c.n.govers@students.uu.nl.

By participating in this survey, you consent to the following:
-Participation is voluntary. If you agree to participate, you may withdraw at any moment without having to justify why.
-The data is collected anonymously and can not be linked to you as participant.
-The collected data from participation in this survey is used for this research as only purpose.
-The collected data is safely stored for at least during the course of this research.

Informed consent I consent to participate in the research study as described above.

- ○ Yes, I consent to participate in the research study. (1)
- ○ No, I do not consent to participate in the research study. (2)

Skip To: End of Survey If I consent to participate in the research study as described above. = No, I do not consent to participate in the research study.

End of Block: Introduction

Start of Block: Explanation

Instructions Goal of the study

Scatterplots are used to visualize large amounts of data points. Below you see such a scatterplot in which we have two types (classes) of points, each represented by one color. How well these colored groups are separated from each other is called visual separation. In this study, we want to measure the visual separation of several such scatterplots.

The first example given below shows excellent visual separation. There is no datapoint of purple or yellow that is mixed within the group of the other color. On a scale from extremely bad to extremely good, this would score extremely good.

The second example given below shows terrible visual separation. The datapoints of the two colors are fully mixed. On a scale from extremely bad to extremely good, this would score extremely bad.

Study structure

In the following, we will ask you to score 27 scatterplot images on their visual separation. Even though there is no time limit for the study, we recommend that you only swiftly look at each image for a few seconds to rank its visual separation. Note also that it is not possible to revisit an already scored image.

End of Block: Explanation

Start of Block: Control exercise

Q29 Practice questions

We start with 4 practice questions to show you how to score scatterplot images on their visual separation.

A legenda on the side of the image will show which colors are present. Furthermore, it is possible to zoom-in on a touch-screen to get a better view.

JS

QEx1 Question

How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

JS

AEx1 Answer

The visual separation in this example is "extremely good". This is because the colors are well separated from each other.

JS

QEx2 Question

How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

JS

AEx2 Answer

The visual separation in this example is "good". This is because even though the colors are very mixed in some small parts, most parts of the colors are well separated from each other.

JS

QEx3 Question

How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

JS

AEx3 Answer

The visual separation in this example is "extremely bad". This is because the colors are not separated from each other anywhere.

JS

QEx4 Question

How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

JS

AEx4 Answer

The visual separation in this more representative example is harder to judge. It would probably score "neutral". This is because the colors are not separated well in the center, while there is some separation on the outer edges.

Q30 These were the 4 practice questions. The same task continues for the 27 images that follow now.

Note that some of the following images contain datapoints that share the same position. This overlap unables you to visually see all datapoints and colors even though they are present.

End of Block: Control exercise

Start of Block: Decision

JS

Q20 kappa 1

- Click to write Choice 1 (1)
- Click to write Choice 2 (2)
- Click to write Choice 3 (3)
- Click to write Choice 4 (11)

Q27 kappa 2

- Click to write Choice 1 (1)
- Click to write Choice 2 (2)
- Click to write Choice 3 (3)
- Click to write Choice 4 (4)

Q33 kappa 3

- ○ Click to write Choice 1 (1)
- ○ Click to write Choice 2 (2)
- ○ Click to write Choice 3 (3)
- ○ Click to write Choice 4 (4)
- ○ Click to write Choice 5 (5)
- ○ Click to write Choice 6 (1000)
- ○ Click to write Choice 7 (1001)
- ○ Click to write Choice 8 (1002)
- ○ Click to write Choice 9 (1003)
- ○ Click to write Choice 10 (1004)
- ○ Click to write Choice 11 (1005)
- ○ Click to write Choice 12 (1006)
- ○ Click to write Choice 13 (1007)
- ○ Click to write Choice 14 (1008)
- ○ Click to write Choice 15 (1009)
- ○ Click to write Choice 16 (1010)
- ○ Click to write Choice 17 (1011)
- ○ Click to write Choice 18 (1012)
- ○ Click to write Choice 19 (1013)

- Click to write Choice 20 (1014)
- Click to write Choice 21 (1015)
- Click to write Choice 22 (1016)
- Click to write Choice 23 (1017)
- Click to write Choice 24 (1018)
- Click to write Choice 25 (1019)
- Click to write Choice 26 (1020)
- Click to write Choice 27 (1021)
- Click to write Choice 28 (1022)
- Click to write Choice 29 (1023)
- Click to write Choice 30 (1024)
- Click to write Choice 31 (1025)
- Click to write Choice 32 (1026)
- Click to write Choice 33 (1027)
- Click to write Choice 34 (1028)
- Click to write Choice 35 (1029)
- Click to write Choice 36 (1030)
- Click to write Choice 37 (1031)
- Click to write Choice 38 (1032)
- Click to write Choice 39 (1033)
- Click to write Choice 40 (1034)

- Click to write Choice 41 (1035)
- Click to write Choice 42 (1036)
- Click to write Choice 43 (1037)
- Click to write Choice 44 (1038)
- Click to write Choice 45 (1039)
- Click to write Choice 46 (1040)
- Click to write Choice 47 (1041)
- Click to write Choice 48 (1042)
- Click to write Choice 49 (1043)
- Click to write Choice 50 (1044)
- Click to write Choice 51 (1045)
- Click to write Choice 52 (1046)
- Click to write Choice 53 (1047)
- Click to write Choice 54 (1048)
- Click to write Choice 55 (1049)
- Click to write Choice 56 (1050)
- Click to write Choice 57 (1051)
- Click to write Choice 58 (1052)
- Click to write Choice 59 (1053)
- Click to write Choice 60 (1054)
- Click to write Choice 61 (1055)

- Click to write Choice 62 (1056)
- Click to write Choice 63 (1057)
- Click to write Choice 64 (1058)
- Click to write Choice 65 (1059)
- Click to write Choice 66 (1060)
- Click to write Choice 67 (1061)
- Click to write Choice 68 (1062)
- Click to write Choice 69 (1063)
- Click to write Choice 70 (1064)
- Click to write Choice 71 (1065)
- Click to write Choice 72 (1066)
- Click to write Choice 73 (1067)
- Click to write Choice 74 (1068)
- Click to write Choice 75 (1069)
- Click to write Choice 76 (1070)
- Click to write Choice 77 (1071)
- Click to write Choice 78 (1072)
- Click to write Choice 79 (1073)
- Click to write Choice 80 (1074)
- Click to write Choice 81 (1075)
- Click to write Choice 82 (1076)

- ○ Click to write Choice 83 (1077)
- ○ Click to write Choice 84 (1078)
- ○ Click to write Choice 85 (1079)
- ○ Click to write Choice 86 (1080)
- ○ Click to write Choice 87 (1081)

JS

… (same for kappa 3 – 9)

End of Block: Decision

Start of Block: Present questions

Display This Question:

If kappa 1 , Click to write Choice 1 Is Displayed

JS

orl_LTSA How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

Display This Question:

If kappa 1 , Click to write Choice 2 Is Displayed

JS

orl_M-LLE How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

Display This Question:

If kappa 1 , Click to write Choice 3 Is Displayed

JS

sms_M-LLE How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

Display This Question:

If kappa 1 , Click to write Choice 4 Is Displayed

JS

sms_S-RP How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2

()	

Display This Question:

If kappa 2 , Click to write Choice 1 Is Displayed

JS

hatespeech_H-LLE How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

Display This Question:

If kappa 2 , Click to write Choice 2 Is Displayed

JS

hatespeech_LPP How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

Display This Question:

If kappa 2 , Click to write Choice 3 Is Displayed

JS

orl_H-LLE How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

Display This Question:

If kappa 2 , Click to write Choice 4 Is Displayed

JS

sms_LTSA How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

Display This Question:

If kappa 3 , Click to write Choice 1 Is Displayed

JS

bank_MC How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

Display This Question:

If kappa 3 , Click to write Choice 2 Is Displayed

JS

bank_N-MDS How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

Display This Question:

If kappa 3 , Click to write Choice 3 Is Displayed

JS

cifar10_H-LLE How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2

()	

Display This Question:

If kappa 3 , Click to write Choice 4 Is Displayed

JS

hatespeech_UMAP How well do you think the colors are visually separated in the image below?

	Extremely bad	Bad	Neutral	Good	Extremely good
	-2	-1	0	1	2
()					

… goes on until all questions of kappa 9 are shown.

End of Block: Present questions

Start of Block: Demographics

Q41 Thank you for completing all the scoring tasks. Lastly there is a set of questions below to finish the survey with.

Age What is your age?

- ○ < 18 (1)
- ○ 18 - 20 (2)
- ○ 21 - 30 (3)
- ○ 31 - 40 (4)
- ○ 41 - 50 (5)
- ○ 51 - 60 (6)
- ○ > 60 (7)

Gender What is your gender?

- ○ Male (1)
- ○ Female (2)
- ○ Non-binary / third gender (3)
- ○ Prefer not to say (4)

Education What is the highest level of school you have completed or the highest degree you have received?

- ○ No degree (1)
- ○ Primary school (2)
- ○ Secondary school (3)
- ○ Bachelor (HBO/WO) (4)
- ○ Master (HBO/WO) (5)
- ○ Doctor, PhD (6)
- ○ Intermediate vocational education (MBO) (7)

Q43 Do you have any knowledge / experience in the field of Data Analytics or Data Visualization? If so, for how many years?

- ○ No (1)
- ○ < 1 (2)
- ○ 1 (3)
- ○ 2 (4)
- ○ 3 (5)
- ○ 4 (6)
- ○ 5 >= (7)

Device On what device did you fill in this survey?

- ○ Mobile phone (1)
- ○ Tablet (2)
- ○ Laptop or desktop (3)
- ○ Other, please specify: (4) _____

Q46 Did you experience any difficulties with scoring the images? Feel free to elaborate.

_____

Q49 Feel free to mention any physical constraints (e.g. vision, color blindness), that may have influenced your ratings.

_____

Q47 Feel free to note any further comments on the survey.

_____

End of Block: Demographics

Appendix B. Dataset Format

Table #.

variable	data category	value
dataset	Nominal	[bank, cifar10, cnae9, coil20, epileptic, fashion_mnist, fmd, har, hatespeech, hiva, imdb, orl, secom, seismic, sentiment, sms, spambase, svhn]
projection	Nominal	[DM, UMAP, FA, F-ICA, GPLVM, G-RP, H-LLE, IDMAP, I-PCA, ISO, K-PCA-P, K-PCA-R, K-PCA-S, LAMP, LE, L-ISO, LLC, LLE, L-TSA, L-MDS, L-MVU, LPP, LSP, LTSA, MC, MDS, M-LLE, N-MDS, NMF, PBC, PCA, PLSP, P-PCA, S-PCA, SPE, S-RP, T-SNE, T-SVD, UMAP]
unique_labels	Ordinal	[2, 3, 5, 6, 9, 10, 20, 41]
num_entries	Ordinal	[400, 646, 732, 735, 835, 997, 1080, 1440, 1567, 2059, 2748, 3000, 3076, 3221, 3250, 4601, 5750]
q_number	Nominal	dataset_projection
kappa	Ratio	[0,120833 – 1]
round_kappa	Interval	[1 – 9]
score	Ordinal	[-2, -1, 0, 1, 2]
average_q_score	Interval	[-2 – 2]
scaled_average_q_score	Interval	[0 – 1]
category_single_stack	Nominal	[Yes, No]
category_line_stack	Nominal	[Yes, No]
category_hidden_color	Nominal	[Yes, No]
age	Ordinal	[18-20, 21-30, 31-40, 41-50, 51-60, 60 >]
gender	Nominal	[Female, Male, Non-binary / third gender, Prefer not to say]
education	Nominal	[Secondary school, Intermediate vocational education (MBO), Bachelor (HBO/WO), Master

		(HBO/WO), Doctor/PhD]
experience	Ordinal	[< 1, 1, 2, 3, 4, 5 >=]
device	Nominal	[Laptop or desktop, Mobile phone]
difficulties	-	String
physical_contraints	-	String
comments	-	String