Bridging Objective and Subjective Evaluations in Data Visualization: a Crossover Experiment

Angela Locoro*
angela.locoro@unibs.it
Università degli Studi di Brescia
Brescia, Italy

Silvia Corchs
Università degli Studi dell'Insubria
Varese, Italy
silvia.corchs@uninsubria.it

Paolo Buono
Università degli Studi di Bari Aldo Moro
Bari, Italy
paolo.buono@uniba.it

Paolo Bruscagin
Università degli Studi dell'Insubria
Varese, Italy
pbscagin@studenti.uninsubria.it

ABSTRACT

One of the problems affecting evaluation in the design and adoption of HCI technology is that neither objective nor subjective measures are sufficient when taken alone or individually. This paper proposes a crossover approach, making sense of objective and subjective evaluation methods by hypothesizing them as constitutive of each other's explanation. Objective image features borrowed from image processing may explain or being explained in terms of validated qualitative items for infographics value-in-use and qualitative labelling from users' interaction. These methods are all applied to the evaluation of a small set of Data Vizualizations (Data Viz from now on). Image features are computed first, in order to provide a varied-features Data Viz selection from researchers; the subjective part of the evaluation is accomplished by the 98 participants of an experiment, who interacted with pairs of Data Viz by executing a task, then using the validated items of the Infographics-Value (IGV) short scale, and adding free qualitative comments. Crossing over these dimensions shows that: a high *feature congestion* in a Data Viz can hinder its perceived *intuitiveness* and *clarity*; a poorly distributed *saliency* may impact *intuitiveness* and *clarity* too; a high *colorfulness* may influence the perceived *beauty*; both *saliency* and *colorfulness* may impact on the perceived *usefulness*, *informativity*, and *beauty*. Furthermore, *colorfulness* can improve or worsen the perceived overall *quality of design* and *quality of interaction* when used and combined with *feature congestion*; and *saliency* may improve or worsen the perceived *beauty* when interacting with *colorfulness*. These results show how objective and subjective evaluations may be exploited as each other's explanations for improving the evaluation process during both design and user experience with Data Viz. Based on this experiment, the importance of crossing-over quantitative and qualitative Data Viz evaluation is argued, and motivations to the exploitation of a combination of approaches instead of the application of one approach alone are supported. This contribution intends to lead towards a holistic Data Viz quality assessment method, able to provide a virtuous cycle enforcing both quantitative and qualitative approaches during all the phases of a Data Viz evaluation life.

CCS CONCEPTS

• **Human-centered computing → Visualization**; **Empirical studies in visualization**; **Visualization design and evaluation methods**;

KEYWORDS

Data Visualization, Visualization Metrics, IQA methods, IGV Short Scale, Objective and Subjective Measurement

ACM Reference Format:
Angela Locoro, Silvia Corchs, Paolo Buono, and Paolo Bruscagin. 2023. Bridging Objective and Subjective Evaluations in Data Visualization: a Crossover Experiment. In *15th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2023), September 20–22, 2023, Torino, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3605390.3605408

1 MOTIVATIONS AND BACKGROUND

The design choices of effective and efficient Data Viz is, most of the time, a matter of empirical knowledge left to the judgment of single authors, data communication experts, data scientists, or graphic designers [13]. This heterogeneity, lacking systematic background and specific expertise, raised many problems in the quality of Data Viz outcome due, for example, to the number of arbitrary features added for aesthetics rather than informative purposes [11]. Today, data are produced at an increasing pace, and this problem increases its relevance and cannot be overlooked anymore. Data Visualization should reduce the time to understand the data to anticipate the time to make decisions. One of the reasons for the success of Data Visualization is that it exploits the human visual perception system, which enables people to understand an image at a glance, provided that this task results as cognitively efficient and pleasant as possible (cf the standard notion of usability). Since each feature added to a Data Viz, even for decorative reasons only, vehiculates a perceptual signal (color, shape, orientation, and the like) that is processed by our cognitive system, the risk of useless or confusing information is a concrete one [24].

CHItaly 2023, September 20–22, 2023, Torino, Italy

ACM ISBN 979-8-4007-0806-0/23/09…$15.00
https://doi.org/10.1145/3605390.3605408

As said above, data exploration quality affects user experience (UX). Among the several methods measuring UX [23], a combination of quantitative and qualitative evaluations is deemed one of the more complete in giving a clear assessment of the UX [1]. However, to the best of our knowledge, whether and how to combine those methods has been the object of very few studies, in contrast with the importance of this matter [3, 4]. Moreover, besides UX, even fewer studies tried to address how authors and designers of Data Viz could benefit from this evaluation activity, which would be carried out well in advance concerning the time of users' test. Not many authors took all of these considerations seriously or provided a more organic perspective in the above direction (as counterexamples, see e.g., [6, 22]).

In the domain of image processing, this issue has reached a higher relevance instead, and Image Quality Assessment (IQA) methods were set up in combination. However, they are mainly focused on natural scene images [8]. IQA methods are both subjective and objective. IQA strongly encourages their synergy in order to provide quality metrics for images before or after publication or exploitation in users' tasks. Subjective methods are based on psycho-physical experiments where human observers assign quality scores to an image. The objective methods are computational models that aim to automatically evaluate image quality and have to be correlated with subjective scores. However, either subjective or objective methods are used differently than in user experience frameworks. For example, they are exploited on the same variable construct. Moreover, an image is static, and most of the time UX is not considered among the quality dimensions.

Inspired by both UX studies and the IQA field, we propose a crossover of objective and subjective evaluation approaches for the domain of Data Viz to be used either at design time or during interaction with users. In order to bridge the above gaps, objective and subjective methods are hypothesized as being in a reciprocal explanatory relationship, and their effects are combined. In this way, we should be able to provide a feedback loop solving their respective limitations and improving their combined effects on design.

Starting from the consideration that a Data Viz has a visual part besides a textual one, we claim that the visual part of a Data Viz is an image. It could be treated as such, especially when the visual and diagrammatic features that can be extracted from it should respond to minimum quality standards (see, e.g., [2]). A second characteristic of a Data Viz lies in its purpose: giving users value, while interacting with data. The value-in-use is one of the most specific subjective evaluations for Data Viz. For example, other constructs like usability, although usable as subjective evaluations, are very general purpose. Using a combination of approaches taken from image processing and value-in-use evaluation may serve both designers and users. It can provide objective evaluations of Data Viz features first, and their degree of influence on the subjective UX evaluations next, and vice-versa. Discovering how image features influence interaction may constitute the feedback returned to designers in terms of how their design choices should be affected by users and could improve their attitudes to Data Viz design (e.g., in image features calibration) in a virtuous cycle able to improve the quality of Data Viz design and interaction.

In this paper, we set up an experiment to start choosing the Data Viz stimuli able to influence the perceived quality of Data Viz during UX. The stimuli were chosen based on the analysis of traditional low-level image descriptors like color, orientation, and luminance contrast. The choice of these image descriptors instead of the IQA metrics rootes in that, while the latter focus on the evaluation of natural images and their distortions (e.g. blurriness, noise, jpeg-blockiness), the former are able to give us a description of the image in terms of low-level features for synthetic images like data visualizations.

These image features are investigated in the light of their interaction with the perceived value-in-use of Data Viz, measured through the validated items of the Infographics-Value (IGV) short scale [17]. Furthermore, participants' free comments on their experience with the Data Viz are exploited to provide quality labelling of the interaction experience and triangulate within the objective and the subjective evaluations with validated methods. Our contribution aims to investigate whether these simple image features explain value-in-use or are explained in terms of value-in-use and vice-versa in a virtuous cycle, and how to promote this crossing-over as a new and more holistic methodological asset for Data Viz quality. As said above, these new insights may, for example, help explain to the designers and the UX researchers users' impressions and causes of difficulties with their Data Viz prototypes.
The questions that we would like to investigate in this study are the following:

(1) Are objectively computed image features on Data Viz able to explain and being explained by quality dimensions? If yes, may such quality dimensions be those of value-in-use and/or free labelling subjectively perceived by users?
(2) May this crossing-over be used to better qualify objective evaluation and quantify subjective evaluation in a virtuous cycle?
(3) With reference to Q2, how this crossing-over may be used?

The paper is structured as follows: Section 2 introduces the image features, the qualitative items, and the experiment design, as well as the Data Viz selected for the study; Section 3 presents the results of our experiment, and Section 4 discusses the results by also resuming the above research questions. Section 5 concludes with a recap of the primary outcomes of this study, the limitations of the work conducted, and future directions.

2 METHOD

In what follows, we provide definitions and the application of computed image features to a Data Viz, namely: *colorfulness*, *feature congestion*, and *saliency*; we introduce the Infographics-Value short scale (IGV); we report the selection criteria of the four Data Viz used in the study; and we explain how we designed the experiment to let participants interact with pairs of Data Viz. We describe their subjective evaluation with a set of validated items and free text comments, further enriching the qualitative labelling of the interaction experience.

2.1 Image Feature Computation

The first feature deemed suitable to be computed on Data Viz is the ***colorfulness***, which consists of a linear combination of the mean and standard deviation of the pixel cloud in the sRGB color space. Such a linear combination was fitted to perceptual data collected from the authors within a user controlled experiment performed by Hasler and Suesstrunk [10]. From these correlations, they observed that values of the metric greater than 80 indicate high color perception.

As it is well known, clutter plays a crucial role in the design of user interfaces and data visualization [21]. Therefore, the second feature is the ***feature congestion*** measure of visual clutter as proposed by Rosenholtz et al. [21]. It is based on the analogy that the more cluttered a display is, the more difficult it is to introduce a visually salient object. The congestion measure of clutter considers three key features: color, orientation, and luminance contrast. Clutter maps for each of them were evaluated across scales and combined to get a single measure for each image. Rosenholtz et al. [21] tested this metric for different visual search experiments, obtaining values of the index greater than 4 in the case of more cluttered arrangements.

The third feature is not a single value but a topographic ***saliency*** map of the image, as proposed by Itti et al. [12]. This model is inspired by the visual attention mechanism, combining information across modalities in a bottom-up manner. The maps encode low-level visual features (intensity, orientation, and color) in a center-surround fashion at several spatial scales. The multiscale features are then combined into a single saliency map.

2.2 The IGV Short Scale for Subjective Value-in-Use Evaluation

The Infographics-Value (IGV) short scale measures the value-in-the use of Data Viz [16, 17]. The scale was made to assess the quality dimensions of Data Viz experienced by users during the execution of tasks in a contextualized scenario. Users were asked to retrieve a piece of information by explicitly interacting with the Data Viz. After usage, they were asked to rate the quality dimensions of ***usefulness***, ***intuitiveness***, ***clarity***, ***informativity***, and ***beauty*** of the Data Viz. The scale items can be used as a reference for assessing subjective measures of the quality of Data Viz or for comparing Data Viz quality dimensions. In this study, we adopted the dimensions of the IGV scale using them as a way to compare the two Data Viz shown to the participants. In particular, we used the quality items to compare pairs of Data Viz, by asking users to identify whether one or another of the two was perceived as having that quality, or both (or none) were perceived as having that quality. We obtained a variable, for each of the IGV quality dimensions, that summed up to the number of positive (vs. negative) evaluations for each Data Viz, included the positive (vs. negative) scores assigned to both the Data Viz.

2.3 Data Viz Selection Criteria

Starting from the Data Viz online repository of “Beautiful News”[1], we selected a subset of Data Viz, according to the following criteria:

- being a very “basic type” of Data Viz and be of familiar kind (bar chart, line chart, pictorial chart, area chart);
- having a “great variety” in the computed image feature values (Feature Congestion, Colorfulness, and Saliency — see Section 2.1 for the details of computation);
- using the same feature for different encoding (e.g., colors used either for decorative purpose only or for encoding categorical data; saliency resulting as being correctly maximum where the key information content is located or, at the opposite, being focused in the title or at the bottom of the chart, and the like);
- non-redundancy of charts (e.g., we avoided selecting two Data Viz of the same kind, like two bar charts, and preferred one Data Viz for each of the above-mentioned kinds);
- containing no more than one take-home message to be fast retrieved by users (e.g., a percentage incidence, a trend, a difference between two periods, and the like);
- depicting very general topics (i.e., non-technical), so as to be comprehensible by most part of people (e.g., rate of literacy, rate of poverty, behaviour change rate, and the like)

A total of 10 Data Viz were provisionally selected. These Data Viz were further inspected, and a final selection was carried out. For example, too similar, too complex, or those with lower feature values were discarded. This process was conducted by three of the authors, each one reporting on a table of criteria the values for each Data Viz and comparing the results among them until a final agreement was reached. We ended up with a selection of four Data Viz, reported in Figure 1. The selection criteria are summarized in Table 1.

The selection criteria determined a feature vector:
$\langle kind, saliency, feature\ congestion, colorfulness, color\ use\rangle$
for each Data Viz that we used to combine the Data Viz pairwise according to their most comprehensive kind of variability and to present this pair to participants for online interaction with them.

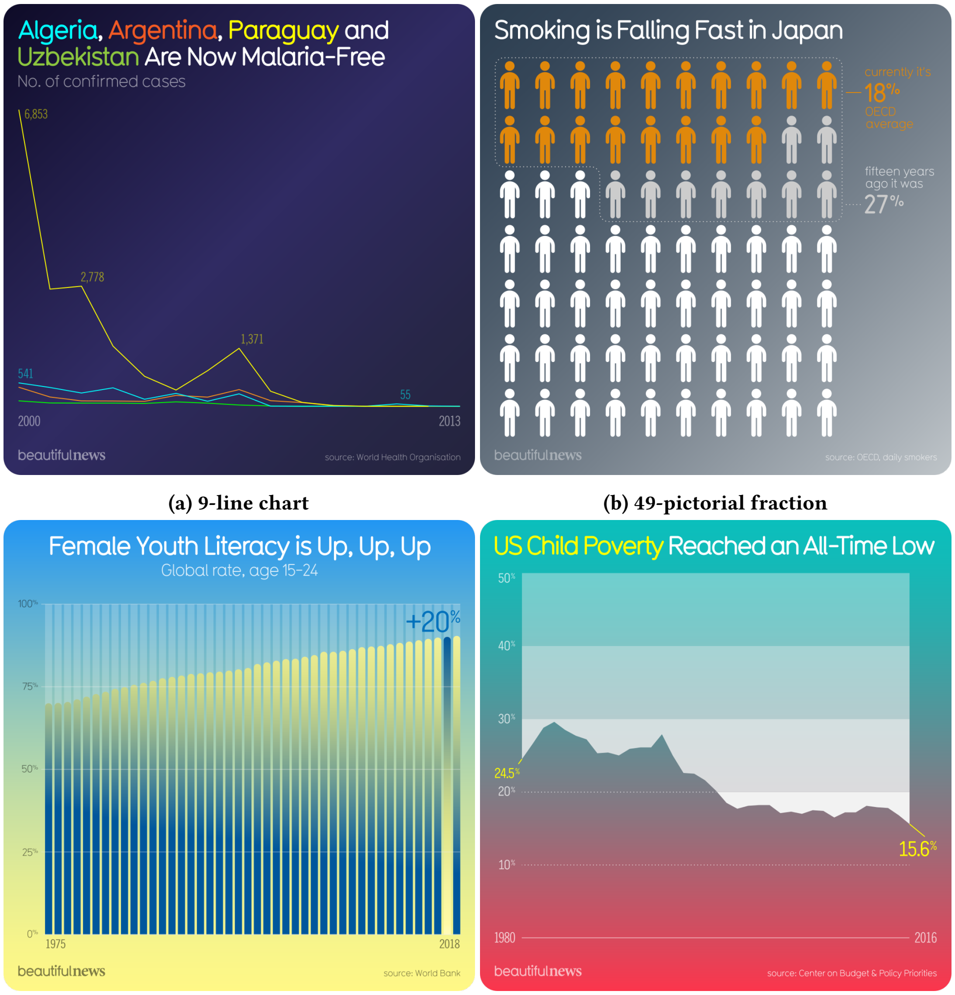
The full feature vectors of our Data Viz pairing activity are depicted in Table2.

They present quite the opposite combinations of image feature values as measured for the four Data Viz selected for our experiment. We deem the variability maximum: in the above matrices, we have four feature pairs with opposite values. Furthermore, this choice maximizes the opposition of chart kinds. Indeed, line charts and area charts have lines as a standard visual encoding. In pictorial fraction charts icons are often used to encode bar-composing icons, resulting more similar to bar charts than expected. Our combination avoids this possible overlapping of visual encoding features.

2.4 Experiment Design

The experiment was designed as an online form with two Data Viz administered to participants. For each Data Viz, the participants where initially asked to inspect it, then they were asked to reply to a question related to the information included in the Data Viz. We chose to have a “different” task for each Data Viz, i.e., a question related to the specific content of each Data Viz. Though this may be interpreted as a possible bias in the experiment, we argue that for us it is important to let user experience the value-in-use of Data Viz, and this necessarily implies to strictly adhere to the specific content of each Data Viz for the conceived task. Then, a questionnaire with the quality dimensions described in Section 2.2 was administered

[1] https://informationisbeautiful.net/beautifulnews/



(a) 9-line chart

(b) 49-pictorial fraction

(c) 1332-bar chart

(d) 428-area chart

Figure 1: The four Data Viz selected for our study. Their original names contain one ID, the one used in the paper to identify them (and named in their sub-captions, from left to right).

Table 1: Selection Criteria for the four Data Viz. The saliency description was derived by the manual inspection of a "black and white" transformation of each Data Viz where the white represented the saliency areas of the picture.

Data Viz ID	kind of chart	Feature Cong. Index	Colorfulness Index	Saliency description
9	line	2.75	40.61	title and bottom
49	pictorial fraction	4.83	43.34	equal and all over
428	area	2.40	92.71	only at the center
1332	bar	4.33	80.58	bottom and top of bars

to the participants, where, for each quality dimension, they were requested to indicate in which Data Viz the specific quality was more evident. Finally, two free text answers (one for each Data Viz) were required. They regarded positive or negative comments that participants may wish to write about their interaction experience. Either the order of presentation of the two Data Viz or the quality items were randomly shifted each time to avoid response biases. The structure of the online form administered to participants is reported in Appendix A.

As written above, the form was prepared in two variants: the first one included Data Viz ID 1332 and 9, and the second one included Data Viz ID 49 and 428. The choice of pairing the Data Viz in this way derived from the above criteria (see Section 2.3) and further considerations: it regards their kind, their objective measures of

Table 2: The two feature vector matrices of our final pairing for the four Data Viz selected.

	$F_{\{1332\wedge 9\}}$		$F_{\{49\wedge 428\}}$	
kind	bar	line	pictorial	area
saliency	- (top/bottom)	- (top/bottom)	- (all over)	+ (middle)
feature congestion	+	-	+	-
colorfulness	+	-	-	+
color use	decorating	mostly encoding	encoding	mostly decorating

feature congestion, *colorfulness*, and *saliency*, and the need to combine them according to the broadest variability of feature value. It is also related to the problem of administering a questionnaire online, which suffers more than on-site experiments from the fatigue bias, and the need to collect as much data as possible (by varying, for example, the pair of Data Viz compared) with the minimum effort required to the participants. The estimated completion time was 10 minutes. However, we did not consider the completion time in our analysis, because the task to be accomplished was conceived as a device aimed to interact with the Data Viz (i.e., having enough time to observe it and to evaluate it).

On these principles, we selected the first pairs of Data Viz for "Form number 1" including one Data Viz (ID 1332) with a *feature congestion* index among the highest (4.33) and a *colorfulness* index (80.56) among the highest, and another Data Viz (ID 9) with *feature congestion* index (2.75) among the lowest and the lowest *colorfulness* index (40.61) among the selected subset of Data Viz. On the opposite, the second version of the form (Form number 2) was made of a pair of Data Viz (ID 49) characterized by one Data Viz with the highest *feature congestion* index (4.83) and one of the lower *colorfulness* indices (43.34), and another Data Viz (ID 428) with the lowest *feature congestion* index (2.40, i.e., the lowest value in the subset of Data Viz selected for this study), and the highest *colorfulness* index (92.71, i.e., the highest value in the subset of Data Viz selected for this study).

Concerning saliency, the first pair of Data Viz is characterized by having a focused *saliency* at the top and bottom of their layout (IDs 9 and 1332). In contrast, the second pair of Data Viz is characterized by having a higher *saliency* all over (ID 49) or focused in their center (ID 428) (see also 2).

In this way, we were able to collect and test for many combinations of variable values using both between- and within-group tests. For example, we tested the correlation of each quantitative dimension with each qualitative dimension by comparing pairwise results in a "control" vs. "treatment" fashion. Table 3 explains the combination of tests where each time we had a "control" variable (those with similar values) and one or more "treatment" variables (those with opposite values). With within-group tests, we could attempt more holistic results based on a combinations of variables.

In what follows, we present the results for each test carried out and depicted in the above table.

3 RESULTS

The two versions of the form were left online for two weeks (from mid-November to the end of November 2022). The questionnaires were advertised at the University of each of the authors during their classes. The participation was voluntary and anonymous, and duplicate compilations were avoided. A total of 58 participants completed Form number 1, and a total of 40 participants completed Form number 2. After a check of the responses, all were retained as valid ones. After codifying the valid answers to the questions related to the information content of each Data Viz into right (1) and wrong (0), we proceeded to codify the answers related to each IVG scale quality dimension for each Data Viz. For this task, we decided to assign the total score referred to each Data Viz and half of the scores referred to both Data Viz to each of the two. In the end, we also codified the free text comments related to what the participant had appreciated or not appreciated of each Data Viz just experienced. The codification was carried out by a classification of all the answers into topics (e.g., colors) and subtopics (e.g., readability of colors), and a dichotomous value standing for positive comment (1), negative comment (-1) or neutral one (0). The codification was conducted independently by two authors, and a final agreement was reached with a substantial Cohen's k intercoder reliability [19] of 0.65 (considered as moderately satisfying).

3.1 The Task execution

3.1.1 Within-group A and B tests. All participants completed the task by answering one question related to the information content for each Data Viz. Checking for statistical independence of the differences between right and wrong answers for Data Viz 1332 and 9 led to running a χ^2 test, with a significance level of 0.05. The relation between these variables was significant, χ^2 (1, N = 58) = 85.01, p < .001. Participants were much more likely to give the correct answer for Data Viz 9 than for Data Viz 1332, with statistical significance. The effect size, comparable to Cohen's d and calculated on this test [7, 18] is 3.3.

Regarding the Form 2 questionnaire, the χ^2 test run between the right and wrong answers to the question of Data Viz 49 resp. Data Viz 428 was χ^2 (1, N = 40) = 7.31, p < .01. Participants were much more likely to give the correct answer for both Data Viz than the wrong one. The effect size of the test is 1.34.

3.1.2 Between-group C, D, E, and F tests. Comparing the right and wrong answers in test C, i.e., during interaction with Data Viz 1332 and Data Viz 428, led running the χ^2 on the two groups of participants of Form 1 and Form 2. The relation between these variables was significant, with χ^2 (1, N = 98) = 62.62, p < .001. This result means that participants were much more likely to give the correct answer for Data Viz 428 than Data Viz 1332.The effect size of the test is 9.30.

Test D was computed to compare the right and wrong answers of the task related to Data Viz 49 and Data Viz 9, and the χ^2 test on the two groups of participants of Form 1 and Form 2 led to the result

Table 3: Statistical tests with control vs. treatment variables. The sign "+" means that the control (resp. treatment) variable has a higher value for the first Data Viz of the row resp. the second one; the sign "-" means the opposite (see also Section 2.3 for details).

Test no: DV IDs	Group Test	Control variable(s)	Test variable(s)
A: 1332 and 9	within	Saliency (both -)	Feature Congestion (+ and -), Colorfulness (+ and -)
B: 428 and 49	within	Saliency (both +)	Feature Congestion (- and +), Colorfulness (+ and -)
C: 1332 and 428	between	Colorfulness (both +)	Feature Congestion (+ and -)
D: 49 and 9	between	Colorfulness (both -)	Feature Congestion (+ and -)
E: 1332 and 49	between	Feature Congestion (both +)	Colorfulness (+ and -)
F: 428 and 9	between	Feature Congestion (both -)	Colorfulness (+ and -)

of χ^2 (1, N = 98) = 3.53, p=0.06, hence a non statistically significant difference.

Test E aimed to compare the right and wrong answers of the task related to Data Viz 1332 and Data Viz 49. The relation between these variables was significant, with χ^2 (1, N = 98) = 90.87, p < .001. This result means that participants were much more likely to give the correct answer for Data Viz 49 than Data Viz 1332.The effect size of the test is 4.27.

In the end, the last test was F, comparing the right and wrong answers of the task related to Data Viz 428 and Data Viz 9. The relation between these variables was significant, with χ^2 (1, N = 98) = 1.34, p = .24, hence a non statistically significant difference.

3.2 The perceived quality of interaction: IGV Quality Dimensions

The following results aim to assess each quality dimension on each pair of Data Viz, in order to confirm whether one Data Viz was perceived significantly better from the point of its *usefulness*, *informativity*, *clarity*, *beauty*, and *intuitiveness* than another one. This qualitative evaluation from participants may then be put in relation to the characteristics of each Data Viz as measured at the beginning, i.e., the feature congestion level, the colorfulness level, and the saliency level to extract new information about correlations between objective and subjective measures.

3.2.1 Within-group A and B tests comparison. Regarding test A, a binomial test[2]. Throughout the paper, each time we ran a binomial test, we converted it into a z-test subject to the above condition, with a significance level of 0.05. This test was ran in order to compute the polarization of responses for each of the five quality dimensions perceived and rated by participants.

In Test A, for Data Viz 1332, clear negative polarizations have been detected for the following items (negative vs positive): *intuitiveness* (0.71 vs. 0.27, p >.0001, and effect size 0.48); *beauty* (0.70 vs. 0.28, p < .001, and effect size 0.44); *informativity* (0.64 vs. 0.36, p = .03, and effect size 0.29); *clarity* (0.70 vs. 0.28, p < .001, and effect size 0.44). On the opposite, for Data Viz 9, these are all to be intended as clear positive polarizations.

Regarding test B, for Data Viz 49, clear negative polarizations have been found for the following items: *intuitiveness* (0.68 vs. 0.32, p = .04, and effect size 0.36); *clarity* (0.79 vs. 0.21, p <.001, and effect size 0.64). On the contrary, for Data Viz 428, this are all positive polarization for both the above dimensions.

3.2.2 Between-group C, D, E, and F tests comparisons. For test C, results of the two-sample proportion test[3] with significance level 0.05 indicated that there is only one significant difference in the *beauty* item between Data Viz 1332 proportion of positive polarization (29%) and Data Viz 428 proportion of positive polarization (60%), Z = 2.88, p = .004. The effect size is 0.64. Another significant difference (although at a significant level of 0.1) of positive polarization proportions was found for the *informativity* item, between Data Viz 1332 (36%) and Data Viz 428 (55%), z = 1.71, p = .09, with effect size 0.39.

Regarding test D, there is a significant difference in the *beauty* item between Data Viz 9 proportion of positive polarization (71%) and Data Viz 49 proportion of positive polarization (43%), Z = 2.65, p < .01, and effect size is 0.59.

Test E results show a significant difference in positive polarization proportion for item *intuitiveness* between Data Viz 1332 (27%) and Data Viz 49 (68%), Z = 3.77, p < .001, and effect size 0.83; also item *clarity* shows a significant difference between Data Viz 1332 positive polarization proportion (29%) and Data Viz 49 positive polarization proportion (80%), Z = 4.79, p < .001, and effect size 1.08.

Test F showed a significant difference in positive polarization proportions for item *intuitiveness* between Data Viz 9 (71%) and Data Viz 428 (33%), Z = 3.6, p < .001, and effect size 0.8; also item *clarity*, for the same Data Viz, shows a significant difference, with 69% of positive polarization proportion vs. 23% of positive polarization proportion , Z = 4.38, p < .001, and effect size 0.98.

3.3 The Free Quality Labelling from Participants

As written above, each participant was invited to comment briefly, for each used Data Viz to accomplish the task, about what they have appreciated or not appreciated in their experience with it. All comments were codified as reported in Section 2. The topics and subtopics into which they were codified are depicted in Figure 2b. All comments were retained as valid, no comments were codified as neutral comments, there were either positive or negative ones, and some comments contained either positive or negative statements (hence, they were codified as both positive and negative for each

[2] We used the following rule of thumb: if $np \geq 10$ and $nq \geq 10$, the binomial test becomes a z-test, see also [14]

[3] all the sample proportion tests were run using the following calculatorhttps://www.statskingdom.com/index.html.

related subtopic). The proportions for negative and positive comments for each Data Viz related to topics and subtopics are reported in Figure 2a.

We ran a binomial test to identify statistically significant negative vs. positive polarization between comments for each pair of Data Viz. The results are synthesized in the table of Figure 3. This figure reports the result of the test, only in terms of the p-value when below the alpha level of 0.05, for each free text comment label as codified manually, and for each Data Viz. Color and sign of the figure report positive (green and +) and negative (red and -) polarizations. For the sake of clarity, we do not report the results of the binomial statistics in the figure, but only its significance.

These codified comments from participants are used in the next discussion session, as an interpretation support that can triangulate between the objective measurements (image features) computed at the beginning and the subjective measurements of the validated IGV quality items [20]. Appendix B reports the details of the binomial statistics for the sake of completeness.

4 DISCUSSION

The analysis of results show that participants considered some Data Viz as more problematic than others. The evidence of this outcome is discussed in what follows, putting together the feature matrices of the Data Viz criteria of Table 2, the within- and between-group statistical analysis of Section 3 and the free text comments analysis summarized in Figure 3.

The results show that Data Viz 1332 (bar chart) was significantly more complex to interpret than all the other three Data Viz (9-line chart, 49-pictorial and 428-area chart). Most part of participants could not give the correct answer to the question related to this Data Viz, and this phenomenon was significantly quantifiable in orders of magnitude difference[4]. Data Viz 1332 is characterized for having a higher value of both the *feature congestion* index and *colorfulness* index and a saliency which was comparable to that of its companion Data Viz (9), i.e., at its top and bottom, leaving the middle more “undistinguished”.

The IGV qualitative items analysis shows that Data Viz 1332 was considered significantly less *useful*, *informative*, *clear*, and *beautiful* with respect to its companion Data Viz (9). The comments by participants make evident that Data Viz 1332 received more negative comments than chance regarding all of the following aspects and other Data Viz (in parenthesis): *color choice* (compared with 9 and 428); *color encoding* (compared with 428 and 49); *chart choice* (compared with 429); *quality of design* and *quality of interaction* (compared with 9 and 49). Interestingly, Data Viz 1332 is a bar chart, which is deemed as one of the most familiar, the most precise for differences comparison, and most popular ones.[5]

Regarding Data Viz 49, it was considered significantly less *intuitive* and *clear* than its companion Data Viz in test B (ID 428). The former is a pictorial fraction chart with the highest *feature congestion* index value and one of the lower *colorfulness* index values. On the contrary, Data Viz 428 is an area chart with the lowest *feature congestion* and the highest *colorfulness*. In terms of *saliency*, Data Viz 49 was salient all over (without a key area of saliency). In contrast, Data Viz 428 was more salient in the middle (being the key information contained in the middle of the area chart). The qualitative comparison of these two Data Viz may reveal that *feature congestion* is negatively impacting the *intuitiveness* and the *clarity* of a Data Viz, e.g., by making it cognitively harder to be processed. Also, *saliency* seems to have a role in the perceived *clarity* and *intuitiveness* of a Data Viz, hence in the immediacy of interpretation and in lowering the cognitive effort required to process it. This result also seems to be confirmed by test E, where Data Viz 1332 (bar chart with top/bottom saliency) is significantly less *intuitive* and *clear* than 49 (where saliency, although not focused in a defined area of the chart, is nevertheless related to all the key informative data points, i.e., the pictorial icons).

The above results may also suggest that *feature congestion* is the more problematic feature when exceedingly high, at least for the perceived quality of *clarity* in a Data Viz (all of the tests based on the qualitative items tend to confirm this).

When the Data Viz is also charged with another high-value feature (e.g., *colorfulness*), this combination may even worsen the interaction, and the perceived *usefulness*, *informativity*, and *beauty*. This seems to be confirmed whenever color is not used for data encoding but rather for decorative purposes. Evidence of this is given by the fact that Data Viz 9 (the line chart using color for encoding trends) results as more *intuitive* and *clear* than 428 (the area chart where color is mostly used as a decorative device), and more *beautiful* than Data Viz 49 (the pictorial fraction chart where the color is used for data encoding, but with a prevalence of less saturated colors —gray and orange). A piece of evidence in this direction is also shown in test E: Data Viz 9 has significantly more positive comments related to *color choice*, *quality of design*, and *quality of interaction* concerning Data Viz 428. However, Data Viz 49 collects significantly more positive comments regarding *color encoding* than Data Viz 9. This result may hint at the influence of *saliency* on the quality perception of *beauty* (Data Viz 49 has a good saliency for each data points, whereas Data Viz 9 is less salient and more at the top of it — the title and legend, rather than in the area of data encoding — the multi-lines chart).

In test C, Data Viz 1332 was considered less *informative* and *beautiful* than Data Viz 428. These two Data Viz have opposite *feature congestion*. Also, *saliency* location is opposite. This result may reveal that interaction among these two image features may hinder the aesthetic experience and even the data *informativity*. To strengthen this evidence, the participants’ comments seem to confirm the significantly higher *chart information fit* of Data Viz 428 vs. Data Viz 1332. They both depict the rate of change of two phenomena and, indeed, the area chart is perceived as more suitable than the bar chart.

The perceived difference in terms of quality is also significant about the *colorfulness* when comparing significantly positive comments for Data Viz 49 (using color mostly for data encoding) for the *quality of design* subtopic with respect to Data Viz 428 (using color mostly for decoration purposes and partly for data encoding). Interestingly, Data Viz 428 results have significantly more negative

[4] Verifying this is as simple as computing the odds ratio between either the proportions of wrong or the proportions of correct answers for each pair of Data Viz

[5] Looking at the *Google NGram* Viewer, papers with topic bar charts are an order of magnitude more numerous than papers with topic line chart, area chart, and pictorial charts. See for example https://books.google.com/ngrams/graph?content=bar+chart&year_start=1800&year_end=2019&corpus=en-2019&smoothing=3, last accessed January 2023.

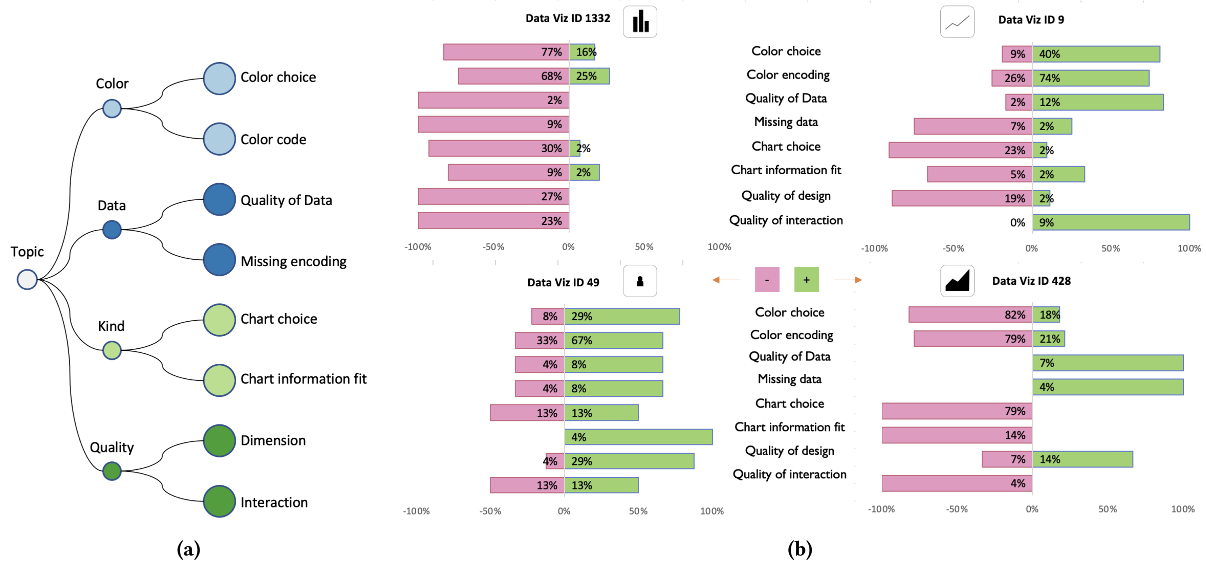
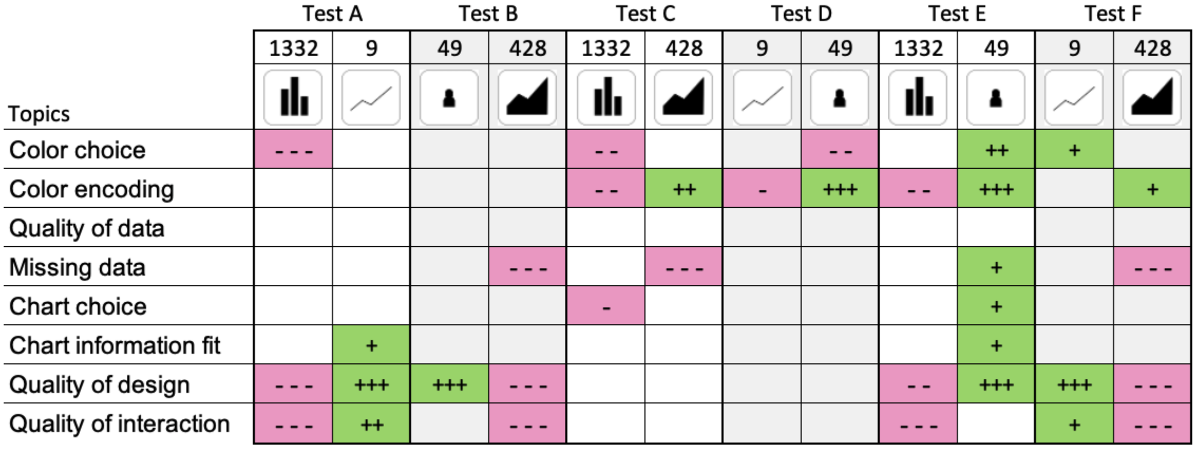


Figure 2: On the left (2a): Topics and subtopics emerged from the codification of free text comments from participants. Each comment was labelled as positive or as negative for each of the topics and subtopics. On the right (2b): Proportion of negative and positive comments for each Data Viz, by subtopics.



	Test A		Test B		Test C		Test D		Test E		Test F	
Topics	1332	9	49	428	1332	428	9	49	1332	49	9	428
Color choice	- - -				- -			- -		++	+	
Color encoding					- -	++	-	+++	- -	+++		+
Quality of data												
Missing data				- - -		- - -				+		- - -
Chart choice					-					+		
Chart information fit		+								+		
Quality of design	- - -	+++	+++	- - -					- -	+++	+++	- - -
Quality of interaction	- - -	++		- - -					- - -		+	- - -

Figure 3: Binomial test run on each pair of Data Viz; p-values are reported only if statistically significant. The number of '-' or '+' determines the p-values: -(+) stands for p < .1, - -(++) stands for p. < .05, and ---(+++) stands for p < .001. The + or - sign identifies a positive (resp. negative) comment polarization for the corresponding Data Viz (see also Table1).

comments for *missing data* than Data Viz 49, 1332, and 9, and more negative comments related to the *quality of design* and the *quality of interaction* than Data Viz 49 and 9 (both using color for data encoding). This is surprising considering that Data Viz 49 is more *feature congested* than Data Viz 428. This result may suggest that participants were sensible to *colorfulness*, up to the point that this feature can emphasize or mitigate the effect of *feature congestion* and to overall and negatively impact a good design perception and the quality of interaction experience. Further considerations in this direction may regard the influence of the aesthetics quality of Data Viz (e.g., the perceived *beauty*) on the participants' perception and of the quality of interaction experience overall.

Resuming the research questions of Section **??**, we may conclude that some statistically significant reciprocal explanatory relations exist between image features measured quantitatively and the value-in-use of Data Viz measured qualitatively; furthermore, also some qualitative labelling from users during their experience with Data Viz were significantly related with both image features quantification and value-in-use qualification items. Ways to exploit these each other's explanatory relationships evaluations are, for example, considering them in tandem, either in the design phase or during a user experience study, by adding the right mix of features in a Data Viz and by raising awareness of their interaction with the perception of the quality of experience with them. We may conclude that studies of this kind provide proof of concept of the virtuous cycle triggered by crossing-over evaluation methods as constituents of a whole construct of Data Viz quality rather than using a one-way approach or different methods taken individually.

5 LIMITATIONS AND CONCLUSIONS

One of the limitations of our work lies in the number of features considered, and the number of Data Viz taken into exam. Our choice is motivated by the fact that we preferred to start with a minimum but meaningful set of image features and focus on a depth-first analysis of a few Data Viz rather than a width-first analysis involving many kinds of Data Viz. This choice also reduced the variables to be controlled and the possible confounders. Another limitation regards the feature metrics we have considered, for example, the fact that only bottom-up saliency models were identified. It should be interesting to integrate top-down factors/elements within our framework. Also considering 3D visualizations could be considered an extension to the present work, adding to the features set the one of "navigability" [15]. A third limitation depends on the experiment design choice, which is only to compare objective measurements with subjective measurements, and not, for example, setting a-priori hypotheses about objective or subjective measurement interactions. This is again due to the need to delimit the experiment complexity and the variables to be controlled in order to give a deeper rather than wider view.

That said, our contribution may be identified clearly in the importance of attenuating the *feature congestion* in Data Viz in favor of the perceived *intuitiveness* and *clarity* of Data Viz, as well as to reach a good distributed *saliency* not to impact on these quality dimensions; also *colorfulness* may influence the perceived quality of Data Viz, especially in the perceived *beauty*; and both may impact on the perceived *usefulness*, *informativity* and *beauty*. *Colorfulness* can improve or worsen the above quality dimensions and the perceived overall *quality of design* and *quality of interaction* when used and combined with *feature congestion*. In contrast, *saliency* may attenuate the harmful impact of a high *colorfulness* on the quality dimension of *beauty*. Research questions found a satisfactory answer in that taking into account the whole design of this study, from the selection criteria to the user experience study, can be seen as a methodological framework towards the definition of a Data Viz quality assessment method based on a virtuous cycle of quantitative and qualitative evaluations.

As future work, we consider taking into account high-level image features like, for example, *memorability* [5], *aesthetics* [25] and *complexity* [9]. Another step will regard the exploitation of eye-tracking tools that will provide us with a ground-truth map of the *saliency*. Also, making a more detailed analysis and integrating the *textual* and the *visual* modalities present in the Data Viz, would enrich the evaluation frame.

We also consider widening the target of Data Viz to be analyzed, by taking into exam other kinds of charts, representing data of different nature and granularity, Data Viz containing more than one chart, 3D visualizations, and the like.

REFERENCES

[1] Carmelo Ardito, Paolo Buono, Maria F. Costabile, Antonella De Angeli, and Rosa Lanzilotti. 2008. Combining Quantitative and Qualitative Data for Measuring User Experience of an Educational Game. *EL-C. Law, N. Bevan, G. Christou, M. Springett & M. L{á}rusd{ó}ttir (eds.) Meaningful Measures: Valid Useful User Experience Measurement (VUUM)* (2008).

[2] Carlo Batini, L. Furlani, and Enrico Nardelli. 1985. What is a Good Diagram? A Pragmatic Approach. In *Proceedings of the Fourth International Conference on Entity-Relationship Approach*. IEEE Computer Society, USA, 312–319.

[3] Tanja Blascheck, Frank Bentley, Eun Kyoung Choe, Tom Horak, and Petra Isenberg. 2021. Characterizing Glanceable Visualizations: From Perception to Behavior Change. In *Mobile Data Visualization*. Chapman and Hall/CRC, 151–176. https://doi.org/10.1201/9781003090823-5

[4] Tanja Blascheck, Lonni Besançon, Anastasia Bezerianos, Bongshin Lee, Alaul Islam, Tingying He, and Petra Isenberg. 2023. Studies of Part-to-Whole Glanceable Visualizations on Smartwatch Faces. In *PacificVis 2023 - The16th IEEE Pacific Visualization Symposium*. Seoul, South Korea. https://hal.inria.fr/hal-04018448

[5] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What makes a visualization memorable? *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2306–2315.

[6] P. Buono, D. Caivano, M. F. Costabile, G. Desolda, and R. Lanzilotti. 2020. Towards the Detection of UX Smells: The Support of Visualizations. *IEEE Access* 8 (2020), 6901–6914. https://doi.org/10.1109/ACCESS.2019.2961768

[7] Susan Chinn. 2000. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in medicine* 19, 22 (2000), 3127–3131.

[8] Gianluigi Ciocca, Silvia Corchs, Francesca Gasparini, Carlo Batini, and Raimondo Schettini. 2016. *Quality of Images*. Springer International Publishing, Cham, 113–135. https://doi.org/10.1007/978-3-319-24106-7_5

[9] Silvia Elena Corchs, Gianluigi Ciocca, Emanuela Bricolo, and Francesca Gasparini. 2016. Predicting complexity perception of real world images. *PloS one* 11, 6 (2016), e0157986.

[10] David Hasler and Sabine E. Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, Vol. 5007. SPIE, 87–95.

[11] Jessica Hullman. 2019. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 130–139.

[12] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20, 11 (1998), 1254–1259.

[13] Peter R Keller, Mary M Keller, Scott Markel, A John Mallinckrodt, and Susan McKay. 1994. Visual cues: practical data visualization. *Computers in Physics* 8, 3 (1994), 297–298.

[14] Jeffrey E Kottemann. 2017. *Illuminating statistical analysis using scenarios and simulations*. John Wiley & Sons.

[15] Matthias Kraus, Johannes Fuchs, Björn Sommer, Karsten Klein, Ulrich Engelke, Daniel Keim, and Falk Schreiber. 2022. Immersive analytics with abstract 3D visualizations: A survey. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 201–229.

[16] Angela Locoro, Federico Cabitza, Rossana Actis-Grosso, and Carlo Batini. 2017. Static and interactive infographics in daily tasks: A value-in-use and quality of interaction user study. *Computers in Human Behavior* 71 (2017), 240–257.

[17] Angela Locoro, Federico Cabitza, Aurelio Ravarini, and Paolo Buono. 2020. IGV short scale to assess implicit value of visualizations through explicit interaction. *Applied Sciences* 10, 18 (2020), 6189.

[18] Jake Olivier and Melanie L Bell. 2013. Effect sizes for 2× 2 contingency tables. *PloS one* 8, 3 (2013), e58777.

[19] Cliodhna O'Connor and Helene Joffe. 2020. Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods* 19 (2020), 1609406919899220.

[20] Michael Quinn Patton. 1999. Enhancing the quality and credibility of qualitative analysis. *Health services research* 34, 5 Pt 2 (1999), 1189.

[21] Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. 2007. Measuring visual clutter. *Journal of vision* 7, 2 (2007), 17–17.

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

[23] Thomas Tullis and William Albert. 2013. *Measuring the User Experience, Second Edition: Collecting, Analyzing, and Presenting Usability Metrics* (2nd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[24] Colin Ware. 2019. *Information visualization: perception for design*. Morgan Kaufmann.

[25] Lu Zhao and Haimin Sun. 2022. Technical Aesthetics Strategy of Information Visualization. In *Cross-Cultural Design. Interaction Design Across Cultures: 14th International Conference, CCD 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part I*. Springer, 302–311.

A ONE OF THE TWO FORMS USED IN OUR EXPERIMENT

Here we depict the structure of one of the two forms used in our experiment. The form is divided in two parts: part one contains the two Data Viz and the task (question) to be accomplished by interacting with it; part two contains the qualitative evaluation (the IGV items + open questions about pros and cons of the two Data Viz).

A.1 Part One: Data Viz and a task to be accomplished

Look at the 2 Data Viz below and answer a question under each one (answers are required). Then evaluate the two infographics with the 5 proposed comparative quality dimensions (mandatory choice). Finally, add your considerations on each of the two infographics in the dedicated free comment parts (optional).

Viz ID 1332

Question: Since which year has the phenomenon shown in the graphic grown without decreasing?

Viz ID 9

Question: Which country had the lowest incidence of cases over the period?

A.2 Part Two: Qualitative Evaluation

Based on your use of the two Data Viz, evaluate which dimension you would attribute more to one or more to the other.

Table 4: The IGV items presented for the qualitative and controlled evaluation phase of the experiment.

Dimension	Plus the first	Same as	Plus the second
Useful	○	○	○
Informative	○	○	○
Beautiful	○	○	○
Clear	○	○	○
Intuitive	○	○	○

Comment on the first Data Viz (is there anything you particularly liked or disliked about it)?

free text answer

The same question is proposed for the second Data Viz

B BINOMIAL TESTS ON THE CODIFIED PARTICIPANTS' COMMENTS

The following table reports the result of the binomial test statistics for participants' comments.

Table 5: Binomial statistics with p-value depicted as one, two or three asterisk depicting the significance.

Topic	Test A		Test B		Test C		Test D		Test E		Test F	
	1332	9	49	428	1332	428	9	49	1332	49	9	428
Color choice	3.3***				2.2**		2.7**		2.4**		1.7*	
Color encoding					2.6**	2.1**	3.3***	1.7*	3.8***	2.3**	1.9*	
Quality of data												
Missing data			4.8***		4.1***				1.7*		4.6***	
Chart choice					1.6*				1.9*			
Chart information fit	2.3*								1.9*			
Quality of design	4.6***	4***	3.3***	3.3***					3.4***	2.8**	4.4***	4.4***
Quality of interaction	2.5**	6.4***	5.3***						5.4***		1.9*	6.2***