International Journal of Human-Computer Interaction

NaVi-Q: A Quality Model for Evaluating Narrative Information Visualizations

--Manuscript Draft--

Manuscript Number:	IJHC-D-23-01540
Full Title:	NaVi-Q: A Quality Model for Evaluating Narrative Information Visualizations
Article Type:	Research Article
Section/Category:	Big Data visualization
Manuscript Region of Origin:	
Abstract:	Evaluating narrative visualizations can be a complex task due to their unique characteristics, which poses difficulties in formulating objective evaluation criteria and developing standard evaluation procedures. This study introduces an evaluation model designed to assess the quality of narrative visualizations. Our goal was to capture the fundamental features of these visualizations to enable a comprehensive evaluation, allowing teams to identify areas for improvement. We draw upon existing models in Software Engineering and Information Visualization. To test its feasibility, we conducted a small-scale evaluation with three evaluators of varying levels of expertise. The results demonstrated high reliability, suggesting the model is robust and capable of producing consistent results regardless of the evaluators' background. The model also serves as a valuable tool in defining clear quality criteria and providing a systematic approach to evaluation. It holds practical implications for creating standardized designs, bridging the gap between technical and non-technical teams.

NaVi-Q: A Quality Model for Evaluating Narrative Information Visualizations

Andrea Lezcano Airaldi[a]*, E. Irrazábal[a], and J. A. Diaz-Pace[b]

[a]Department of Informatics, National University of the Northeast, Corrientes, Argentina; [b]ISISTAN, CONICET-UNICEN, Tandil, Argentina

Andrea Lezcano Airaldi * Corresponding author

alezcano@exa.unne.edu.ar

https://orcid.org/0000-0003-1361-2797

Emanuel Irrazábal

eirrazabal@exa.unne.edu.ar

https://orcid.org/0000-0003-2096-5638

Jorge Andrés Diaz-Pace

andres.diazpace@isistan.unicen.edu.ar

https://orcid.org/0000-0002-1765-7872

Disclosure statement. The authors have no competing interests to declare.

Acknowledgements. This work was supported by the National Council on Scientific and Technical Research (CONICET) under a PhD Fellowship (RESOL-2021-154-APN-DIR#CONICET) and the National University of the North-East (SCyT-UNNE) under Grant 21F005. It is a part of the research conducted under the Computer Science Doctorate Program at UNNE, UNaM, and UTN.

Data availability: The datasets generated during and/or analysed during the current study are available in the OSF repository, https://doi.org/10.17605/OSF.IO/BT98H.

NaVi-Q: A Quality Model for Evaluating Narrative Information Visualizations

Abstract. Evaluating narrative visualizations can be a complex task due to their unique characteristics, which poses difficulties in formulating objective evaluation criteria and developing standard evaluation procedures. This study introduces an evaluation model designed to assess the quality of narrative visualizations. Our goal was to capture the fundamental features of these visualizations to enable a comprehensive evaluation, allowing teams to identify areas for improvement. We draw upon existing models in Software Engineering and Information Visualization. To test its feasibility, we conducted a small-scale evaluation with three evaluators of varying levels of expertise. The results demonstrated high reliability, suggesting the model is robust and capable of producing consistent results regardless of the evaluators' background. The model also serves as a valuable tool in defining clear quality criteria and providing a systematic approach to evaluation. It holds practical implications for creating standardized designs, bridging the gap between technical and non-technical teams.

Keywords: information visualization, data storytelling, narrative visualization, evaluation, quality model.

1 Introduction

Narrative visualizations are a type of information visualization that use storytelling techniques to communicate data and insights (Segel & Heer, 2010). They combine images, text, and annotations to convey the meaning and importance of the data, making complex information accessible and engaging for users. Some examples include (Cairo, n.d.; *Food Prices Are Outpacing Wider Inflation across Most of the World | The Economist*, n.d.; *Paramount Pushes in: New Streamers Are Still Finding Ways to Grow*, n.d.; Jacobson, 2019).

Evaluating a visualization technique is important for understanding how users perceive and interpret the underlying information. It is often described as a complex task (Carpendale, 2008; Elmqvist & Yi, 2015; Lam et al., 2012; Plaisant, 2004), because it involves assessing the visualizations themselves and the cognitive activities they support, such as exploratory analysis, communication, or decision-making. In addition to the well-known difficulties associated to traditional visualizations (Carpendale, 2008; Plaisant et al., 2008), narrative visualizations are challenging as they often depend on the context in which they are used and the characteristics of their users (Hullman & Diakopoulos, 2011).

Although some evaluation methods have been developed for assessing visualizations (Forsell & Johansson, 2010; Lan et al., 2021; Padda et al., 2007, 2008; Wall et al., 2019; Zuk et al., 2006), they often fail to address the complexities of real-world tasks (Carpendale, 2008). Therefore, there is a growing need for more standardized, automated evaluation tools that allow for holistic, iterative assessment of narrative visualizations, as shown in our previous study (Lezcano Airaldi et al., 2022). This includes developing more objective metrics for assessing aspects such as aesthetics, comprehension, or user engagement.

In the field of Software Engineering, the ISO/IEC 25010 standard (ISO/IEC 25010, 2011) defines quality models for software products, a hierarchy of quality properties of interest and how to quantify and assess them. Due to the different nature of narrative visualizations, these quality models cannot be applied directly as they are. To ensure the intended quality of a visualization and develop a meaningful evaluation model, it is necessary to understand the application context and adjust its definition.

The contribution of this work is the development of a quality model for the evaluation of narrative information visualizations, drawing upon existing models in SE. We describe the process for the development and construction of the model. Then, we present a small-scale evaluation with 18 visualizations, static as well as interactive, to assess the feasibility of the model.

The rest of this paper is organized as follows. Section 2 reviews background and related works. Section 3 describes the process for constructing the quality model for narrative visualizations. Section 4 presents the small-scale evaluation. Section 5 discusses the main findings of the study and their implications. Finally, Section 6 presents the conclusions and outlines future research.

2 Background and Related Works

This section presents an overview of evaluation in the field of information visualization, including existing methodologies, to contextualize the aim of this study.

2.1 Quality of Visualizations

To build a quality model, one needs to define the components of an effective, high-quality visualization and determine the set of attributes to be assessed. Although there is a general consensus on what constitutes an effective visualization, the definition is often incomplete and inconsistent (Zhu, 2007). In general, a visualization is considered effective if it helps users extract accurate information (Card et al., 1999; Zhang et al., 2015).

Mackinlay (Mackinlay, 1986) defines effectiveness as a measure of how well a visualization exploits the display media and the visual perception of the user. Zhu (Zhu, 2007) reviews the existing definitions and characterizes effectiveness in terms of three principles: accuracy, utility, and efficiency. Bai et al. (Xiaoyan Bai et al., 2010) states that in order to be effective, visualizations need to provide useful, relevant and actionable information that allows users to address their requirements.

In Software Engineering (SE), the ISO/IEC 25010 standard (ISO/IEC 25010, 2011) provides a systematic approach for modeling quality requirements. These standards propose different definitions of properties grouped into several categories with a decomposition structure. For the development of our model, we define the quality of a narrative visualization as *"the degree to which the visualization satisfies the declared and implicit needs of the various stakeholders and, therefore, adds value"*, in accordance with ISO/IEC 25010.

2.2 Evaluation in Information Visualization

Several methodologies have been developed to evaluate information visualizations. The methods can be classified into quantitative and qualitative. On the one hand, quantitative studies, also known as controlled experiments, focus on collecting performance measurements such as time and errors and require a high degree of precision (Carpendale, 2008). Qualitative studies, on the other hand, are generally less formal and consider the interplay between several factors to achieve a richer understanding (Patton, 2014).

Some examples of quantitative evaluation include studies on graphical perception (Heer & Bostock, 2010; Nowell et al., 2002; Perin et al., 2018), interaction (Feng et al., 2019; Yi et al., 2007), or aesthetics in visualizations (Cawthon & Moere, 2007; Lei et al., 2018; Quispel et al., 2016). For qualitative evaluation, the most common approach is usability heuristics, that while useful, do not capture all the dimensions of quality being relevant to information visualizations (Santos et al., 2015; Williams et al., 2018; Zhu, 2007).

As various authors have pointed out (Carpendale, 2008; Elmqvist & Yi, 2015; Lam et al., 2012; Plaisant, 2004) an empirical evaluation beyond traditional metrics is difficult because the effectiveness of visualizations depends on several subjective factors, such as the prior knowledge of the user (Tory & Moller, 2005; Yi, 2010), goals, or context (Hullman & Diakopoulos, 2011; Padda et al., 2007). Additionally, the interpretation of visualizations can be influenced by individual biases and preferences, which makes it difficult to develop objective evaluation criteria and standardized procedures.

In their review, Lam et al. (Lam et al., 2012) provide an overview of evaluation scenarios, categorized into those for understanding data analysis processes and those that evaluate visualizations themselves. They base their categorization on questions and goals, rather than on existing methods. Our proposed model falls within the latter group and is closer in essence to the User Experience (UE) scenario, which aims to "understand how well the visualization supports the intended tasks" and can help uncover limitations in the visualization design.

2.3 Existing approaches to evaluation

In the literature, many authors have approached the lack of evaluation tools for visualizations and put forward techniques and models to address this issue. Early attempts include the Q-VIS Reference Model (Haase, 1998), which characterizes quality in terms of six categories or "sub-qualities": data resolution, semantic, mapping, image, presentation, and user. For each of the six sub-qualities, a weight value represents the importance of the sub-quality, and a value measures how well the visualization meets such a subquality. However, this model is a theoretical proposal that lacks empirical validation, and its usefulness in practice remains uncertain.

In a previous study (Lezcano Airaldi et al., 2022), we identified several methods that varied in their approach and focus, ranging from user questionnaires to more objective performance measures. For instance, Padda (Padda et al., 2008) proposed a set of criteria for evaluating how well visualizations aid comprehension. The authors modeled comprehension in terms of perception, cognition, and presentation qualities, each of them with its own sub-criteria.

Hung and Parsons (Hung & Parsons, 2017) introduced VisEngage, a self-assessment questionnaire designed to measure different aspects of user engagement such as aesthetics, user control, and exploration. The questionnaire consists of 22 questions rated on a seven-point Likert scale and is intended to assess engagement levels after the user interacts with a visualization.

Recent studies have also highlighted the impact of affective responses on visualization (Chen et al., 2022; Lan et al., 2022). In this context, Lan et al. (Lan et al., 2021) developed a set of heuristics to assess usability and expressiveness of infographics. Similarly, Wall (Wall et al., 2019) developed a methodology based on the value framework by Stasko (Stasko, 2014) to estimate and quantify the potential value of a visualization, i.e., its ability to provide a proper understanding of the data.

Bai et al. (Bai et al., 2009) suggest measuring the Visual Intelligence Density (VID) of a visualization to quantify the amount of useful information it provides and to evaluate its capabilities in terms of supporting decision making. In another study (Xiaoyan Bai et al., 2010), the authors proposed the Purposeful Visualization model, that covers seven assessment areas: visual representation, information presentation, psychology of the observer, information quality, visual impact, overall design style, and overall performance. Each area is further broken down into several sub-categories.

More recently, Evergreen developed an evaluation instrument for data visualizations (Evergreen, 2012). This instrument was designed to gauge the extent of graphic design use in evaluation reporting and focused on five design areas: type, alignment, graphics, and color. It uses a rating scale consisting of three response options: fully met, partly met, and not met. Although this tool is useful in assessing some aspects of data visualization design, it has limitations in terms of the number of items and accuracy of the response scale.

Many of the models above have focused on specific aspects of visualizations, such as comprehension support or user engagement, whereas others merely propose a set of criteria, but they need to be operationalized to be useful.

In the field of SE, authors have faced similar issues in assessing the quality of software products. In (Wagner et al., 2015) Wagner et al. developed the Quamoco Model, aimed at closing the gap between concrete measurements and abstract quality aspects. More recently, Siebert et al. (Siebert et al., 2022) introduced a quality model for machine learning systems, providing a systematic approach for the development and operationalization of the model.

Overall, our review highlights the need for a comprehensive and practical evaluation framework that can effectively capture the multidimensional aspects of narrative visualizations and allow authors to optimize them for their intended audience.

In this context, our proposed quality model draws upon the ISO/IEC 25010 (ISO/IEC 25010, 2011) standard for software product quality and existing models in SE, particularly the work by Siebert (Siebert et al., 2022). These frameworks were adapted to the visualization domain. This ensures that the resulting model is grounded in established research, while also addressing the unique challenges and requirements for evaluating narrative visualizations.

The model combines quantitative and qualitative approaches to provide more comprehensive results. On the one hand, it incorporates the views of experts and non-experts to determine the evaluation criteria and their level of importance via the AHP-Express method. On the other hand, we define evaluation rules based on the formative construct described in ISO/IEC 33003 (ISO/IEC 33003, 2015). By incorporating aspects of previous studies our model seeks to provide a thorough evaluation process.

3 Quality Model Construction Process

In this section, we describe the steps followed for constructing the quality model. It consists of five steps, which are performed iteratively. This process is summarized in Fig. 1. The outcome of each step provided feedback to improve the previous steps.

(1) **Define scope and application context:** Before developing the quality model, it is necessary to define the specific settings in which the model will be applied. As a first step, we identified the different user types of the model and the visualization techniques under evaluation.

(2) **Define the quality meta-model:** Here, we described the features of the quality model, that is, the basic structure used to document all quality properties of interest as well as the measures and metrics to quantify them. This led to the development of a quality metamodel which provides a systematic approach for specifying quality models.

(3) **Define assessment rules:** To ensure quality, the model must be associated with an approach that synthesizes and interprets the measurement data collected from the visualization. In this step, we specified the evaluation method applied to measure the different elements of the quality model.

(4) **Identify relevant quality properties:** To build a comprehensive quality model is important to cover as many relevant quality properties as possible. To achieve this, (Siebert et al., 2022) suggest using a systematic approach for defining and refining which properties should be modeled. For this step, we conducted a systematic mapping study (Lezcano Airaldi et al., 2022) as well as an analysis of various information visualization sources (Munzner, 2014; Nussbaumer Knaflic, 2015; Ware, 2020).

(5) **Instantiate the model for a use case:** The final step involves customizing the model for a concrete use case, based on the meta-model described in the first step. This includes adjusting the weights of the elements for a given context to reflect the distinct requirements of each scenario.

In the following sections, we elaborate on each of the steps above.

Fig. 1: Overview of the design process for the proposed quality model.

3.1 Step 1: Define scope and application context

The proposed quality model focuses on the external quality properties of a narrative visualization, as highlighted in Fig. 2, which is adapted from the ISO/IEC 25010 standard.

Fig. 2: Quality in the lifecycle according to (ISO/IEC 25010, 2011).

The figure shows the relationship between the different elements that constitute the quality of a narrative visualization. Process quality, internal properties and quality in use are outside the scope of this work, as they have been addressed by other approaches, which we discussed in Section 2.

An example of an internal property of a visualization is its source code. If developers do not adhere to best programming practices or use outdated libraries, the resulting code can be difficult to maintain. However, the visualization can still exhibit a high degree of quality.

External quality properties include aspects such as the clarity of the narrative structure, the accuracy of the data presented, and the ability for the visualization to engage the user. These qualities might support the requirements of a narrative visualization, but are generally non-exhaustive, as some features might be desirable depending on the context of use of a visualization and its environment.

3.1.1 User types

We have identified potential users of the model based on the categorization of stakeholders proposed by Amini et al. (Amini et al., 2018), and further refined our user types with considerations from the ISO/IEC 25010 standard of direct and indirect users. They are categorized into Authors, Audiences and Evaluators. These user types and their roles are summarized in Table 1.

Table 1: Users of the quality model

User type	**Description**	**Examples**
Author	Individuals who are responsible for creating narrative visualizations.	Developers, Designers, Practitioners
Audience	Individuals who interact with narrative visualizations to accomplish their goals.	Decision-makers, Students, General Public
Evaluator	Individuals who assess the quality of visualizations but do not interact with or create them.	Researchers, Editors, Supervisors

The model is primarily intended for non-expert users with basic knowledge of data visualization, so that authors, audience, or evaluators can create high-quality narrative visualizations by using the quality properties to identify the appropriate requirements.

When using the quality model, all users will assume the role of evaluators. However, their goals will vary depending on their perspective. For example, the author of a visualization might need it to be engaging so that readers interact with it, while an audience member might need to understand the visualization to make appropriate decisions. The activities that can benefit from the use of the quality model include:

(1) Identifying visualization requirements.

(2) Validating the comprehensiveness of the requirement definition.

(3) Identifying visualization design objectives.

(4) Identify the acceptance criteria for a visualization.

(5) Establishing quality property measures to support these activities.

(6) Evaluating the quality of the visualization.

(7) Identifying areas of improvement based on the results of the evaluation.

As (Wagner et al., 2015) points out, the results of the evaluation might be used either in a summative or in a formative way. On the one hand, using the results in a summative way allows evaluating the quality of a visualization and incorporating the corresponding improvements. Additionally, the results can be interpreted on their own, or they can be compared with those of other visualizations to perform analysis and identify trends. However, using the results in a formative manner might inform design decisions made during the development process.

3.1.2 Visualization techniques

A fundamental step in developing the quality model is to define what types of visualizations will be evaluated. In the field of Information Visualization, there are a wide variety of techniques with different levels of complexity, ranging from simple visualizations, such as a line charts, to more complex ones, like parallel coordinates (Johansson & Forsell, 2016) or radial bar charts (Waldner et al., 2019).

When considering which visualization techniques to include, the goal is to balance comprehensiveness and simplicity. Our aim was to include those that potential users might encounter and use more frequently, which are also familiar to them, since the intended audience is a general, non-expert user.

Borner et al. (Börner et al., 2016) found that most people have difficulty interpreting visualizations that they are not normally exposed to through media, books, or websites. Studies such as (Lee et al., 2017) and (Battle et al., 2017) have identified that line, bar, scatter, and pie charts are the most frequent in websites and visualization tools.

In light of these findings, the model was primarily designed to support the evaluation of common techniques – scatter plots, line, bar and pie charts, choropleth maps and treemaps – while still being applicable to other narrative visualizations as well. This emphasis is intended to ensure that the model can be readily applied by a wide range of individuals, including those who might not have a significant experience with complex visualizations.

It is important to note that, for more complex visualization techniques, there might be additional quality properties not included in the model. Therefore, while our model provides a comprehensive set of quality criteria for narrative visualizations, it does not fully capture the specific requirements and nuances of every visualization technique.

3.2 Step 2: Define Quality Meta-model

The central element of the model is a **property**; that is, a measurable attribute that characterizes a visualization and is indicative of its quality. As Wagner (Wagner et al., 2015) indicates, the concept of property is general, and it can be applied at different levels of abstraction.

To describe quality from an abstract level down to concrete measurements, we differentiate between two types of properties: **quality criteria** and **visualization**

properties. Quality criteria express abstract quality goals, such as memorability or comprehension. Visualization properties, in turn, refer to measurable attributes of specific visualization components.

Each visualization property is associated with one or more **applications**. An application is a concrete description of how a specific visualization property should be quantified in a particular context. Furthermore, each application has an **instrument**, which describes the concrete implementation of the application, and enables the transformation of user input into a computable value for calculating the result.

The value assigned to each option in the instrument is based on the correctness or appropriateness of the option. A value of 1 is assigned to the option that is considered the most correct or best choice for the specific application. This indicates that the option aligns well with the quality criteria and properties being evaluated. Other options may receive a value of 0 if they are deemed less appropriate or incorrect for the given application. The "N/A" option may also be selected in cases where a specific application is not suitable for a given visualization.

An example of a quality criterion is "comprehension," which includes the property of "simplicity." Simplicity can be achieved by limiting the number of series displayed. The evaluation instrument measures how well the visualization meets this condition.

Each element in the model (criteria, property, and application) carries a weight that signifies its relative importance or influence in the evaluation process. Some quality criteria might be considered more critical than others for specific use cases, thus carrying a higher weight. Similarly, certain visualization properties might have more impact on quality criteria, which is reflected in their respective weights.

Fig. 3 shows the proposed quality metamodel and the hierarchy of elements.

Fig. 3: Hierarchical structure of the quality model.

3.3 Step 3: Define Assessment Rules

To support the evaluation of the quality of a visualization, the model must be associated with an approach to synthesize and interpret measurement data collected from the visualization.

3.3.1 Quality as a Formative Construct

For the development of assessment rules, we draw upon the ISO/IEC 25010 standard, which specifies the relationship between a quality model and a quality assessment method, and the ISO/IEC 33000 family of standards which outlines best practices for the design of measurement models.

The relationship between a multidimensional construct – in this case, the quality of a narrative visualization – and its dimensions can be explained by construct specifications (Bagozzi, 2011; Diamantopoulos et al., 2008; Diamantopoulos & Winklhofer, 2001; Edwards & Bagozzi, 2000). ISO/IEC 33003 (ISO/IEC 33003, 2015) provides guidance on the choice of reflective versus formative measurement models.

In reflective models, the construct is a cause of the observed variables or indicators, and changes in the construct will be reflected in changes in the indicators. In formative models, the construct is created by the combination of the measured variables or indicators.

We selected the formative measurement model given that the quality of a narrative visualization is constructed based on the quality criteria and their corresponding properties, rather than causing them. This allows for a more accurate assessment of how each criterion impacts the overall quality of a visualization, leading to informed, actionable results. This type of model is often used when the research involves the development of a construct that is not well defined or understood (Diamantopoulos et al., 2008), which is the case of the quality of a visualization, as discussed in Section 2.

Fig. 4 shows the structure of the formative measurement model, in which constructs are represented as ovals, observed measures as rectangles, causal paths as single-headed arrows, and correlations as double-headed arrows.

Fig. 4: Relationship between a formative construct and its measures (adapted from (ISO/IEC 33003, 2015)).

The formative construct can be regarded as an index generated by the observed measures (Edwards & Bagozzi, 2000). Formative measures are not interchangeable, and each measure captures a specific aspect of the narrative visualization. Thus, omitting any measure might alter the result. The formative construct η (eta) can be represented as follows:

$$C = \gamma_1 x_1 + \cdots + \gamma_q x_q$$

where η is the construct being estimated by its formative measure x_i, and coefficient γ_i denotes the effect of measure x_i on the variable η. This equation works as the multi-attribute decision-making (MADM) method, in which a composite measure C[1] is determined by a set of formative measures weighted by the importance or priority of those measures (Bollen & Ting, 2000).

[1] A composite measure is defined as a measure derived from a combination of various measures of a multidimensional theoretical model that cannot be captured by a single measure (OECD, 2008).

MADM (Yoon & Hwang, 2011; Zeleny, 1982) is a well-established methodology for constructing preference decisions among available alternatives characterized by multiple criteria or attributes. MADM models can be divided into compensatory and non-compensatory approaches (Hwang & Yoon, 1981). In compensatory techniques, the strengths of one criterion might offset the weaknesses of another, and the exchange between criteria is allowed. In contrast, non-compensatory approaches do not allow the exchange between criteria.

Given the need to customize the quality model based on the application context by defining quality profiles for each assessment, a compensatory approach fits better with our perspective, as it allows evaluators to make trade-offs between quality criteria.

There are several compensatory MADM methods for aggregating a set of values to derive a composite measure. According to (Yoon & Hwang, 2011) and (Jung, 2013), the most suitable for a formative measurement model are the simple additive weighting (SAW) and the weighted product (WP) methods. These methods involve assigning weights to criteria to reflect their relative importance and then calculating scores for each alternative.

The Analytic Hierarchy Process (AHP) (Saaty, 1980) is a decision-making framework that can be used as a weight assigning method for SAW or WP to establish the relative importance of each criterion. It involves breaking down a problem into a hierarchy of decision criteria and alternatives, and then performing pairwise comparisons to determine their priority.

3.3.2 Simplified Analytic Hierarchy Process

One of the main difficulties in traditional AHP is maintaining the consistency rate within a given threshold (Saaty, 1980). However, this does not always depend on the experts' knowledge, as pointed out in (Tavana et al., 2021). Additionally, as more decision elements are added, the process can become slower due to the increasing number of comparisons required (Nasution et al., 2022).

Because of these issues, several improvements to the method have been proposed, such as Fuzzy AHP (Aliyev et al., 2020; Helmy et al., 2021), or Voting AHP (Liu & Hai, 2005), among others. In 2020, Leal (Leal, 2020) proposed the AHP-Express method, which simplifies the classical approach and reduces the number of comparisons from $\frac{n(n-1)}{2}$ to $(n-1)$ for n criteria, making the overall process work faster. The main advantage of AHP-Express is that it requires only one comparison for each criterion. The process involves three steps: (a) a single comparison of alternatives for each criterion, (b) calculation of the reciprocal value and its summation, and (c) normalization. The comparison values are then combined into a matrix, and the decision-making process follows the same steps as in a classic AHP.

Fig. 5 shows an example of a hierarchical structure in AHP (left) and a partial view of the quality model (right).

Fig. 5: Comparison between a standard hierarchy in AHP (on the left) and an extract of the proposed quality model (on the right).

The hierarchical structures of the quality model and the AHP method share a similar conceptual foundation. In the quality model, the criteria represent abstract aspects of quality, while the properties represent more concrete, measurable attributes. Similarly, in AHP, the higher-level criteria represent broader decision elements, while the lower-level sub-criteria represent more specific elements. This similarity reinforces the use of AHP as a suitable tool for deriving the weights of the criteria in the quality model.

3.3.3 Scale Definition

To measure the fulfillment of quality properties, we need to select an appropriate measurement scale. Stevens (Stevens, 1951) defined four types of measurement scales, from a lower to a higher level: nominal or categorical scales (attributes are only named), ordinal scales (attributes have a significant order), interval scales (equal distances correspond to equal quantities of the attribute), and ratio scales (equal distances correspond to equal quantities of the attribute where the value of zero corresponds to none of the attributes).

It is possible to perform transformations from the highest to the lowest level, such that a ratio scale can be transformed into an interval, ordinal or nominal scale; an interval scale can be transformed into an ordinal or nominal scale, and an ordinal scale can be transformed into a nominal scale. The inverse direction is not allowed. Along this line, the quality properties of the model should be defined with a ratio scale with values ranging from 0 to 1, because they are used to compute the percentage of achievement for each property. This allows for a consistent interpretation of the measurements, with 0 indicating complete absence or failure to meet the desired property, and 1 representing full achievement of a property. This approach is consistent with the recommendations of other studies (Fayers & Hand, 2002). In the case of applications, the value between 0 and 1 is obtained through its corresponding instrument.

3.3.4 Measurement Functions

The elements of the quality model can be quantified by applying a **measurement function**, that is, a sequence of operations that combine quality measurement elements. The quality measurement elements are, in turn, lower-order measures, up to the data obtained directly from the visualization, as illustrated in Fig. 6.

Fig. 6: Relationship between the quality model and the measurement model.

The model identifies four types of functions, from the highest to the lowest level of the hierarchy:

(1) The **Quality** measurement function, which aggregates the values of the quality criteria (V_{C1}, V_{C2}, …, V_{Cn}) and is formulated as follows:

$$fQ\left(C_{1,}C_{2,}\dots C_{n,}\right) = W_{C1} \cdot V_{C1} + W_{C2} \cdot V_{C2} + \cdots + W_{Cn} \cdot V_{Cn}$$

(2) The **Criteria** measurement function, which aggregates the values of the quality properties (V_{P1}, V_{P2}, …, V_{Pn}) and is formulated as follows:

$$fC\left(P_{1,}P_{2,}\dots P_{n,}\right) = W_{P1} \cdot V_{P1} + W_{P2} \cdot V_{P2} + \cdots + W_{Pn} \cdot V_{Pn}$$

(3) The **Property** measurement function, which aggregates the values of the applications (V_{A1}, V_{A2}, …, V_{An}) and is formulated as follows:

$$fP\left(A_{1,}A_{2,}\dots A_{n,}\right) = W_{A1} \cdot V_{A1} + W_{A2} \cdot V_{A2} + \cdots + W_{An} \cdot V_{An}$$

(4) The **Application** measurement function, which takes the data obtained directly from the visualization via instruments.

These functions are based on the Simple Additive Weighting (SAW) technique, and the weights, denoted by W, are calculated using the Analytic Hierarchy Process. This reflects the causal relationship between the criteria, properties and applications of narrative visualization, indicating the order in which they are connected. Therefore, we have that:

- The quality of a visualization is determined by the quality of its criteria.
- The quality of a criterion is determined by the quality of its associated properties.

- The quality of a property is determined by the quality of its associated applications.
- The quality of an application is determined by the direct values obtained from the visualization through the instruments.

3.4 Step 4: Identify relevant quality properties

To select the set of quality properties, we conducted a Systematic Mapping Study (Lezcano Airaldi et al., 2022) and analyzed various information visualization sources (Munzner, 2014; Nussbaumer Knaflic, 2015; Ware, 2020). In this section we describe the quality criteria, visualization properties, applications, and instruments that constitute the model.

3.4.1 Quality criteria

The proposed quality criteria are comprehension, engagement, information, memorability, and usability, which are described below.

C01: Comprehension. The goal of an information visualization is to facilitate knowledge extraction (Card et al., 1999). To achieve this, visual representations should exploit the human perceptual capabilities. Several authors have studied graph comprehension. They define it as “the ability to understand and interpret a graph” (Friel et al., 2001) and propose a three-level framework(Bertin, 2010; Carswell, 1992; Wainer, 1992): the elementary level, where the user can find a specific value in the graph; the intermediate level, where the user is able to identify trends and relationships, and the advanced level, where the user can read beyond what is presented in the graph. Some charts might be easier to understand than others. This is due to the fact that comprehension is not only based on the characteristics of the graph itself but is also influenced by how these characteristics interact with the knowledge and goals of the user (Shah & Hoeffner, 2002). This is closely related to the concept of visual literacy, which refers to “the ability to extract, interpret, and give meaning to the information presented in a visualization” (Lee et al., 2017). As the value and availability of data continues to grow, so does the need for comprehension. Therefore, it is key to assess how well users understand a visualization and the insights they gain during its consumption.

C02: Engagement. According to Bach et al. , engagement refers to the degree to which a viewer is involved and interested (Bach et al., 2018) in the story being told through the visualization. This can include the users’ emotional response, their level of attention and focus, and the extent to which they actively interact with the visualization (such as clicking or exploring parts of the story). Moreover, if we consider engagement as "the time the user spends with the content during a given period of time" (Cherubini & Nielsen, 2016), interaction becomes a key factor in achieving a greater engagement. With today’s explosion of information, maintaining the attention of the audience becomes a challenge [87]; thus, engagement is a fundamental aspect to assess the effectiveness of narrative visualizations.

C03: Information. To develop narrative visualizations, it is crucial to understand the source of the data used, the methodological decisions, and the potential interests of the authors (Diakopoulos, 2018). This criterion refers to the information source and the ethical aspects behind it, such as who produced the data and with what intention, the degree of accuracy and completeness, among others. While it is important to provide transparency by citing or linking to data sources (Hullman & Diakopoulos, 2011), there is also the ethical question of whether to use a dataset and how the data was obtained. In addition, the graphical mapping and data transformations can influence its interpretation. For example, in some cases it is appropriate to start an axis at zero (such as in bar charts), because values are read according to the length of the bar), while in others (such as scatterplots or line charts), an axis starting at a value other than zero is acceptable and can even clarify variations in the data. In this context, Tufte (Tufte, 2001) proposes the concept of "graphical integrity," which states that a graph is not distorted when the visual representation of the data matches its numerical representation. Therefore, is important to learn to identify common distortion techniques in order to avoid them.

C04: Memorability. Memorability is the ability to retain and recall information (Brown et al., 1977). This criterion refers to a basic cognitive concept that has direct implications for the design of narrative visualizations and influences higher cognitive functions such as comprehension (Borkin et al., 2016). To create a memorable visualization, it is necessary to understand which elements are more likely to stick in users' minds. It is not desirable to remember incorrect information (also known as *chartjunk*) but rather the relevant aspects of the data and patterns that the visualization author intends to convey. It is essential to consider factors such as the organization and sequence of the different components of the visualization, as well as the use of graphic resources such as colors and shapes.

C05: Usability. This criterion refers to the degree to which a visualization can be used by specific users to achieve their objectives with effectiveness, efficiency, and perceived satisfaction in a specific context of use (ISO 9241-11, 2018). If the overall visualization design helps users achieve their objectives with less effort, they are likely to react positively. Otherwise, negative responses can be generated (Norman, 2005). Lan et al. (Lan et al., 2021) consider accessibility as one of the main aspects to increase usability. Several studies have also shown the relationship between aesthetics and ease of use (Cawthon & Moere, 2007; Kurosu & Kashimura, 1995; Quispel et al., 2016).

Furthermore, aspects related to interaction and user control should also be considered, as they contribute to both the engagement of a visualization and an effective information consumption.

3.4.2 Visualization properties

Building upon the findings of the SMS (Lezcano Airaldi et al., 2022), we identified 17 quality properties, which are presented in Table 2. They were organized in a sequence that reflects the typical process of constructing a visualization. These properties cover the encoding and graphic mapping of information, the quality of the underlying data, as well as the design and narrative elements.

Table 2: Visualization properties

ID	Property	Description
P01	Accuracy	Use of an appropriate scale to represent the data, avoiding any distortion or alteration that could mislead or confuse the viewer.
P02	Saliency	The extent to which the visualization associates important information with prominent visual features. This property implies that the importance of the data should match the effectiveness of the visual channel, based on the ranking presented in (Munzner, 2014) .
P03	Color	Strategic use of color, stated as a dominant preattentive attribute[2]. It involves consideration of the number of colors as well as their purpose.
P04	Layout	The arrangement of visual elements within a narrative visualization, such

[2] Preattentive processing is the automatic and rapid processing of visual stimuli that occurs without conscious effort. It is an important aspect of visual perception and can be leveraged to enhance visual communication (Healey et al., 1993; Nussbaumer Knaflic, 2015).

ID	Property	Description
		as aspect ratio, contrast, and white space, which can influence how viewers interpret the visualization.
P05	Visual clutter	The number of unnecessary, obtrusive elements that do not add new information and could confuse the viewer, also known as “chart junk” (Bateman et al., 2010).
P06	Focus	The use of visual cues such as color, contrast, or annotations to highlight important patterns or pieces of information and direct the attention of the user.
P07	Accessibility	The degree to which a visualization is designed to be comprehensible by a diverse range of users, including those with visual or hearing impairments, or varying levels of technical proficiency.
P08	Reliability	The degree to which the source and quality of the data used in a visualization is transparent and well-documented. This includes information about where the data came from, how it was collected, and any potential limitations or biases that may be present.
P09	Redundancy	The extent to which the same the same information is presented in multiple ways within a single visualization. While too much redundancy can lead to clutter, a certain degree can facilitate comprehension of complex concepts and reinforce key takeaways.
P10	Integrity	Representation of information in a truthful, accurate and complete manner, without any intentional or unintentional omissions or obscurations
P11	Clarity	The ability of a narrative visualization to convey complex ideas in a clear and concise manner, avoiding unnecessary or distracting elements.

ID	Property	Description
P12	Context	The degree to which the visualization provides relevant background information and supporting details that help viewers understand the data even with little prior knowledge or expertise in the subject matter.
P13	Narrative	The use of storytelling techniques to convey the intended message, using a structured and coherent sequence of events, and visual cues such as color, typography, and transitions to guide and engage the viewer.
P14	User guidance	The degree to which the visualization provides clear and sufficient instructions or cues to the viewer on how to navigate and interpret the data being presented, particularly for multiple-panel visualizations.
P15	Metonymy	The use of visual elements that are representative of the data or concepts being presented. It involves using an intuitive graphic mapping that goes in line with preexisting mental constructs i.e., left = past, right = future.
P16	Consistency	The degree to which the visual elements of a visualization are uniform and cohesive. This can be accomplished in a variety of ways, such as using consistent color schemes, fonts, and iconography throughout the visualization.
P17	Interactivity	The degree to which the user is able to interact with and manipulate the data being presented. Interactive features can include a range of techniques, such as zooming, panning, filtering, and highlighting, among others.

3.4.3 Applications

Lastly, we identified 55 applications, which form the basis of evaluation. An excerpt of the applications is presented in Table 3. Each application has an instrument with a series of options, that allow the user to enter evaluation data. The complete list of applications and their corresponding instruments can be found in the Supplementary Materials[3].

Table 3: Applications

ID	Application	Description	Options	Value
AP03	Axes titles	Axes should be named appropriately.	Descriptive axes titles	1
			Generic axes titles	0.5
			No axes titles	0
AP10	Data labels	Data should be labeled directly.	Data is labeled directly	1
			Data is labeled by legends	0.5
			Data is labeled both ways	0.25
			Data is not labeled	0
AP19	Limited colors	The number of colors should be within working memory limits.	Up to 2 colors	1
			Between 3 and 5 colors	0.5
			More than 5 colors	0
AP23	Main message	Key takeaways should be made explicit in titles and subtitles.	Main message in title	1
			Main message not in title	0
			N/A	
AP32	Gridlines	Grids should be usefully visible, unobtrusive.	Gridlines are not used	1
			Gridlines are used sparingly	0.5

[3] Supplementary Materials

ID	Application	Description	Options	Value
			Gridlines are moderately used	0.2
			Gridlines are heavily used	0
			N/A	
AP46	Messages and summaries	Text elements should be used to summarize and explain information.	Fulfilled	1
			Partially fulfilled	0.5
			Unfulfilled	0
			N/A	
AP47	Illustrations	Illustrations, if present, should be topic-relevant	Illustrations are related to the concepts	1
			Illustrations are decorative	0.5
			Illustrations are irrelevant or confusing	0
			N/A	

As mentioned in Section 3.2, each application has an instrument, which consists of a list of options and values. The options were derived from the literature and designed to be as tangible and precise as possible. To account for complexities that might make the evaluation exceedingly challenging or impractical, we have incorporated an unidimensional scale that offers three classifications: fulfilled, "partially fulfilled", and "unfulfilled". This scale is intended to provide a more manageable means to assess and quantify the implementation of the application.

3.4.4 Relations between quality properties

Based on the previous definitions, Tables 3 and 4 summarize the quality criteria and properties. They indicate the corresponding lower-order elements associated with each property and outline their specific measurement functions.

Table 4: Quality criteria and properties

ID	**Description**	**Quality Properties**	**Measurement Function**
C01	Comprehension	P01; P02; P03; P05; P06; P07; P11; P12; P13	$f\mathrm{C}_{01}=W_{\mathrm{P01}}\cdot V_{\mathrm{P01}}+W_{\mathrm{P02}}\cdot V_{\mathrm{P02}}+W_{\mathrm{P03}}\cdot V_{\mathrm{P03}}+W_{\mathrm{P05}}\cdot V_{\mathrm{P05}}+W_{\mathrm{P06}}\cdot V_{\mathrm{P06}}+W_{\mathrm{P07}}\cdot V_{\mathrm{P07}}+W_{\mathrm{P11}}\cdot V_{\mathrm{P11}}+W_{\mathrm{P12}}\cdot V_{\mathrm{P12}}+W_{\mathrm{P13}}\cdot V_{\mathrm{P13}}$
C02	Engagement	P03; P05; P11; P12; P13; P14; P16; P17	$f\mathrm{C}_{02}=W_{\mathrm{P03}}\cdot V_{\mathrm{P03}}+W_{\mathrm{P05}}\cdot V_{\mathrm{P05}}+W_{\mathrm{P11}}\cdot V_{\mathrm{P11}}+W_{\mathrm{P12}}\cdot V_{\mathrm{P12}}+W_{\mathrm{P13}}\cdot V_{\mathrm{P13}}+W_{\mathrm{P14}}\cdot V_{\mathrm{P14}}+W_{\mathrm{P16}}\cdot V_{\mathrm{P16}}+W_{\mathrm{P17}}\cdot V_{\mathrm{P17}}$
C03	Information	P01; P04; P06, P08; P09; P10; P12	$f\mathrm{C}_{03}=W_{\mathrm{P01}}\cdot V_{\mathrm{P01}}+W_{\mathrm{P04}}\cdot V_{\mathrm{P04}}+W_{\mathrm{P06}}\cdot V_{\mathrm{P06}}+W_{\mathrm{P08}}\cdot V_{\mathrm{P08}}+W_{\mathrm{P09}}\cdot V_{\mathrm{P09}}+W_{\mathrm{P10}}\cdot V_{\mathrm{P10}}+W_{\mathrm{P12}}\cdot V_{\mathrm{P12}}$
C04	Memorability	P03; P06; P09; P11; P12; P13; P14; P15	$f\mathrm{C}_{04}=W_{\mathrm{P03}}\cdot V_{\mathrm{P03}}+W_{\mathrm{P06}}\cdot V_{\mathrm{P06}}+W_{\mathrm{P09}}\cdot V_{\mathrm{P09}}+W_{\mathrm{P11}}\cdot V_{\mathrm{P11}}+W_{\mathrm{P12}}\cdot V_{\mathrm{P12}}+W_{\mathrm{P13}}\cdot V_{\mathrm{P13}}+W_{\mathrm{P14}}\cdot V_{\mathrm{P14}}+W_{\mathrm{P15}}\cdot V_{\mathrm{P15}}$
C05	Usability	P02; P04; P07; P08; P11; P14; P16; P17	$f\mathrm{C}_{05}=W_{\mathrm{P02}}\cdot V_{\mathrm{P02}}+W_{\mathrm{P04}}\cdot V_{\mathrm{P04}}+W_{\mathrm{P07}}\cdot V_{\mathrm{P07}}+W_{\mathrm{P08}}\cdot V_{\mathrm{P08}}+W_{\mathrm{P11}}\cdot V_{\mathrm{P11}}+W_{\mathrm{P14}}\cdot V_{\mathrm{P14}}+W_{\mathrm{P16}}\cdot V_{\mathrm{P16}}+W_{\mathrm{P17}}\cdot V_{\mathrm{P17}}$

Table 5: Quality properties and applications

ID	**Description**	**Applications**	**Measurement Function**
P01	Accuracy	AP01; AP03; AP04; AP05; AP08	$f\mathrm{P}_{01}=W_{\mathrm{A01}}\cdot V_{\mathrm{A01}}+W_{\mathrm{A03}}\cdot V_{\mathrm{A03}}+W_{\mathrm{A04}}\cdot V_{\mathrm{A04}}+W_{\mathrm{A05}}\cdot V_{\mathrm{A05}}+W_{\mathrm{A08}}\cdot V_{\mathrm{A08}}$
P02	Saliency	AP14; AP16; AP20	$f\mathrm{P}_{02}=W_{\mathrm{A14}}\cdot V_{\mathrm{A14}}+W_{\mathrm{A16}}\cdot V_{\mathrm{A16}}+W_{\mathrm{A20}}\cdot V_{\mathrm{A20}}$
P03	Color	AP18; AP19; AP20; AP21; AP22	$f\mathrm{P}_{03}=W_{\mathrm{A18}}\cdot V_{\mathrm{A18}}+W_{\mathrm{A19}}\cdot V_{\mathrm{A19}}+W_{\mathrm{A20}}\cdot V_{\mathrm{A20}}+W_{\mathrm{A21}}\cdot V_{\mathrm{A21}}+W_{\mathrm{A22}}\cdot V_{\mathrm{A22}}$
P04	Layout	AP08; AP28; AP29; AP35, AP36, AP37, AP38, AP39	$f\mathrm{P}_{04}=W_{\mathrm{A08}}\cdot V_{\mathrm{A08}}+W_{\mathrm{A28}}\cdot V_{\mathrm{A28}}+W_{\mathrm{A29}}\cdot V_{\mathrm{A29}}+W_{\mathrm{A35}}\cdot V_{\mathrm{A35}}+W_{\mathrm{A36}}\cdot V_{\mathrm{A36}}+W_{\mathrm{A37}}\cdot V_{\mathrm{A37}}+W_{\mathrm{A38}}\cdot V_{\mathrm{A38}}+W_{\mathrm{A39}}\cdot V_{\mathrm{A39}}$
P05	Visual clutter	AP06; AP19; AP28; AP29; AP32; AP33; AP34; AP35, AP36, AP37, AP38, AP39	$f\mathrm{P}_{05}=W_{\mathrm{A06}}\cdot V_{\mathrm{A06}}+W_{\mathrm{A19}}\cdot V_{\mathrm{A19}}+W_{\mathrm{A28}}\cdot V_{\mathrm{A28}}+W_{\mathrm{A29}}\cdot V_{\mathrm{A29}}+W_{\mathrm{A32}}\cdot V_{\mathrm{A32}}+W_{\mathrm{A33}}\cdot V_{\mathrm{A33}}+W_{\mathrm{A34}}\cdot V_{\mathrm{A34}}+W_{\mathrm{A35}}\cdot V_{\mathrm{A35}}+W_{\mathrm{A36}}\cdot V_{\mathrm{A36}}+W_{\mathrm{A37}}\cdot V_{\mathrm{A37}}+W_{\mathrm{A38}}\cdot V_{\mathrm{A38}}+W_{\mathrm{A39}}\cdot V_{\mathrm{A39}}$
P06	Focus	AP16; AP20, AP34	$f\mathrm{P}_{06}=W_{\mathrm{A16}}\cdot V_{\mathrm{A16}}+W_{\mathrm{A20}}\cdot V_{\mathrm{A20}}+W_{\mathrm{A34}}\cdot V_{\mathrm{A34}}$
P07	Accessibility	AP10; AP15; AP21; AP22; AP23; AP26, AP27; AP29; AP31; AP35, AP36, AP37, AP38, AP39; AP55	$f\mathrm{P}_{07}=W_{\mathrm{A10}}\cdot V_{\mathrm{A10}}+W_{\mathrm{A15}}\cdot V_{\mathrm{A15}}+W_{\mathrm{A21}}\cdot V_{\mathrm{A21}}+W_{\mathrm{A22}}\cdot V_{\mathrm{A22}}+W_{\mathrm{A23}}\cdot V_{\mathrm{A23}}+W_{\mathrm{A26}}\cdot V_{\mathrm{A26}}+W_{\mathrm{A27}}\cdot V_{\mathrm{A27}}+W_{\mathrm{A29}}\cdot V_{\mathrm{A29}}+W_{\mathrm{A31}}\cdot V_{\mathrm{A31}}+W_{\mathrm{A35}}\cdot V_{\mathrm{A35}}+W_{\mathrm{A36}}\cdot V_{\mathrm{A36}}+W_{\mathrm{A37}}\cdot V_{\mathrm{A37}}+W_{\mathrm{A38}}\cdot V_{\mathrm{A38}}+W_{\mathrm{A39}}\cdot V_{\mathrm{A39}}+W_{\mathrm{A55}}\cdot V_{\mathrm{A55}}$
P08	Reliability	AP07; AP09; AP10; AP11; AP12; AP13; AP40; AP41; AP42; AP43; AP44	$f\mathrm{P}_{08}=W_{\mathrm{A07}}\cdot V_{\mathrm{A07}}+W_{\mathrm{A09}}\cdot V_{\mathrm{A09}}+W_{\mathrm{A10}}\cdot V_{\mathrm{A10}}+W_{\mathrm{A11}}\cdot V_{\mathrm{A11}}+W_{\mathrm{A12}}\cdot V_{\mathrm{A12}}+W_{\mathrm{A13}}\cdot V_{\mathrm{A13}}+W_{\mathrm{A40}}\cdot V_{\mathrm{A40}}+W_{\mathrm{A41}}\cdot V_{\mathrm{A41}}+W_{\mathrm{A42}}\cdot V_{\mathrm{A42}}+W_{\mathrm{A43}}\cdot V_{\mathrm{A43}}+W_{\mathrm{A44}}\cdot V_{\mathrm{A44}}$

P09	Redundancy	AP03, AP10, AP15, AP16, AP21, AP23, AP30, AP46, AP47	$f\mathrm{P}_{09}=W_{\mathrm{A03}}\cdot V_{\mathrm{A03}}+W_{\mathrm{A10}}\cdot V_{\mathrm{A10}}+W_{\mathrm{A15}}\cdot V_{\mathrm{A15}}+W_{\mathrm{A16}}\cdot V_{\mathrm{A16}}+W_{\mathrm{A21}}\cdot V_{\mathrm{A21}}+W_{\mathrm{A23}}\cdot V_{\mathrm{A23}}+W_{\mathrm{A30}}\cdot V_{\mathrm{A30}}+W_{\mathrm{A46}}\cdot V_{\mathrm{A46}}+W_{\mathrm{A47}}\cdot V_{\mathrm{A47}}$
P10	Integrity	AP02, AP07, AP33, AP40, AP42, AP43, AP44	$f\mathrm{P}_{10}=W_{\mathrm{A02}}\cdot V_{\mathrm{A02}}+W_{\mathrm{A07}}\cdot V_{\mathrm{A07}}+W_{\mathrm{A33}}\cdot V_{\mathrm{A33}}+W_{\mathrm{A40}}\cdot V_{\mathrm{A40}}+W_{\mathrm{A42}}\cdot V_{\mathrm{A42}}+W_{\mathrm{A43}}\cdot V_{\mathrm{A43}}+W_{\mathrm{A44}}\cdot V_{\mathrm{A44}}$
P11	Clarity	AP06, AP07, AP17, AP23, AP24, AP35, AP36, AP37, AP38, AP39, AP46, AP47	$f\mathrm{P}_{11}=W_{\mathrm{A06}}\cdot V_{\mathrm{A06}}+W_{\mathrm{A07}}\cdot V_{\mathrm{A07}}+W_{\mathrm{A17}}\cdot V_{\mathrm{A17}}+W_{\mathrm{A23}}\cdot V_{\mathrm{A23}}+W_{\mathrm{A24}}\cdot V_{\mathrm{A24}}+W_{\mathrm{A35}}\cdot V_{\mathrm{A35}}+W_{\mathrm{A36}}\cdot V_{\mathrm{A36}}+W_{\mathrm{A37}}\cdot V_{\mathrm{A37}}+W_{\mathrm{A38}}\cdot V_{\mathrm{A38}}+W_{\mathrm{A39}}\cdot V_{\mathrm{A39}}+W_{\mathrm{A46}}\cdot V_{\mathrm{A46}}+W_{\mathrm{A47}}\cdot V_{\mathrm{A47}}$
P12	Context	AP03, AP23, AP25, AP30, AP44, AP46, AP47, AP50	$f\mathrm{P}_{12}=W_{\mathrm{A03}}\cdot V_{\mathrm{A03}}+W_{\mathrm{A23}}\cdot V_{\mathrm{A23}}+W_{\mathrm{A25}}\cdot V_{\mathrm{A25}}+W_{\mathrm{A30}}\cdot V_{\mathrm{A30}}+W_{\mathrm{A44}}\cdot V_{\mathrm{A44}}+W_{\mathrm{A46}}\cdot V_{\mathrm{A46}}+W_{\mathrm{A47}}\cdot V_{\mathrm{A47}}+W_{\mathrm{A50}}\cdot V_{\mathrm{A50}}$
P13	Narrative	AP16, AP21, AP23, AP25, AP30, AP45, AP46, AP47, AP50, AP51, AP52, AP54	$f\mathrm{P}_{13}=W_{\mathrm{A16}}\cdot V_{\mathrm{A16}}+W_{\mathrm{A21}}\cdot V_{\mathrm{A21}}+W_{\mathrm{A23}}\cdot V_{\mathrm{A23}}+W_{\mathrm{A25}}\cdot V_{\mathrm{A25}}+W_{\mathrm{A30}}\cdot V_{\mathrm{A30}}+W_{\mathrm{A45}}\cdot V_{\mathrm{A45}}+W_{\mathrm{A46}}\cdot V_{\mathrm{A46}}+W_{\mathrm{A47}}\cdot V_{\mathrm{A47}}+W_{\mathrm{A50}}\cdot V_{\mathrm{A50}}+W_{\mathrm{A51}}\cdot V_{\mathrm{A51}}+W_{\mathrm{A52}}\cdot V_{\mathrm{A52}}+W_{\mathrm{A54}}\cdot V_{\mathrm{A54}}$
P14	User guidance	AP11, AP30, AP34, AP45, AP46, AP48, AP49, AP50, AP51, AP54	$f\mathrm{P}_{14}=W_{\mathrm{A11}}\cdot V_{\mathrm{A11}}+W_{\mathrm{A30}}\cdot V_{\mathrm{A30}}+W_{\mathrm{A34}}\cdot V_{\mathrm{A34}}+W_{\mathrm{A45}}\cdot V_{\mathrm{A45}}+W_{\mathrm{A46}}\cdot V_{\mathrm{A46}}+W_{\mathrm{A48}}\cdot V_{\mathrm{A48}}+W_{\mathrm{A49}}\cdot V_{\mathrm{A49}}+W_{\mathrm{A50}}\cdot V_{\mathrm{A50}}+W_{\mathrm{A51}}\cdot V_{\mathrm{A51}}+W_{\mathrm{A54}}\cdot V_{\mathrm{A54}}$
P15	Metonimy	AP15, AP16, AP21, AP47	$f\mathrm{P}_{15}=W_{\mathrm{A15}}\cdot V_{\mathrm{A15}}+W_{\mathrm{A16}}\cdot V_{\mathrm{A16}}+W_{\mathrm{A21}}\cdot V_{\mathrm{A21}}+W_{\mathrm{A47}}\cdot V_{\mathrm{A47}}$
P16	Consistency	AP14, AP16, AP17, AP18, AP19	$f\mathrm{P}_{16}=W_{\mathrm{A14}}\cdot V_{\mathrm{A14}}+W_{\mathrm{A16}}\cdot V_{\mathrm{A16}}+W_{\mathrm{A17}}\cdot V_{\mathrm{A17}}+W_{\mathrm{A18}}\cdot V_{\mathrm{A18}}+W_{\mathrm{A19}}\cdot V_{\mathrm{A19}}$
P17	Interactivity	AP48, AP49, AP52, AP53, AP54	$f\mathrm{P}_{17}=W_{\mathrm{A48}}\cdot V_{\mathrm{A48}}+W_{\mathrm{A49}}\cdot V_{\mathrm{A49}}+W_{\mathrm{A52}}\cdot V_{\mathrm{A52}}$

	$+W_{A53}\cdot V_{A53}+W_{A54}\cdot V_{A54}$

In our evaluation model, the weights of properties and applications are distributed evenly. For instance, if a property includes three applications, each will have a weight of 1/3. In cases where a property includes one or more optional applications – which could or could not be included in a visualization, the weight of these optional applications is re-distributed among the remaining applications for balance.

For instance, let us consider Property P01 – Accuracy, which includes applications AP01, AP03, AP04, AP05, and AP08. If the visualization in question does not incorporate a scale break (AP05), this application becomes non-applicable and cannot be evaluated. As a result, its weight must be redistributed among the remaining four applications. Thus, instead of having each application carrying a weight of 1/5, each one would now have a weight of 1/4. This adjustment ensures that the sum of the weights of all applications remains equal to 1, regardless of the applications being applicable in a particular instance.

3.5 Step 5: Instantiate the model for a use case

While the metamodel provides a foundation for evaluation, not every element will hold the same relevance across all contexts; the importance of the quality properties will depend on the needs and requirements of the stakeholders. Therefore, the model must be customized prior to its use to reflect the specifics of each case and identify significant properties.

For this purpose, we propose the concept of **quality profiles**. A quality profile as a set of properties that are tailored to a particular context or use case (Morisio et al., 2002). For example, a standard quality profile might consider all quality criteria equally relevant, while a decision-making profile could weigh specific criteria, such as comprehension and information above all others.

By defining a quality profile, evaluators can more effectively assess the quality of a narrative visualization and identify areas for improvement. Quality profiles can also help developers and designers to ensure that they are addressing the needs of their stakeholders and creating a visualization that meets their specific requirements.

To establish quality profiles, we followed a goal-oriented approach in which each profile is determined according to the purpose of the visualization. Table 6 presents a summary of the proposed quality profiles.

Table 6: Quality Profiles

Quality Profile	**Description**	**Priority Criteria**
Standard	General-use profile. Provides comprehensive coverage for various contexts.	All quality criteria
Decision-making	Prioritizes the properties that aid understanding in decision-making processes.	Comprehension, Information
Education	Emphasizes criteria that enhances learning experiences.	Comprehension, Engagement, Memorability
Communication	Prioritizes properties improving clarity and succinctness of the message.	Comprehension, Engagement, Memorability
Exploration	Suitable for scenarios where users need to	Usability,

Quality Profile	Description	Priority Criteria
	explore and discover patterns in data.	Information

It is important to note that the priority criteria outlined for each profile are presented as initial suggestions and can be further refined to align with specific requirements. The weights of the criteria are assigned using the AHP Express method based on their relative importance within the context of the intended purpose. In addition, the model will allow users to define custom profiles by adjusting the weights of the relevant criteria based on their objective and needs.

4 Small-scale Evaluation

To test the feasibility of the model, we conducted a small-scale evaluation following the guidelines by Wohlin (Wohlin & Rainer, 2022) and Robson (Robson, 2017). This type of evaluation is appropriate as it is intended for preliminary assessments and involve a limited number of data sources (Robson, 2017).

4.1 Research Questions

The following research questions were formulated to assess the performance of the quality model:

RQ1. How well does the quality model capture the important aspects of narrative visualizations?

RQ2. What are the strengths and limitations of the quality model in assessing the quality of narrative visualizations?

RQ3. What is the reliability of the quality model in terms of producing consistent evaluations by different evaluators?

RQ4. What is the perceived level of difficulty when applying the evaluation model?

Questions 1 to 3 aim to provide insights into the prospective value of the model from different perspectives, while question 4 focuses on user experience and aims to inform improvements for future evaluations.

4.2 Evaluation

4.2.1 Quality Profile Set Up

To determine the weights of the criteria for the Standard quality profile, we performed the AHP Express method with 51 participants from diverse backgrounds and perspectives. Of the 51 participants, 18 were male and 33 were female, with ages ranging from 20 to 35 years old. In a collaborative classroom setting, participants shared their views on the importance of each criterion.

While the AHP Express method assumes evaluation consistency among decision-makers, we took precautions to minimize potential inconsistencies. Clear instructions were provided to the participants, and efforts were made to ensure a shared understanding of the criteria. Additionally, discussions and deliberations were encouraged to resolve any discrepancies or conflicting opinions.

The resulting weights for each quality criterion are listed in Table 7.

Table 7: Criteria weight for the Standard quality profile

Criteria ID	**Description**	**Criteria Weight**
C01	Comprehension	0,23
C02	Engagement	0,15

Criteria ID	**Description**	**Criteria Weight**
C03	Information	0,18
C04	Memorability	0,26
C05	Usability	0,18

These weights served as the basis for conducting the evaluation of the quality model, which is described in subsequent sections.

4.2.2 Dataset selection

To conduct the evaluation, we selected a set of static as well as interactive visualizations. For static visualizations, we used the MASSVIS Dataset (*MASSVIS - Massachusetts (Massive) Visualization Dataset*, n.d.), a diverse collection of visualizations that can help assess the generalizability of the model across different visualization types. This dataset is openly available and widely used in the research community, which increases the reproducibility and comparability of our results. We focused particularly on the Government and News visualizations, due to their narrative nature, as the Science subset often requires specialized knowledge and lacks the storytelling component, which is the focus of our model. To ensure a representative sample, we used a Python script to select a random subset of visualizations in each group.

For interactive visualizations, we drew on the example galleries of visualization libraries and online tools, as well as on several visualization portals known for their storytelling components. We decided to include interactive visualizations given that it is a central aspect of visual data analysis, as discussed in (Saket et al., 2018). These sources offer a diverse range of interactive visualizations and provide a convenient way to assess the applicability of our model.

4.2.3 Evaluator Background

Three evaluators participated in the assessment process. Each evaluator had distinct levels of experience in the field, with Evaluator 1 having extensive background knowledge regarding Information Visualization, Evaluator 2 having intermediate experience, and Evaluator 3 having basic to no prior experience. The ages of the evaluators ranged from 26 to 33 years.

By involving evaluators with different backgrounds and levels of experience, we aimed to account for a wide range of perspectives and potential biases. This approach sought to enhance the validity and reliability of the evaluation results, as it considered the viewpoints of evaluators with varying levels of familiarity and expertise in the domain.

4.2.4 Procedure

To conduct the evaluation, we set up a spreadsheet with the criteria, properties, and applications, and the corresponding formulas to calculate their value. We used the Standard quality profile, with the weights derived at a previous stage, as explained in Section 4.2.1.

At the beginning of the evaluation process, evaluators were introduced to the Quality Model and its components. An investigator explained the purpose of the study and guided the evaluators through the key aspects of the model. The evaluators received a brief in-person training that covered an overview of how to conduct an evaluation, how the spreadsheet worked and how the results were calculated. They were encouraged to ask questions about any areas requiring clarification.

For each visualization, the evaluators responded to a series of questions (referred to as “applications”) by selecting options from a dropdown list. After completing an evaluation, the corresponding score was calculated and recorded in a separate sheet. Once all evaluators completed the evaluation process, the individual scores were aggregated and processed in an additional spreadsheet.

After completing the evaluations, participants filled in a survey designed to gather feedback about the model. The survey utilized a 5-point Likert scale and to capture the evaluators’ perspectives on various aspects of the quality model. This allowed for a quantitative assessment of their satisfaction and perception of the model. Furthermore, the survey included comment sections to encourage participants to provide additional qualitative feedback, allowing for a more comprehensive understanding of their experiences and suggestions for improvement.

All materials used for this study, including the Python script, evaluation spreadsheets and the survey form can be found in the Supplementary Materials.

4.3 Results

In this section we present the results for the small-scale evaluation, addressing each of the research questions.

4.3.1 RQ1: How well does the quality model capture the important aspects of narrative visualizations?

Table 8 provides a summary of the evaluation results. Each row corresponds to a specific visualization, along with its corresponding quality scores from the three evaluators. The remaining columns contain the average quality score and standard deviation. Highlighted cells indicate those visualizations that received an above average score.

Table 8: Evaluation results of the visualizations using the quality model.

ID_Vis	Evaluator 1	Evaluator 2	Evaluator 3	Average Score	Standard Dev
01_news	0,97	0,82	0,73	0,84	0,12
02_news	0,83	0,82	0,83	0,83	0,01
03_news	0,68	0,63	0,69	0,67	0,03
04_news	0,69	0,65	0,62	0,65	0,04
05_news	0,84	0,76	0,89	0,83	0,07
06_news	0,93	0,84	0,87	0,88	0,05
01_gov	0,55	0,57	0,53	0,55	0,02
02_gov	0,54	0,48	0,45	0,49	0,05
03_gov	0,50	0,56	0,50	0,52	0,03
04_gov	0,51	0,40	0,39	0,43	0,07
05_gov	0,48	0,43	0,46	0,46	0,03
06_gov	0,85	0,80	0,90	0,85	0,05
01_int	0,91	0,80	0,87	0,86	0,06
02_int	0,98	1,00	0,98	0,99	0,01
03_int	0,94	0,84	0,88	0,89	0,05
04_int	0,97	0,95	0,95	0,96	0,01
05_int	0,99	0,97	1,00	0,99	0,02
06_int	0,97	0,95	0,92	0,95	0,03

When evaluating the effectiveness of the quality model in capturing important aspects of narrative visualizations, several observations were made. First, the model proved comprehensive in its coverage, leaving no significant features unassessed. This suggests that the model can successfully encapsulate the inherent diversity and complexity of various types of visualizations.

The model makes particular emphasis on narrative aspects, such as the amount of context provided in the visualization, or the use and purpose of color and textual elements. Therefore, visualizations that did not comply with these aspects tended to receive lower scores, which suggests that the model is working as intended – distinguishing between narrative visualizations from more traditional ones.

In terms of scope, the evaluation model was applied to a diverse set of visualizations, ranging from basic line charts to maps and multi-panel visuals. The model maintained a consistent performance across all formats, underlining its versatility to evaluate a broad range of visualization types.

Observing the overall scores, we identified a common pattern. Among static visualizations, the ones from the news category, in general, scored higher than those from the government group. This could be attributed to the nature of news visualizations, which are often designed for specific audiences requiring explanatory elements. This finding provides insights into the nature of different visualization categories and also highlights the sensitivity of the model to these nuances.

On the other hand, interactive visualizations received the highest scores among all three evaluators. This could be explained by the vast array of technologies available, which enables designers to construct narratives in diverse and innovative ways. Interactive components facilitate exploration and increase user engagement. In addition, interactivity allows for the manipulation of multiple variables, leading to a more dynamic representation of information. Therefore, the combination of narrative elements with interactive features appears to increase the quality of visualizations.

Fig. 7 illustrates two examples of evaluated visualizations, depicting the varying average quality scores between (A) a static visualization and (B) an interactive visualization.

Fig. 7: Example of evaluated visualizations. A) a static visualization with an average quality score of 0.46. B) an interactive visualization with an average quality score of 0.99.

4.3.2. RQ2: What are the stre0ngths and limitations of the quality model in assessing the quality of narrative visualizations?

To identify the strengths and limitations of the quality model, we gathered feedback from the evaluators who participated in the assessment process. The survey provided insights into the perspective of evaluators regarding the performance of the model.

Among the primary strengths of the model, as reported in the survey responses, was its diagnostic capability. Evaluators found that the model served to identify areas in visualizations that could be improved, such as the clarity of the data, or visual clutter. This is useful to enhance the quality of existing visualizations and to guide the development of future designs. Moreover, the specific insights gained through this process allow for targeted improvements, resulting in visualizations being more aesthetically pleasing, informative and engaging for the viewer.

The survey also revealed that the model demonstrated sensitivity to subtle differences in the quality of visualizations. Evaluators appreciated the granularity of evaluation, which allowed them to assess design elements such as label placement, annotation orientation, and narrative structure. By providing topic-specific response options for each application, the model supported precise, detailed improvements. The coverage of the model was highlighted as another positive aspect, incorporating crucial aspects of narrative visualizations, from scale accuracy to narrative flow and event sequence.

Flexibility was identified as a noteworthy feature of the model. The survey responses indicated that the model could be applied across different domains where the narrative component plays a role in communicating complex information. This capability goes beyond specific subject areas, indicating the potential of the model for uses in real world settings.

However, the survey feedback also shed light on certain challenges. The wide coverage and sensitivity to nuances introduced a level of complexity that some users could find overwhelming, especially those without background knowledge. The evaluators noted that assessing a visualization in an accurate, thorough manner might require a certain level of expertise.

Additionally, the survey responses highlighted the potential time investment required for in-depth assessments. While the average completion time for an evaluation was 10 minutes, evaluators acknowledged that conducting thorough assessments could be time-consuming, particularly when evaluating a large volume of visualizations. This could limit the feasibility of the model for quick assessments.

The survey feedback also emphasized that the model was subject to some degree of subjectivity. Despite efforts to define clear and objective response options, depending on the evaluator's perspective, certain applications can be open to interpretation.

Overall, the survey results highlighted the need for continuous refinement and iterative improvements. By addressing these limitations, the model can better support the evaluation process for narrative visualizations.

4.3.3 RQ3: What is the reliability of the quality model in terms of producing consistent evaluations by different evaluators?

To assess inter-rater reliability in our quality model, we used Intraclass Correlation Coefficient (ICC). This measure is used to assess the reliability of ratings, particularly in studies where numerical or continuous measurements are made on the same subjects by different observers (Koo & Li, 2016). The ICC value is interpreted according to the following thresholds:

- Less than 0.5: poor reliability.
- Between 0.5 and 0.75: moderate reliability.
- Between 0.75 and 0.9: good reliability.
- Greater than 0.9: excellent reliability.

To calculate the ICC value, we utilized the results obtained from the evaluations conducted by the different evaluators. After setting up a summary spreadsheet with the evaluation data, we applied the ICC formula to assess the inter-rater reliability of the model.

The computed ICC value was 0.94, indicating excellent reliability. This suggests that the model is robust and capable of producing consistent results regardless of the background or expertise of evaluators.

This consistency can also be attributed to the precision of the evaluation criteria. We aimed to define each criterion as clear and unambiguous as possible, explaining evaluators what each of them entailed, to minimize discrepancies in interpretation.

It is worth mentioning that these findings consider the varying levels of experience and knowledge of evaluators, which we discuss further in the following RQ.

This highlights the potential use of the model in real-world settings, in which different evaluators might be involved.

4.3.4 RQ4: What is the perceived level of difficulty when applying the evaluation model?

To assess the perceived level of difficulty when applying the evaluation model, we included a specific question in the survey to gather feedback from the evaluators. The question asked them to rate their overall experience using the Narrative Visualization Quality Model on a scale of 1 to 5, with 1 indicating "Very difficult" and 5 indicating "Very easy." The survey also provided an optional section for additional comments.

Based on the survey responses from the three evaluators, we could observe that the perceived level of difficulty varied among them based on their background knowledge and experience.

Evaluator 1, with extensive background knowledge, found the model relatively easy to apply, completing each evaluation in an average of 10 minutes without requiring additional training. This evaluator stated that "*the model was straightforward to use*", which suggests it is accessible for users with a higher understanding of the topic.

Evaluator 2, with intermediate knowledge, also required an average of 10 minutes for each evaluation. While she found the model intuitive overall, she noted that "*some of applications were difficult to understand*". This feedback highlights the importance of providing appropriate training and support to users with varying levels of experience.

On the other hand, Evaluator 3 had little familiarity with the subject and required the most time to complete an evaluation, with an average of 15 minutes. This evaluator expressed that "*the model was challenging to use without prior knowledge. An* *'unclear' option would be helpful for the more puzzling applications.*" This suggests a possible refinement for future iterations of the model.

These finding indicate that the model can be used across varying degrees of knowledge and experience, with a reasonable time commitment. However, it also highlights the importance of an initial training, particularly for less experienced users, to make the model more intuitive and easier to use.

5 Discussion

This section discusses the implications derived from our work. We reflect on the main findings of the research questions and address the threats to the validity of the results.

5.1 Lessons learned

After analyzing the results of the evaluations presented in Section 4.3.1 and the evaluators' feedback gathered from the surveys, we structured the lessons learned along three dimensions, described below.

Enhanced clarity of applications and response options: The evaluation revealed that certain applications, such as AP03, AP08, AP22, AP28, and AP43, could benefit from a revised framing and clearer language. Simplifying the wording and introducing an "unclear" option, as suggested by Evaluator 2, would improve interpretation, and reduce confusion, particularly for non-expert users.

Comprehensive training materials: The evaluation highlighted the importance of providing comprehensive training for users of the quality model. Detailed explanations and illustrative examples for each application should be incorporated to guide evaluators in their assessments. While integrating such training mechanisms into a software tool may be more practical than a spreadsheet format, it would significantly enhance the usability and effectiveness of the model.

Annotation or comment section: Based on feedback from evaluators, the inclusion of an annotation or comment section for each application is recommended. This feature would allow evaluators to provide context-specific notes or explanations during their assessments. It would facilitate the generation of comprehensive and explanatory reports, capturing valuable insights about the evaluation process and unique aspects encountered in specific visualizations.

By incorporating these improvements in future works, we expect the quality model to be accessible to a wider range of users, capable of generating comprehensive and informative evaluations, and enabling discussions that help build consensus among evaluators. These changes aim to address the challenges identified during the small-scale evaluation, fostering a more effective and reliable assessment process for narrative visualizations.

5.2 Practical implications

The following are practical implications that using the model could have for researchers and practitioners. These implications are derived from the findings and lessons learned during the small-scale evaluation.

Improved decision making: Higher quality visualizations tend to be more memorable and comprehensible. The user-friendly nature of such visualizations allows users to interact with them in an intuitive manner, reducing cognitive load and freeing up mental resources to focus on the task (Dimara & Stasko, 2022). Therefore, they could lead to a more effective decision-making process by end users. The model can also serve as a foundation for exploratory studies that delve deeper into the connection between the quality of a visualization and its decision-making support.

Consistent design practices: By defining clear quality criteria and providing a systematic evaluation approach, the model can assist authors with a set of best practices.

Within organizations, this could generate consistent results that adhere to a certain quality standard. The model can also function as a training tool to understand the key components of narrative visualizations, thus bridging the gap between technical and non-technical teams. For researchers, it provides a framework for studying and advancing these design practices across different fields and applications.

Benchmarking: The model can serve as a reference point against which future visualizations can be measured. In an industry setting, this involves setting a standard to compare results over time, allowing to monitor progress and identify areas of improvement. In academia, this can be valuable in experimental contexts. For example, researchers could use the model to compare different approaches to narrative design and assess new techniques or methods.

Time and resource management: A deeper understanding of narrative visualization best practices can lead to a more efficient management of time and resources. By identifying issues early in the design process, both practitioners and researchers can prioritize their resources to address them, optimizing the overall workflow.

User engagement: A final implication of this model is its capacity to increase user engagement. By educating about the best practices in narrative visualization design, it enables the creation of comprehensible visualizations. In an academic context, this could lead to impactful research findings, as engaged users have a greater chance to retain and recall the information.

It is important to note that the quality score generated by the model is not intended to measure the aesthetic of a visualization, but rather the extent to which a narrative data visualization adheres to best practices for effective communication. For example, if a user selects a chart type that is not optimal for the data but still satisfies all the quality criteria and properties, the resulting score will reflect this situation.

5.3 Threats to validity

Validity refers to the reliability of the findings – the extent to which the results are accurate and not influenced by the perspective of the researchers. We addressed four aspects of validity, as proposed by (Runeson & Höst, 2009).

Construct validity: It reflects the extent to which the research methodology represents what the researchers have in mind and what is investigated in relation to the research questions. In this case, evaluators might bring their own biases into the assessment process, which could have affected how well the evaluations reflected the intention of the study. To minimize this threat, we provided detailed explanations before the process to assure the evaluators understood each element of the model. In addition, each element was defined as clear and unambiguous as possible, following well-established guidelines from the literature.

Internal validity: It assesses the risks associated with studying cause-and-effect relationships. When we try to determine if one factor influences another, we must also consider that there may be an unseen third factor influencing the situation. If we are not aware of this third factor or do not understand its impact, it can pose a threat to our understanding of the relationship between the initial two factors. The differing levels of expertise and knowledge among the evaluators might have influenced the evaluation results. To address this, we incorporated evaluators with different levels of expertise and knowledge. By doing so, we aimed to emulate a realistic scenario in which a variety of individuals would be using the model.

External validity: It refers to what extent it is possible to generalize our findings and how relevant they are to the public. One potential concern is our limited sample size in terms of visualization techniques to assess. To mitigate this threat and ensure a diverse and representative sample, we used a Python script for random selection. This allowed us to select a subset of visualizations from a dataset of over 5000 files, thereby improving the range of visualization types included.

Reliability: It pertains to the degree to which the research findings and data are independent from the researchers conducting the study. If a different researcher were to replicate the study, the results should align closely or be comparable to the original findings. Although our study did not involve external evaluators, the high inter-rater agreement achieved suggests that the model can provide consistent results across different evaluators. Future iterations of this work could involve external evaluators to further establish the reliability of the model.

6 Conclusions and future work

This paper presented a quality model for evaluating narrative visualizations. We constructed the model following the systematic approach proposed in (Siebert et al., 2022), drawing upon existing quality models in SE and Information Visualization. We aimed to capture the fundamental aspects that characterize this type of visualizations. To validate the model, we conducted a small-scale evaluation with three evaluators using spreadsheets and a set of static and interactive visualizations.

Our findings indicate that the model is reliable, robust and capable of producing consistent results regardless of the background or level of expertise of evaluators, as determined by the ICC coefficient. The granularity embedded in the model allows to identify precise and detailed improvements in a visualization.

Perceived limitations of the model include an inherent complexity due to the number of applications, as well as a degree of subjectivity. Additionally, the small-scale evaluation used to assess the feasibility of the model was conducted on a limited sample size, which leaves room for potential discrepancies in larger and more diverse evaluations.

In terms of practical implications, the model contributes to an improved decision-making, encourages consistent design practices and serves as a benchmarking tool, among others. This highlights the value of our contribution to both academic researchers and industry practitioners.

Future research will be directed towards refining the model and exploring its application in a more diverse range of domains and contexts. For this purpose, we plan on developing a software application to operationalize the model, implementing the improvements we discussed and validating its use in real-world settings.

References

Aliyev, R., Temizkan, H., & Aliyev, R. (2020). Fuzzy Analytic Hierarchy Process-Based Multi-Criteria Decision Making for Universities Ranking. *Symmetry 2020, Vol. 12, Page 1351*, *12*(8), 1351. https://doi.org/10.3390/SYM12081351

Amini, F., Brehmer, M., Bolduan, G., Elmer, C., & Wiederkehr, B. (2018). Evaluating Data-Driven Stories and Storytelling Tools *. In *Data-Driven Storytelling* (1st ed., pp. 249–286). A K Peters/CRC Press. https://doi.org/10.1201/9781315281575-11

Bach, B., Stefaner, M., Boy, J., Drucker, S., Bartram, L., Wood, J., Ciuccarelli, P., Engelhardt, Y., Köppen, U., & Tversky, B. (2018). Narrative Design Patterns for Data-Driven Storytelling. In *Data-Driven Storytelling* (pp. 107–133). CRC Press (Taylor & Francis). https://doi.org/10.1201/9781315281575-5

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: Methodological and philosophical foundations. *MIS Quarterly: Management Information Systems*, *35*(2), 261–292. https://doi.org/10.2307/23044044

Bai, X., White, D., & Sundaram, D. (2009). Visual intelligence density. *Proceedings of the 10th International Conference NZ Chapter of the ACM's Special Interest*

Group on Human-Computer Interaction - CHINZ '09, 93–100. https://doi.org/10.1145/1577782.1577799

Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., & Brooks, C. (2010). Useful junk? The effects of visual embellishment on comprehension and memorability of charts. *Conference on Human Factors in Computing Systems - Proceedings*, *4*, 2573–2582. https://doi.org/10.1145/1753326.1753716

Battle, L., Duan, P., Miranda, Z., Mukusheva, D., Chang, R., & Stonebraker, M. (2017). Beagle: Automated Extraction and Interpretation of Visualizations from the Web. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, *2018-April*, 1–8. https://doi.org/10.1145/3173574.3174168

Bertin, J. (2010). *Semiology of Graphics: Diagrams, Networks, Maps* (1st ed.). Esri Press.

Bollen, K. A., & Ting, K. F. (2000). A tetrad test for causal indicators. *Psychological Methods*, *5*(1), 3–22. https://doi.org/10.1037/1082-989X.5.1.3

Borkin, M. A., Bylinskii, Z., Kim, N. W., Bainbridge, C. M., Yeh, C. S., Borkin, D., Pfister, H., & Oliva, A. (2016). Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 519–528. https://doi.org/10.1109/TVCG.2015.2467732

Börner, K., Maltese, A., Balliet, R. N., & Heimlich, J. (2016). Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, *15*(3), 198–213. https://doi.org/10.1177/1473871615594652

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology*, *29*(3), 461–473. https://doi.org/10.1080/14640747708400622

Cairo, A. (n.d.). *The New Normal*. Retrieved February 21, 2023, from https://thenewnormal.is/

Card, S., Mackinlay, J., & Shneiderman, B. (1999). Readings in Information Visualisation. Using Vision to Think.: Using Vision to Think. In *Morgan Kaufmann* (Issue January). Morgan Kaufmann Publishers.

Carpendale, S. (2008). Evaluating Information Visualizations. In *Information Visualization: Vol. 4950 LNCS* (Issue January 1970, pp. 19–45). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_2

Carswell, C. M. (1992). Choosing Specifiers: An Evaluation of the Basic Tasks Model of Graphical Perception. *Https://Doi.Org/10.1177/001872089203400503*, *34*(5), 535–554. https://doi.org/10.1177/001872089203400503

Cawthon, N., & Moere, A. Vande. (2007). The effect of aesthetic on the usability of data visualization. *Proceedings of the International Conference on Information Visualisation*, 637–645. https://doi.org/10.1109/IV.2007.147

Chen, Q., Cao, S., Wang, J., & Cao, N. (2022). *How Does Automation Shape the Process of Narrative Visualization: A Survey on Tools*. 1–20. http://arxiv.org/abs/2206.12118

Cherubini, F., & Nielsen, R. K. (2016). Editorial Analytics: How News Media are Developing and Using Audience Data and Metrics. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2739328

Diakopoulos, N. (2018). Ethics in Data-Driven Visual Storytelling. In *Data-Driven Storytelling* (pp. 233–248). A K Peters/CRC Press. https://doi.org/10.1201/9781315281575-10

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*(12), 1203–1218. https://doi.org/10.1016/J.JBUSRES.2008.01.009

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index Construction with Formative Indicators: An Alternative to Scale Development. *Https://Doi.Org/10.1509/Jmkr.38.2.269.18845*, *38*(2), 269–277. https://doi.org/10.1509/JMKR.38.2.269.18845

Dimara, E., & Stasko, J. (2022). A Critical Reflection on Visualization Research: Where Do Decision Making Tasks Hide? *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 1128–1138. https://doi.org/10.1109/TVCG.2021.3114813

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Elmqvist, N., & Yi, J. S. (2015). Patterns for visualization evaluation. *Information Visualization*, *14*(3), 250–269. https://doi.org/10.1177/1473871613513228

Evergreen, S. D. H. (2012). Death by boredom: The role of visual processing theory in written evaluation communication. [Western Michigan University]. In

Dissertation Abstracts International Section A: Humanities and Social Sciences. https://scholarworks.wmich.edu/dissertations/403

Fayers, P. M., & Hand, D. J. (2002). Causal variables, indicator variables and measurement scales: an example from quality of life. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *165*(2), 233–253. https://doi.org/10.1111/1467-985X.02020

Feng, M., Peck, E., & Harrison, L. (2019). Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 501–511. https://doi.org/10.1109/TVCG.2018.2865117

Food prices are outpacing wider inflation across most of the world | The Economist. (n.d.). Retrieved February 21, 2023, from https://www.economist.com/graphic-detail/2022/10/07/food-prices-are-outpacing-wider-inflation-across-most-of-the-world

Forsell, C., & Johansson, J. (2010). An heuristic set for evaluation in information visualization. *Proceedings of the International Conference on Advanced Visual Interfaces - AVI '10*, 199. https://doi.org/10.1145/1842993.1843029

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. *Journal for Research in Mathematics Education*, *32*(2), 124. https://doi.org/10.2307/749671

Haase, H. (1998). Evaluating the quality of scientific visualizations: the Q-VIS reference model. In R. F. Erbacher & A. Pang (Eds.), *Visual Data Exploration and Analysis V* (Vol. 3298, Issue January, pp. 123–131). https://doi.org/10.1117/12.309534

Healey, C. G., Booth, K. S., & Enns, J. T. (1993). Harnessing preattentive processes for multivariate data visualization. *Proceedings - Graphics Interface*, 107–117.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using mechanical Turk to assess visualization design. *Conference on Human Factors in Computing Systems - Proceedings*, *1*, 203–212. https://doi.org/10.1145/1753326.1753357

Helmy, S. E., Eladl, G. H., & Eisa, M. (2021). Fuzzy Analytical Hierarchy Process (FAHP) Using Geometric Mean Method to Select Best Processing Framework

Adequate to Big Data. *Journal of Theoretical and Applied Information Technology*, *15*(1). www.jatit.org

Hullman, J., & Diakopoulos, N. (2011). Visualization rhetoric: framing effects in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2231–2240. https://doi.org/10.1109/TVCG.2011.255

Hung, Y. H., & Parsons, P. (2017). Assessing user engagement in information visualization. *Conference on Human Factors in Computing Systems - Proceedings*, *Part F1276*, 1708–1717. https://doi.org/10.1145/3027063.3053113

Hwang, C.-L., & Yoon, K. (1981). *Multiple Attribute Decision Making Methods and Applications A State-of-the-Art Survey* (1st ed., Vol. 186). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-48318-9

ISO 9241-11. (2018). *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts*. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

ISO/IEC 25010. (2011). *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*. https://www.iso.org/standard/35733.html

ISO/IEC 33003. (2015). *Information technology — Process assessment — Requirements for process measurement frameworks*. https://www.iso.org/standard/54177.html

Jacobson, A. (2019). *The Value Proposition Of Good Government — Information is Beautiful Awards*. https://www.informationisbeautifulawards.com/showcase/4010-the-value-proposition-of-good-government

Johansson, J., & Forsell, C. (2016). Evaluation of Parallel Coordinates: Overview, Categorization and Guidelines for Future Research. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 579–588. https://doi.org/10.1109/TVCG.2015.2466992

Jung, H. W. (2013). Investigating measurement scales and aggregation methods in SPICE assessment method. *Information and Software Technology*, *55*(8), 1450–1461. https://doi.org/10.1016/j.infsof.2013.02.004

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/J.JCM.2016.02.012

Kurosu, M., & Kashimura, K. (1995). Apparent usability vs. inherent usability. *Conference Companion on Human Factors in Computing Systems - CHI '95*, 292–293. https://doi.org/10.1145/223355.223680

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., & Carpendale, S. (2012). Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics*, *18*(9), 1520–1536. https://doi.org/10.1109/TVCG.2011.279

Lan, X., Shi, Y., Zhang, Y., & Cao, N. (2021). Smile or Scowl? Looking at Infographic Design Through the Affective Lens. *IEEE Transactions on Visualization and Computer Graphics*, *27*(6), 2796–2807. https://doi.org/10.1109/TVCG.2021.3074582

Lan, X., Wu, Y., Chen, Q., & Cao, N. (2022). *The Chart Excites Me! Exploring How Data Visualization Design Influences Affective Arousal*. 1–12. http://arxiv.org/abs/2211.03296

Leal, J. E. (2020). AHP-express: A simplified version of the analytical hierarchy process method. *MethodsX*, *7*, 100748. https://doi.org/10.1016/j.mex.2019.11.021

Lee, S., Kim, S.-H., & Kwon, B. C. (2017). VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 551–560. https://doi.org/10.1109/TVCG.2016.2598920

Lei, T., Ni, N., Zhu, Q., & Zhang, S. (2018). Aesthetic experimental study on information visualization design under the background of big data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10919 LNCS*, 218–226. https://doi.org/10.1007/978-3-319-91803-7_16

Lezcano Airaldi, A., Irrazábal, E., & Diaz-Pace, J. A. (2022). *Narrative Visualizations Best Practices and Evaluation: A Systematic Mapping Study*. https://doi.org/10.21203/RS.3.RS-1735564/V1

Liu, F. H. F., & Hai, H. L. (2005). The voting analytic hierarchy process method for selecting supplier. *International Journal of Production Economics*, *97*(3), 308–317. https://doi.org/10.1016/J.IJPE.2004.09.005

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics (TOG)*, *5*(2), 110–141. https://doi.org/10.1145/22949.22950

MASSVIS - Massachusetts (Massive) Visualization Dataset. (n.d.). Retrieved March 5, 2023, from http://massvis.mit.edu/

Morisio, M., Stamelos, I., & Tsoukias, A. (2002). A new method to evaluate software artifacts against predefined profiles. *Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering*, *27*, 811–818. https://doi.org/10.1145/568760.568899

Munzner, T. (2014). *Visualization Analysis and Design* (1st ed.). A K Peters/CRC Press.

Nasution, S. M., Husni, E., Kuspriyanto, K., & Yusuf, R. (2022). Personalized Route Recommendation Using F-AHP-Express. *Sustainability*, *14*(17), 10831. https://doi.org/10.3390/su141710831

Norman, D. A. (2005). *Emotional Design: Why We Love (or Hate) Everyday Things* (1st ed.).

Nowell, L., Schulman, R., & Hix, D. (2002). Graphical encoding for information visualization: an empirical study. *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, *2002-Janua*, 43–50. https://doi.org/10.1109/INFVIS.2002.1173146

Nussbaumer Knaflic, C. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley.

OECD. (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide. In *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD. https://doi.org/10.1787/9789264043466-en

Padda, H., Mudur, S., Seffah, A., & Joshi, Y. (2008). Comprehension of Visualization Systems - Towards Quantitative Assessment. *First International Conference on Advances in Computer-Human Interaction*, 83–88. https://doi.org/10.1109/ACHI.2008.19

Padda, H., Seffah, A., & Mudur, S. (2007). Visualization Patterns: A Context-Sensitive Tool to Evaluate Visualization Techniques. *2007 4th IEEE International Workshop on Visualizing Software for Understanding and Analysis*, 88–91. https://doi.org/10.1109/VISSOF.2007.4290705

Paramount pushes in: New streamers are still finding ways to grow. (n.d.). Retrieved February 21, 2023, from https://www.chartr.co/stories/2023-02-17-1-paramount-streaming-platform-is-growing

Patton, M. Q. (2014). Qualitative research and evaluation methods: Integrating Theory and Practice. In *SAGE Publications, Inc.* (4th ed.). Sage publications.

Perin, C., Wun, T., Pusch, R., & Carpendale, S. (2018). Assessing the Graphical Perception of Time and Speed on 2D+Time Trajectories. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 698–708. https://doi.org/10.1109/TVCG.2017.2743918

Plaisant, C. (2004). The challenge of information visualization evaluation. *Proceedings of the Workshop on Advanced Visual Interfaces AVI*, 109–116. https://doi.org/10.1145/989863.989880

Plaisant, C., Fekete, J.-D., & Grinstein, G. (2008). Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository. *IEEE Transactions on Visualization and Computer Graphics*, *14*(1), 120–134. https://doi.org/10.1109/TVCG.2007.70412

Quispel, A., Maes, A., & Schilperoord, J. (2016). Graph and chart aesthetics for experts and laymen in design: The role of familiarity and perceived ease of use. *Information Visualization*, *15*(3), 238–252. https://doi.org/10.1177/1473871615606478

Robson, C. (2017). *Small-Scale Evaluation: Principles and Practice*. Sage Publications.

Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, *14*(2), 131–164. https://doi.org/10.1007/s10664-008-9102-8

Saaty, T. L. (1980). *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. McGraw-Hill.

Saket, B., Srinivasan, A., Ragan, E. D., & Endert, A. (2018). Evaluating Interactive Graphical Encodings for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *24*(3), 1316–1330. https://doi.org/10.1109/TVCG.2017.2680452

Santos, B. S., Ferreira, B. Q., & Dias, P. (2015). Heuristic Evaluation in Information Visualization Using Three Sets of Heuristics: An Exploratory Study. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial*

Intelligence and Lecture Notes in Bioinformatics) (Vol. 9169, pp. 259–270). https://doi.org/10.1007/978-3-319-20901-2_24

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1139–1148. https://doi.org/10.1109/TVCG.2010.179

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, *14*(1), 47–69. https://doi.org/10.1023/A:1013180410169

Siebert, J., Joeckel, L., Heidrich, J., Trendowicz, A., Nakamichi, K., Ohashi, K., Namba, I., Yamamoto, R., & Aoyama, M. (2022). Construction of a quality model for machine learning systems. *Software Quality Journal*, *30*(2), 307–335. https://doi.org/10.1007/s11219-021-09557-y

Stasko, J. (2014). Value-driven evaluation of visualizations. *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, *10-Novembe*, 46–53. https://doi.org/10.1145/2669557.2669579

Stevens, S. S. (1951). *Mathematics, Measurement, and Psychophysics*. John Wiley & Sons, Inc.

Tavana, M., Soltanifar, M., & Santos-Arteaga, F. J. (2021). Analytical hierarchy process: revolution and evolution. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04432-2

Tory, M., & Moller, T. (2005). Evaluating visualizations: do expert reviews work? *IEEE Computer Graphics and Applications*, *25*(5), 8–11. https://doi.org/10.1109/MCG.2005.102

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Graphics Press. https://www.jstor.org/stable/530384?origin=crossref

Wagner, S., Goeb, A., Heinemann, L., Kläs, M., Lampasona, C., Lochmann, K., Mayr, A., Plösch, R., Seidl, A., Streit, J., & Trendowicz, A. (2015). Operationalised product quality models and assessment: The Quamoco approach. *Information and Software Technology*, *62*(1), 101–123. https://doi.org/10.1016/j.infsof.2015.02.009

Wainer, H. (1992). Understanding Graphs and Tables. *Http://Dx.Doi.Org/10.3102/0013189X021001014*, *21*(1), 14–23. https://doi.org/10.3102/0013189X021001014

Waldner, M., Diehl, A., Gracanin, D., Splechtna, R., Delrieux, C., & Matkovic, K. (2019). A Comparison of Radial and Linear Charts for Visualizing Daily Patterns. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 1–1. https://doi.org/10.1109/TVCG.2019.2934784

Wall, E., Agnihotri, M., Matzen, L., Divis, K., Haass, M., Endert, A., & Stasko, J. (2019). A Heuristic Approach to Value-Driven Evaluation of Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 491–500. https://doi.org/10.1109/TVCG.2018.2865146

Ware, C. (2020). *Information Visualization: Perception for Design (Interactive Technologies)* (4th ed.). Elsevier. https://doi.org/10.1016/B978-0-12-812875-6.01001-X

Williams, R., Scholtz, J., Blaha, L. M., Franklin, L., & Huang, Z. (2018). Evaluation of Visualization Heuristics. In M. Kurosu (Ed.), *Human-Computer Interaction. Theories, Methods, and Human Issues* (Vol. 10901, Issue January, pp. 208–224). Springer International Publishing. https://doi.org/10.1007/978-3-319-91238-7_18

Wohlin, C., & Rainer, A. (2022). Is it a case study?—A critical analysis and guidance. *Journal of Systems and Software*, *192*, 111395. https://doi.org/10.1016/j.jss.2022.111395

Xiaoyan Bai, White, D., & Sundaram, D. (2010). Purposeful Visualization System. *2010 Second World Congress on Software Engineering*, *1*, 241–244. https://doi.org/10.1109/WCSE.2010.100

Yi, J. S. (2010, April). Implications of Individual Differences on Evaluating Information Visualization Techniques. *Proceedings of beyond Time and Errors: Novel Evaluation Methods for Visualization*.

Yi, J. S., Kang, Y., Stasko, J. T., & Jacko, J. A. (2007). *Toward a Deeper Understanding of the Role of Interaction in Information Visualization*. *October*.

Yoon, K., & Hwang, C.-L. (2011). Multiple Attribute Decision Making: An Introduction. In *Multiple Attribute Decision Making* (Vol. 10). SAGE Publications, Inc. https://doi.org/10.4135/9781412985161

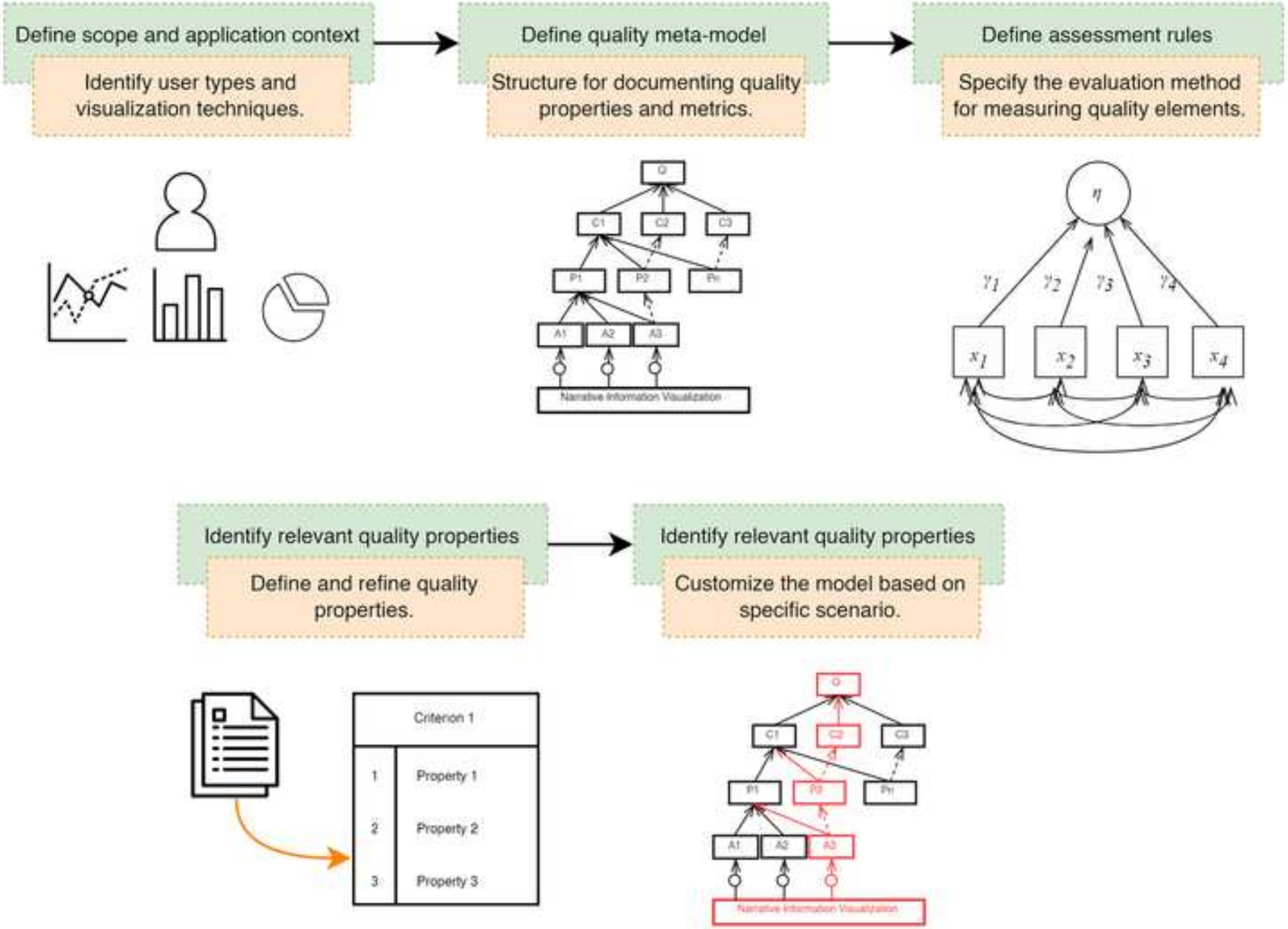
Zeleny, M. (1982). *Multiple Criteria Decision Making*. McGraw-Hill.

Zhang, Y., Bellamy, R. K. E., & Kellogg, W. A. (2015). Designing information for remediating cognitive biases in decision-making. *CHI '15: Proceedings of the*

33rd Annual ACM Conference on Human Factors in Computing Systems, *2015-April*, 2211–2220. https://doi.org/10.1145/2702123.2702239

Zhu, Y. (2007). Measuring Effective Data Visualization. In *Advances in Visual Computing* (pp. 652–661). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-76856-2_64

Zuk, T., Schlesier, L., Neumann, P., Hancock, M. S., & Carpendale, S. (2006). Heuristics for information visualization evaluation. *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors Novel Evaluation Methods for Information Visualization - BELIV '06*, 1. https://doi.org/10.1145/1168149.1168162



Define scope and application context

Identify user types and visualization techniques.

Define quality meta-model

Structure for documenting quality properties and metrics.

Define assessment rules

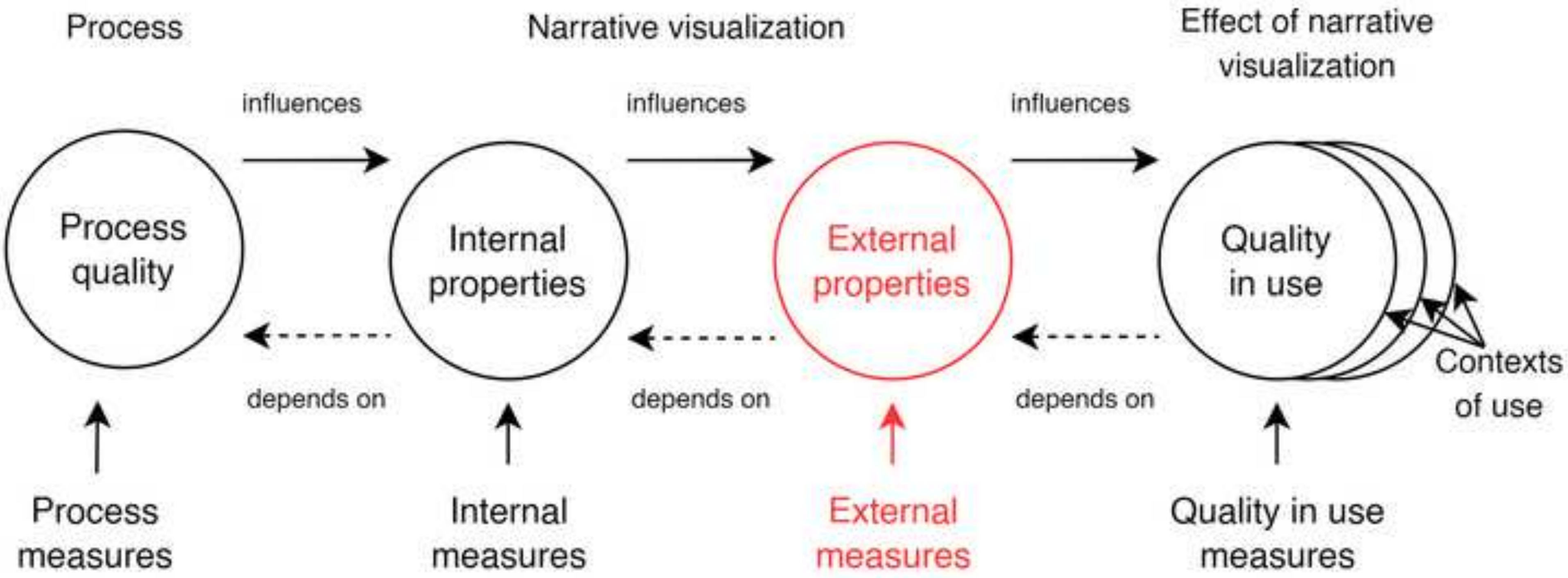
Specify the evaluation method for measuring quality elements.

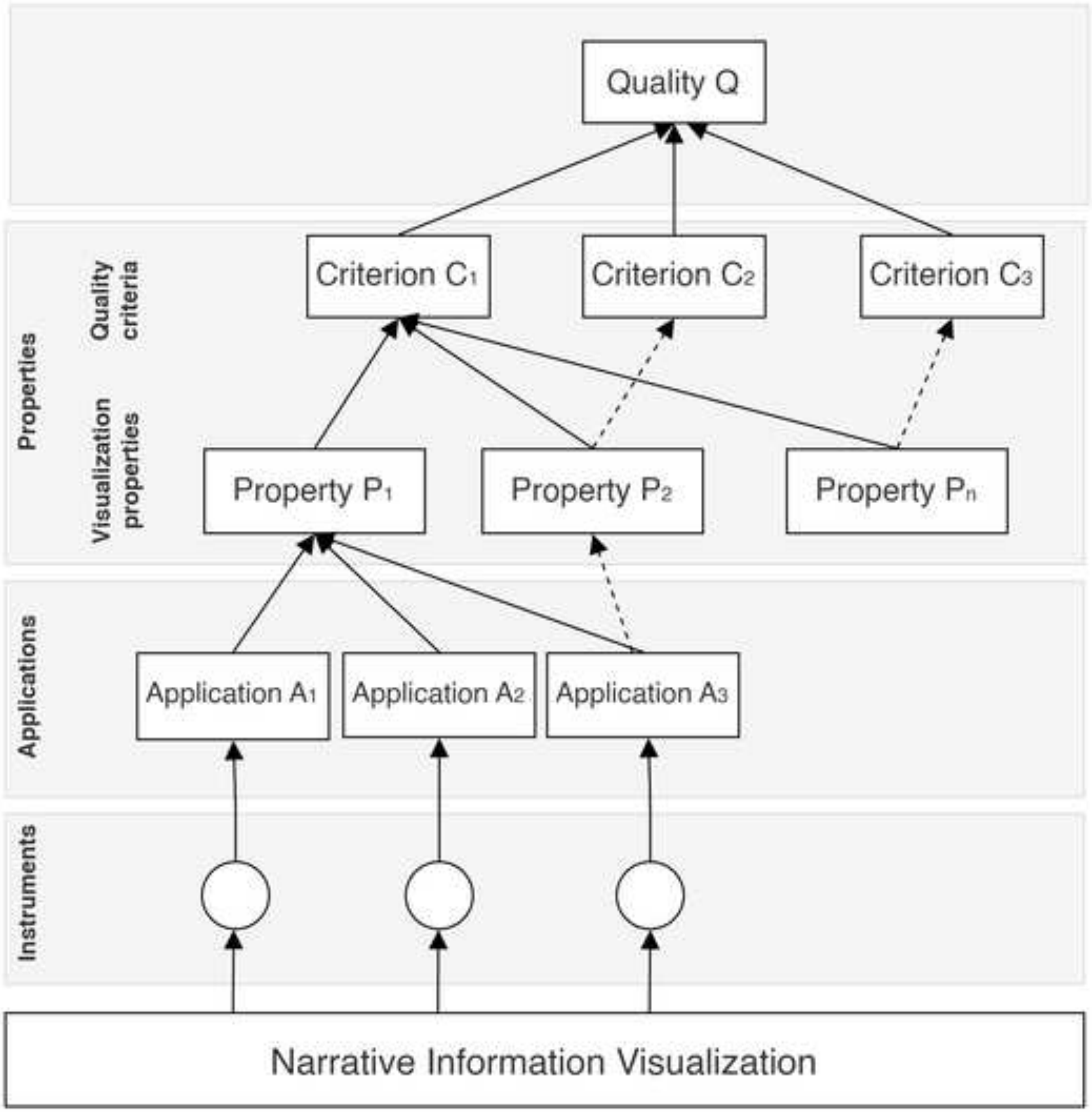
Identify relevant quality properties

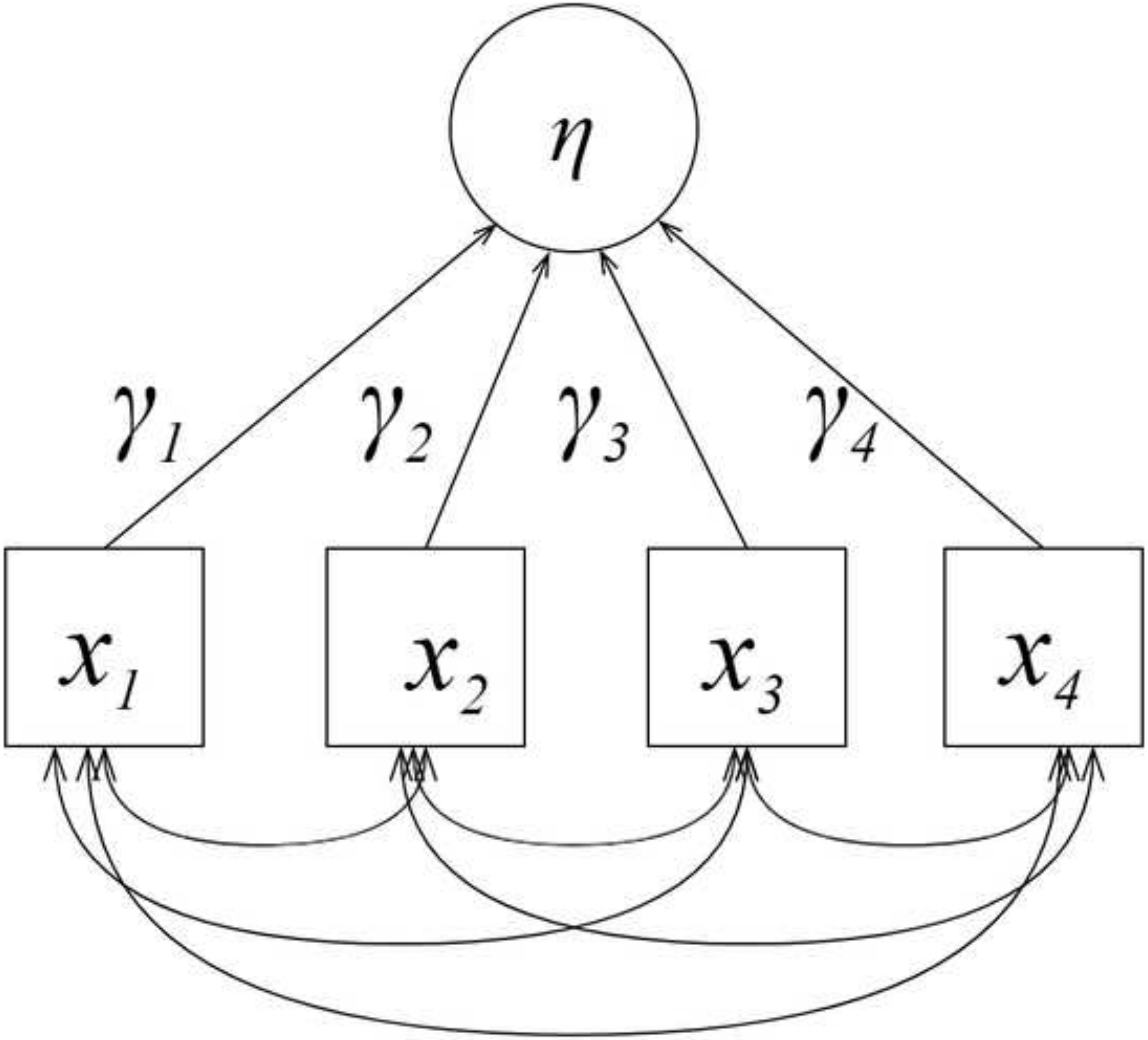
Define and refine quality properties.

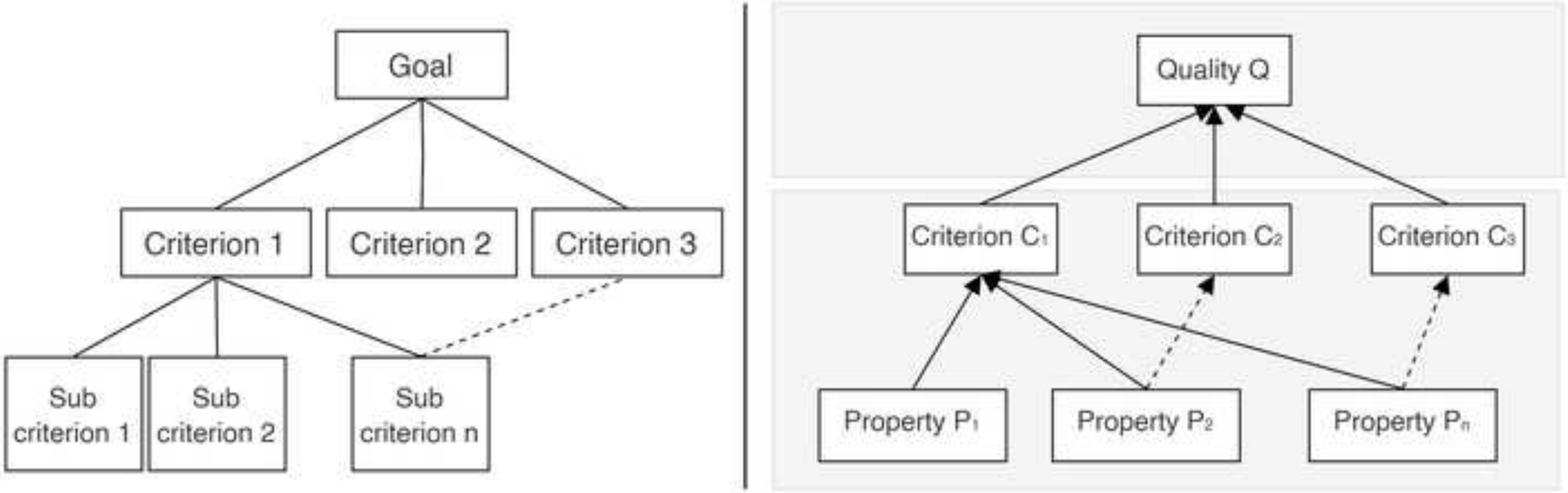
Identify relevant quality properties

Customize the model based on specific scenario.

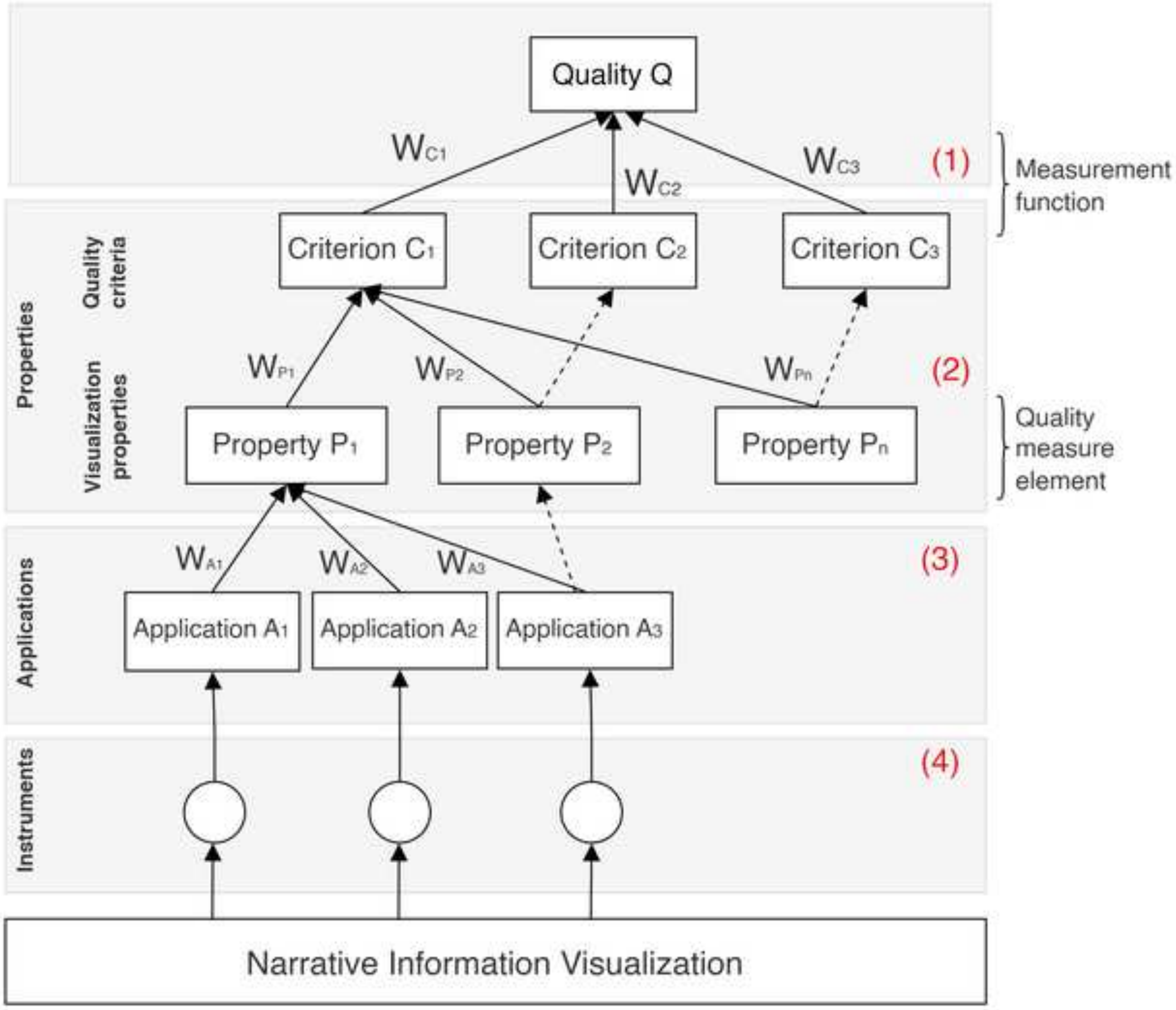

Process
Narrative visualization
Effect of narrative visualization
influences
influences
influences
Process quality
Internal properties
External properties
Quality in use
depends on
depends on
depends on
Contexts of use
Process measures
Internal measures
External measures
Quality in use measures


Quality Q
Criterion C1
Criterion C2
Criterion C3
Property P1
Property P2
Property Pn
Application A1
Application A2
Application A3
Narrative Information Visualization
Quality criteria
Visualization properties
Properties
Applications
Instruments


η
γ_1
γ_2
γ_3
γ_4
x_1
x_2
x_3
x_4

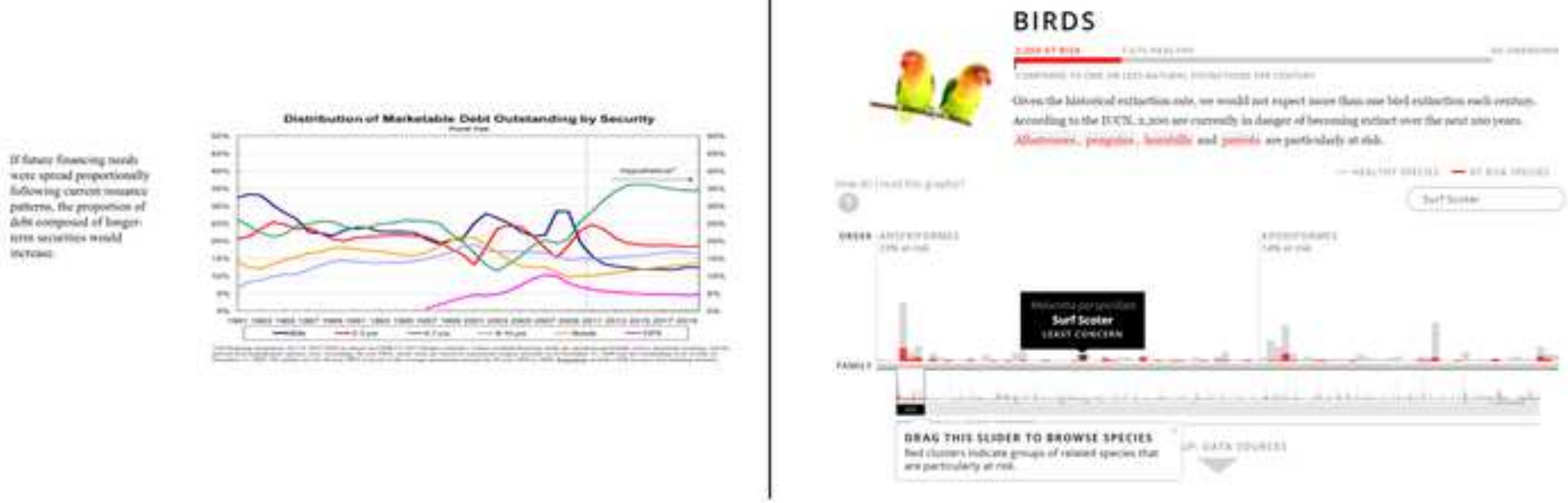

Goal
Criterion 1
Criterion 2
Criterion 3
Sub criterion 1
Sub criterion 2
Sub criterion n
Quality Q
Criterion C_1
Criterion C_2
Criterion C_3
Property P_1
Property P_2
Property P_n

Click here to access/download;Figure;Figure 6.jpeg



(A) ID: 05_gov

(B) ID: 02_int



If future financing needs were spread proportionally following current issuance patterns, the proportion of debt composed of longer-term securities would increase.

BIRDS

Given the historical extinction rate, we would not expect more than one bird extinction each century. According to the IUCN, 2,200 are currently in danger of becoming extinct over the next 100 years. Albatrosses, penguins, hornbills and parrots are particularly at risk.

DRAG THIS SLIDER TO BROWSE SPECIES

Red clusters indicate groups of related species that are particularly at risk.