applied sciences

MDPI

Article

DRIVE-T: A Methodology for Discriminative and Representative Items Selection for Design Quality Constructs and Assessments

Angela Locoro [1,*], Silvia Golia [1] and Davide Falessi [2]

[1] Department of Economics and Management, Universitá degli Studi di Brescia, Contrada Santa Chiara, 50, 25122 Brescia, Italy
[2] Department of Informatics Engineering, Universitá di Roma Tor Vergata, Viale del Politecnico 1, 00133 Roma, Italy
* Correspondence: angela.locoro@unibs.it

Abstract

The lack of measurement constructs for both users' literacy and artifacts difficulty in HCI hinders the quality of assessment tests. This paper proposes DRIVE-T (Discriminating and Representative Items for Validating Expressive Tests), a methodology designed to construct and evaluate items for measuring progressive levels of users'skills and artifacts usability. Given an artifact (a text, a data visualization, an interface prototype), DRIVE-T supports the identification of items discriminability and representativeness for measuring distinct traits of a quality property to be measured and its levels of progression in users and artifacts. DRIVE-T consists of three steps: (1) tagging task-based items associated with an artifact; (2) rating them by independent raters for their difficulty; (3) analyzing raters' raw scores through a Many-Facet Rasch Measurement model. The emergence of difficulty levels of the measurement construct can be derived from the discriminability and representativeness of items for each artifact, ordered into Many-Facets construct levels. The DRIVE-T methodology operationalizes an inductive, practice-based measurement construct design. Based on the expertise of the authors in the data visualization domain, visualization literacy is the quality property of users that is exploited as a case scenario for applying DRIVE-T. Results from a pilot test show the validity of the approach.

Keywords: digital literacies; visualization literacy; construct modelling and assessment; task-based item design and evaluation; HCI quality assessment and design; many-facet rasch measurement model

1. Introduction and Motivation

Design quality is a measurable property related to human's skills and digital artifacts affordances. Open issues largely debated in the HCI research community are related to how to go beyond usability [1,2]. How to accurately capture all of the aspects of human's skills, including whether they change with time and whether invariances exist remains a challenge [3].

An example of human skill is data visualization literacy (data viz literacy from now on), the capability to read and write data visualizations (data viz from now on) with varying levels of difficulty. Earlier studies proposing data viz literacy assessment items explicitly relied on existing cognitive models of learning evolution [4,5]. Those models failed to articulate a clear construct mapping between them and item characteristics, such as discriminability or representativeness. Item discriminability is the ability of items to

Check for updates

Academic Editor: Zhonghua Sun

Received: 9 January 2026
Revised: 31 January 2026
Accepted: 2 February 2026
Published: 4 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

express progressive levels of literacy for users. Item representativeness is the ability of items to represent all traits or levels of artifacts difficulty.

In more recent data viz literacy studies and interface design, the need for a methodology for item preparation has become a compelling requirement [6,7]. So far, the mere definition of data viz literacy was used in place of a latent construct, and no further details into a more precise model [6] were given. Literacy should be discussed in terms of: levels of it, elements composing it, what should be measured, what underlying assumptions should be made explicit about the target population, the measuring intent, and the assessment tools. It is crucial that cautious stances [3] were adopted to avoid common biases when dealing with the problem of measuring a human's skill [8]. On the other hand, a construct based justification of items is necessary. A good tradeoff would yield literacy latent constructs that should remain underspecified, empirically grounded, and flexible enough to adapt to the complexity of evolving interfaces and real world applications [9].

In this paper, we propose a methodology, DRIVE-T (Discriminating and Representative Items for Validating Expressive Tests), whose steps fall within item modelling design [9]. Our proposal could alleviate the limitations of: (i) not making explicit the underlying assumptions of technology design [10]; (ii) nor making them too rigid and narrow. During design and prototyping, it is necessary to investigate both human skills and the affordances of digital artifacts. Methodological support is required to enhance the quality of assessment items and to make the overall process more transparent. Furthermore, post-design analysis is crucial to validate and refine the resulting instruments. Together, these actions can strengthen literacy measurement efforts and help them better meet the expectations of the HCI community.

Our research questions aimed at investigating:

1. How to proceed inductively from a set of users' tasks up to levels of literacy?
2. Is it possible to design and validate measurement constructs without involving respondents?
3. Can we determine, at design time, effectively discriminating and representative items for measuring progressive levels of users' skills and artifacts affordances?

In this study, we show and apply each step of the methodology to model the difficulty levels of a measurement construct approximating a latent construct for data viz literacy. This measurement construct is drawn from semiotics, i.e., based on syntax, semantics and pragmatics tasks that each data visualization may require to be mastered as object of literacy. These objects of literacy are generalizable to any digital artifact. Semiotics provides an operational framework for relating digital artifact properties to observable task performance. Artifacts interpretation relies on task-based shared conventions [11]. Level of difficulties emerge while interacting and managing them while executing tasks, and digital artifacts properties determine the quality of their design. DRIVE-T is intended to be used with task-based items related to an object of literacy (being it a text, a data viz, or a prototype interface). In this paper, we showcase its effectiveness in the domain of data visualization, though its rationale and rigorous validation make it generalizable to the domain of HCI quality design.

To the best of our knowledge, the MFRM model has neither been applied to the data viz literacy domain nor to the problem of task-based measurement constructs design. In DRIVE-T model, examinees are the data viz, and what is measured is their ability to discriminate task-based skills and represent data viz literacy levels. Following DRIVE-T, we were able to select a minimum but complete set of task-based items per data viz from our item bank, based on their discriminability and representativeness. We administered them to a sample of 75 high-school students. The aim of this pilot study was to show the

quality of the assessment test proposed based on the selected items, and their ability to measure data viz literacy properly, thus serving as a proof of concept for the DRIVE-T.

The paper is structured as follows: Section 2 analyzes past work on task-based item design in relation to computational literacies and design quality; Section 3 describes DRIVE-T; Section 4 reports the results of our methodology, and those of the pilot study; Section 5 discusses the results of applying DRIVE-T to the design of an assessment test, of using a pilot study, and outlines the limitations of the study. Section 6 reports final considerations.

2. Related Work

2.1. Quality Design: A Matter of Literacy

Designing high-quality digital interfaces requires more than technical usability: it requires the systematic ability to intercept (i.e., elicit, understand, and model) the diverse digital literacies that users bring into interaction. Digital literacy is not an external user attribute that can be assumed away; it is entangled with interface quality because interface artifacts actively instantiate literacy demands, shaping what counts as intelligible action, what is perceivable, and what forms of reasoning are possible [12]. This entanglement becomes especially visible when interfaces rely on data visualizations, where comprehension depends on a specialized form of digital literacy.

In this context, failing to account for visualization literacy risks producing interfaces that amplify inequalities: digital health systems, for example, can expand access in principle, yet still reproduce exclusion when users cannot interpret the visual or informational formats through which care is mediated [13].

To mitigate this gap, literacy assessments and screening instruments can function as boundary objects [14,15]: standardized artifacts that are robust enough to be shared across communities (designers, stakeholders, and users) while remaining interpretable within different perspectives and professional languages. Nevertheless, the methodological limitations regarding assessment tests design remain substantial. First, many instruments rely predominantly on self-assessment rather than performance evidence. Second, even when multiple dimensions are defined, they are often treated as separable traits, without a robust theoretical and psychometric account of their interdependence, ordering, or cognitive difficulty [9].

A promising direction towards a construct-based view is the attempt to integrate literacy dimensions and prioritize them within a developmental or structural model (e.g., [16]). Unlike purely descriptive definitions, constructs-driven approaches are aligned with a measurement goal: they implicitly pose questions about what the meaningful levels of CL might be, how acquired levels of skills support each other, and which skills precede others. However, even in these more integrated proposals, the field still faces a major construct modelling gap. Computational Literacy (CL) is often invoked as a label that groups skills, rather than as a rigorously defined latent construct with an explicit cognitive model. The above gap becomes even more visible when CL is examined through a broader cognitive lens. In practice, CL performance is shaped not only by problem-solving steps but also by background knowledge, domain familiarity, exposure to similar problem structures, and fluency in manipulating representations. These aspects overlap with notions of computational literacy and computational fluency, which often remain implicit in CT models [17].

This leads to a second problem: a measurement maturity gap. Existing design quality assessment instruments are often weakly connected to explicit cognitive or psychometric models that can explain why certain problems are hard, why certain skills cluster, and how mastery develops. Literacy definitions frequently describe what individuals should be able to do, but do not clarify what makes a task more or less demanding. These contributions

build a rich horizon of perspectives, but still do not converge into a shared model that supports stable measurement and comparability.

A third and often overlooked limitation concerns the fact that usability is assessed and enacted through interactive systems. This introduces HCI quality dependency gap: performance in usability tasks may reflect not only the users' skill, but also the quality of interaction design that mediates the task. If the interface hides representations, imposes unnecessary cognitive friction, or makes exploration and debugging difficult, then observed "users' deficits" may actually be usability failures. Conversely, well designed tools can scaffold decomposition, make abstraction visible, support iterative exploration, and reduce extraneous load, thereby enabling users' skills to manifest more clearly. Therefore, any serious modelling and assessment of users' skills must treat interaction quality as a structural component of validity, rather than as a secondary implementation detail.

The concept of "EUDability" has been proposed precisely to bridge these gaps, by modelling how an End-User Development (EUD) environment enables or inhibits the enactment of CL-related skills and development actions [1,7]. Instead of asking only "how skilled is the user?," EUDability also asks "how well does the system support the skill?," explicitly integrating user capability and tool affordances. This approach provides a framework for interpreting performance evidence in a design aware manner, thereby helping to address both the measurement maturity gap (by connecting CL skills to observable behaviours and system conditions) and the HCI quality dependency gap (by treating interface quality as a determinant of measurable CL expression). In such contexts, users' literacy should be operationalized as an interaction aware construct: skills must be measurable through observable patterns of action, supported by adequate representations, feedback mechanisms, and affordances for reuse, debugging, and evaluation. In this sense, the quality design of HCI is strictly related to construct modelling and assessment, because it controls what counts as evidence of quality in the first place.

De Souza [12] first introduced her semiotic engineering framework to real interactive systems seen as meta-communicative artifacts, emphasizing that designers and users engage in an indirect yet profound communication through the system's interface. According to De Souza, the interface embodies the designers' communicative intent, expressing their understanding of the users' needs, preferences, and contexts of use. By adopting this perspective, the foregrounded idea is that effective interaction design is inherently about effectiveness in this communicative relationship, ensuring that users correctly interpret the designers' intended messages, which are encoded using syntactic, semantic, and pragmatic dimensions. Furthermore, designers do deeply reflect on the meanings embedded in their interfaces, advocating that interfaces should explicitly facilitate user interpretation and sensemaking. Usability extends beyond mere functionality to include the clarity and coherence of the interface communicative content. Therefore, in semiotic engineering, evaluation of interactive systems explicitly involves assessing how clearly the designers' intentions are communicated and how easily users can interpret and apply these intentions to accomplish their tasks effectively.

2.2. Mapping the Data Viz Literacy Terrain

As outlined in the introduction, in data viz literacy modelling many theories were proposed, but none seemed to have survived nor have become the ground truth in the community [6]. As a measurable property, data viz literacy should be conceptualised, operationalised, and assessed by proper tests [18–29]. However, as anticipated in the Introduction, Bloom's taxonomy, Curcio's levels of visualization literacy, and Pinker's model of graphical comprehension offer important cognitive insights, but were criticised for being too abstract and excessively rigid [23]. Bloom's taxonomy, while useful for educational

objectives, presents cognitive processes in rigid hierarchical structures that overlook the dynamic and fluid nature of human cognition and learning. Curcio's visualization literacy levels tend to compartmentalise visual understanding, inadequately addressing the holistic integration of cognitive and communicative skills needed for effective data viz literacy. Similarly, Pinker's graphical comprehension model primarily addresses cognitive decoding of graphical elements but lacks a comprehensive integration of communicative and social value dimensions essential for evolving human interaction models [30].

2.3. Contribution to the State of the Art

In sum, the current landscape of HCI quality design research advocated for the importance of assessing users' skills and artifacts affordances, but lacks integration at the level required for construct modelling and mature assessment. EUD and EUDability provide a particularly relevant pathway to advance this agenda, because they frame computational literacy as a situated and design mediated capability, opening the way to construct driven, methodologically mature, and HCI aware assessment. Also semiotic engineering delve into the importance of an effective communication between designer and user, relying on the expressive potential of language and the mastery of it. In this work, we pose that users' skills and artifacts affordances may be the embroidery of visual and symbolic knowledge, educational background and practical experience. Semiotics is a fluid enough framework, where the mastery of one layer (e.g., syntactic) of a language of signs does not necessarily entail the mastery of a "subsequent" layer (e.g., semantics). This underspecification assumption and relative independence of levels makes semiotics a more robust, well-grounded and generalizable hypothesis to validate models of users' literacy and HCI quality design in general.

Existing task-based literacy approaches in visualization and adjacent HCI domains define task families and item banks to probe comprehension (e.g., VLAT/MINI-VLAT and related instruments [5,21,31]), but they typically do not provide a design-time methodology to (i) verify that items are discriminative across progressive levels and (ii) ensure that a minimal item subset is representative of the construct's targeted traits before administering a test. DRIVE-T addresses this gap by operationalizing design-time item screening through expert difficulty ratings, coupled with a Many-Facet Rasch Measurement model that places artifacts and their task-based item combinations on a common difficulty/discriminability continuum. Traditional Rasch-based test development commonly assumes a sufficiently specified construct model and uses respondent data to calibrate item difficulty and fit post-hoc [9]. In contrast, DRIVE-T explicitly targets the earlier phase of construct operationalization: it supports an inductive, practice-based progression from tagged task-based items to emergent construct levels, using rater-mediated difficulty judgments as the initial evidence base. Furthermore, while MFRM is widely used to account for rater severity/leniency in rater-mediated assessments [32–34], DRIVE-T uses MFRM specifically to jointly model rater effects, task difficulty, and artifact/item-combination difficulty to support item selection for representativeness and discriminability at design time, before respondent administration.

Table 1 helps frame DRIVE-T with existing approaches, positioning our approach and its novelty in the HCI literacy constructs development and assessments landscape. From the table it emerges how existing task-based literacy assessment frameworks and traditional Rasch-based test development approaches differ from DRIVE-T. While prior work primarily focuses on validating items after respondent administration, DRIVE-T explicitly targets the earlier phase of construct operationalization, supporting the inductive emergence of discriminative and representative items at design time.

Table 1. Comparison between DRIVE-T and existing approaches.

Dimension	Task-Based HCI	Rasch Test Design	DRIVE-T
Construct specification	Implicit/theory-driven	Explicit, fixed a priori	Inductive, emergent from practice
Item selection timing	Post hoc/heuristic	Post hoc	Design time (before respondents)
Item evaluation	Performance only	Performance + fit	Expert difficulty + discrim. + represent.
Role of raters	Not modeled	Rare/auxiliary	Explicitly modeled (MFRM)
Object of measurement	Users	Users	Artifacts + task-based item combinations

Our contribution is:

- The identification of task-based literacy items, characterized by discrimnability and representativeness for measuring levels of data viz literacy, before administering them;
- A modelling approach to flexible, situated and formative-style measurement constructs design based on Many-Facets Rasch Measurement (MFRM);
- A set of items selected with DRIVE-T and administered in a pilot test, analyzed and validated with Rasch.

3. Method

This work tries to mitigate the problem of having a full measurement construct model of data viz literacy beforehand, in favor of a more "formative style", lazy-learning and inductive approach based on the design of new/tagging of existing items, which are supposed to measure different levels of the data viz literacy latent construct. In this work, each item should challenge the respondent to express a capability related to its data viz literacy level, but we do not know a priori what is the order of each data viz literacy level on a measurement scale. We only hypothesize that there are levels on a scale, and each level should be measured by a set of items. On the contrary, current methodologies in social science propose to have a full construct hypothesis in mind beforehand [9], and to build items based on it.

In our view, in the mingle of practice, these constraints are often blurred, relaxed, and less systematic than in theory. By considering the possibility of a more situated, still rigorous tool for item reliability checking, we may benefit the whole test-designers and test-takers community. Indeed, we would like to support the co-evolution of designing/refining/adjusting construct traits and items, and to help test-designers validate items/integrate explicit construct traits in their practice. In its turn, the validated construct should reflect and justify the choice of what data viz and items should compose a data viz literacy assessment test, according to the criteria of discriminability and representativeness. As introduced in Section 1, discriminability refers to an item's ability to distinguish among varying levels of data viz literacy in the target population, and representativeness determines what is the minimum set of objects of literacy that should be selected among those available into an item bank, to uniquely but completely represent each measurable trait of the data viz literacy construct. The discriminability and representativeness of items may support test-designers outline on-the-fly hypotheses about how many items should be used for measuring data viz literacy and, based on that, decide for a shorter or longer version of an assessment test, by grounding their decision on objective and justified criteria. As told in the Introduction, semiotics provides an operational framework for relating artifact properties to observable task performance by distinguishing different sources of meaning making that are embedded in interactive artifacts. Following Eco, artifacts can be understood as structured sign systems whose interpretation relies on partially shared conventions rather than on purely formal rules or idiosyncratic interpretations. In the context of interactive and visual artifacts, this implies that difficulty does not stem exclusively from data complexity, but from the extent to which an artifact's representational conventions align with users' acquired semiotic competences. The semi-symbolic perspective is therefore well suited to

support task-based assessment because it allows tasks to be defined in terms of what kind of semiotic work an artifact requires from a user, rather than in terms of abstract cognitive skills. In DRIVE-T, this perspective is operationalized by applying semiotic tasks directly to concrete artifacts, treating them as the primary objects of analysis.

The DRIVE-T methodology is made of three steps, which are depicted in Figure 1. Step 1 outlines a hypothesis of levels of data viz literacy borrowed from semiotics, and describes the process of tagging items according to semiotics tasks. Step 2 describes the process of rating them by judging/rating their difficulty. Step 3 supports the modelling of raters' scores with MFRM. Each step is detailed in the next subsections.

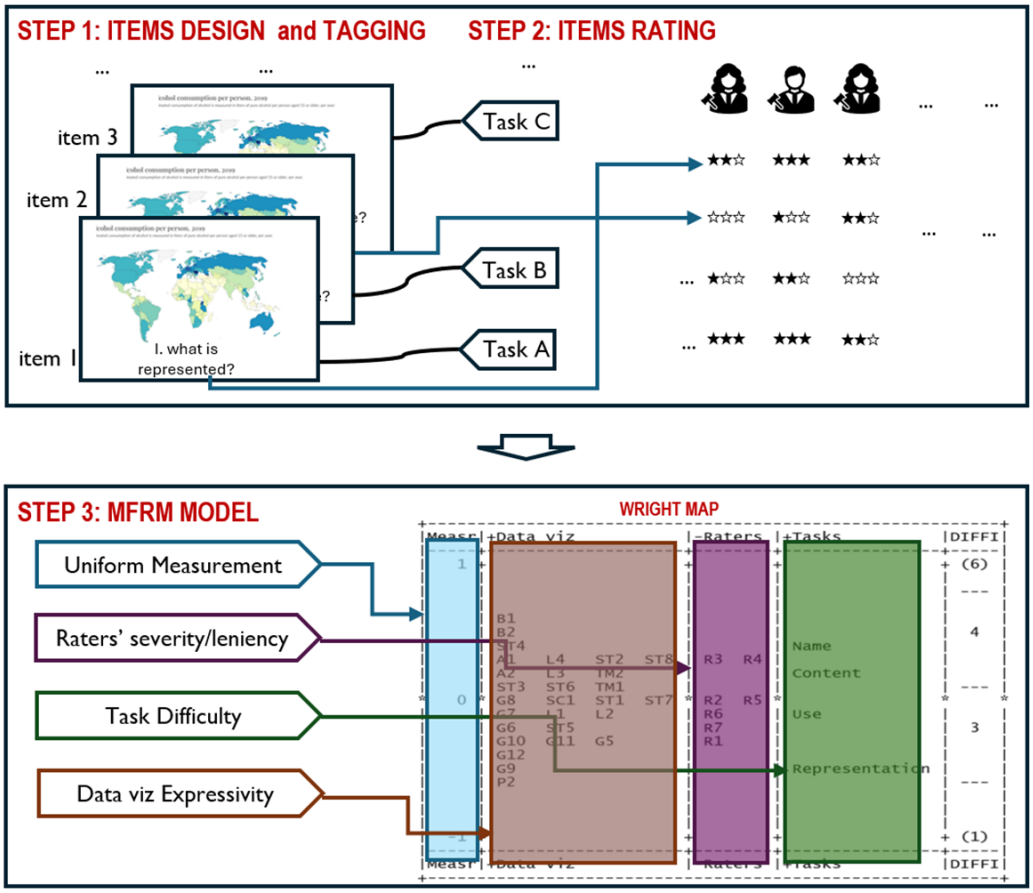


Figure 1. The DRIVE-T methodology and its steps.

3.1. Item Design and Tagging: Step 1

Merging the two aspects of task-based literacy and semiotics, we propose a construct modelling activity characterised by four task-based item design, where each task references one of the semiotic aspects of data viz language. Syntactic tasks are those that query the user about *what a data viz represents* and strongly resonate with the structural, grammatical rule-based understanding of a formal system of signs, such as that of data viz. Semantic tasks have to do with the data viz content, corresponding to a technical understanding of how to engage in a fruitful interaction with them. Such tasks extend beyond simple interaction-level manipulations to deeper interpretative actions, enabling users to extract informational gist. Pragmatic tasks embody reflective judgment in situated interactions. These tasks encourage users to actively consider the implications and appropriateness of a data viz to serve, within a social context. Knowing a data viz name may reveal a further, more profound level of knowledge, as knowing the name of an object is strongly associated with the ideal "knowing it all" about it (cf. the semiotician Umberto Eco in his thesis about

the famous Gertrude Stein's verse "A rose is a rose is a rose", i.e., calling a thing with its name is showing meaning and knowledge of it.).

These task types are not assumed to form a fixed hierarchy; instead, DRIVE-T allows their relative difficulty and discriminative power to emerge empirically through expert ratings and Rasch-based modeling. This alignment enables the identification of difficulty progressions grounded in artifact–task interactions, rather than imposed a priori by cognitive theory.

3.2. Item Scoring: Step 2

Rater-mediated assessments are patently influencing critical decisions in educational, psychological, and hiring contexts, to name only a few. These assessments involve humans judging the performance, literacy, or other skills of the examinees. The problem with raters' evaluation in these kinds of tests is that it is fed with inevitably arbitrary behaviours such as interpretation, discretion, and subjectivity. Raters' evaluations may vary due to differences in interpretation, severity or leniency biases, and tendencies such as central tendency or halo effects biases [35].

In the domain of assessment tests construction, raters may play a crucial role in acting as a self- or external validation device, able to provide a reliable pre-screening of how item design is gauged or needs to be adjusted. When not carefully managed, raters' variability can compromise the validity of this process. As much as rigorous is the raters' training, issues of the above kinds are not avoidable, hence a careful policy should be adopted in order to include raters' modelling in item design and assessment tasks.

Inter-rater (irr) reliability indices, available in common statistical packages, provide critical insights into rater agreement and consistency. Moreover, these indices have some limitations and paradoxes [36]. Furthermore, during item design or post-design analysis, there is the need to critically consider each item singularly. None of the methods above provide a fine-grained analysis of the raters' scoring able to shed light on their level of disagreements by individual items. For this reason, we are not going to carry out any irr computation and analysis. In this work, we apply an alternative analysis to the irr instead, by modelling the severity (resp. leniency) of raters with Rasch-based models. A brief presentation of individual items raw score differences and their visualization are proposed in Section 4. However, this score analysis is intended only to further comment on the main proposal of this work, i.e., providing a unified model of data viz, tasks, and raters. Indeed, some justifications of the model may be derived by a pinpoint identification of the controversies at item-level, where considerations may be drawn about the concurring aspects leading to disagreement (e.g., by comparing type, task-related activity or data viz choice).

Raters Briefing Protocol

Before rating, raters received (i) the definition of each task type (*Represent*, *Content*, *Use*, and *Name*) and its intent within the construct hypothesis, and (ii) the difficulty scale with the associated probability of correct intervals (see Table 2), to anchor interpretations of "easiness/difficulty" to an explicit performance expectation. Raters were instructed to score each item independently, without comparing it to other items, and to treat difficulty as the expected likelihood that a responder in the target population would answer the item correctly.

Table 2. Category labels, scores and probabilities of right response in percentage.

Category Label	Score	Probability %
Very Easy	1	95–100
Fairly Easy	2	75–94
Middle Easy	3	50–74
Middle Difficult	4	30–49
Fairly Difficult	5	6–29
Very Difficult	6	0–5

3.3. Modelling Raters' Scores: Step 3

The family of Rasch models consists of a set of probabilistic models that are able to convert raw scores into linear and reproducible measures of the latent trait of interest, for example data viz literacy [37]. Fundamental assumpions are unidimensionality and local independence. The former refers to the requirement that all items within a questionnaire measure a single underlying construct, for example data viz literacy. The latter implies that, conditional on the latent trait, the responses to any given item are statistically independent of the responses to all other items in a questionnaire. When data align with the model, the yielded measures are supposed to be objective and expressed on a logit (log-odds) scale, which is an interval scale.

In the Rasch model (RM) [38], the probability that subject n answers correctly ($c = 1$) or incorrectly ($c = 0$) to item i, depends on her/his level of latent trait (or ability), θ_n, and on item difficulty, β_i, according to the following formula:

$$P(X_n = c) = \frac{exp\{c(\theta_n - \beta_i)\}}{1 + exp\{\theta_n - \beta_i\}} \qquad c = 0, 1 \tag{1}$$

The original RM was proposed to analyse tests composed exclusively of dichotomous items. Later, RM was extended to account for: (i) polytomous items, such as those on Likert scales, through the Rating Scale Model (RSM) [39], and the Partial Credit Model (PCM) [40]; and (ii) raters, who judged subjects (or examinees), through the Many-facet Rasch Measurement (MFRM) model [32].

In this paper, we will exploit the PCM and the MFRM model; the following subsections will provide their brief introduction.

3.3.1. The Partial Credit Model

The PCM is defined as follows. Given an item i with $m + 1$ response categories ($c = 0, 1, \cdots, m$), the probability of the subject n to respond in category c is given by:

$$P(X_{ni} = c) = \frac{exp\left\{c(\theta_n - \beta_i) - \sum_{j=0}^{c} \tau_{ij}\right\}}{\sum_{l=0}^{m} exp\left\{l(\theta_n - \beta_i) - \sum_{j=0}^{l} \tau_{ij})\right\}} \tag{2}$$

where θ_n denotes the measure of the latent trait (or ability) of the subject n, β_i represents the difficulty of the item i and τ_{ij}, called thresholds, is the point of equal probability of categories $j - 1$ and j ($\tau_{i0} \equiv 0$ and $\sum_{j=1}^{m} \tau_{ij} = 0$). The thresholds can be different for all the items.

Quality of PCM measures is obtained through the Person Reliability Index, the Outfit mean-square statistic (*Outfit MNSQ*), and the Infit mean-square statistic (*Infit MNSQ*) [34,41].

The Person Reliability Index reflects how much the observed differences between subjects can be consistently replicated rather than being attributable to measurement error. It provides an estimate of the replicability of one person's placement along the latent construct that can be expected if the same sample were administered a different set of items measuring the same underlying trait. It ranges from 0 to 1 and is robust to missing data.

Item or subject *Infit MNSQ* and *Outfit MNSQ* assess the degree to which the observed data conform to the expectations of the Rasch analysis. They are based on standardised residuals and range between 0 and infinity, with expected value of 1. Values close to 1 indicate that the item or subject behaves as expected; values below 1 suggest overly predictable behavior, whereas substantially higher values indicate unpredictable or noisy behavior.

3.3.2. The Many-Facet Rasch Measurement Model

The MFRM model can be seen as an extension of the basic RM, to include more variables (or facets), such as, for instance, raters evaluating the performance of examinees on tasks. A minimum of three facets configures the need to use a model of this kind [33]. In this work, three facets were taken into account, i.e., examinees, raters and tasks. Thus, the MFRM model also accounts for raters' variability, i.e., the variability associated more with raters' characteristics than examinees' performance.

The MFRM model is formally expressed as:

$$\ln\left(\frac{p_{nijk}}{p_{nijk-1}}\right) = \theta_n - \beta_i - \alpha_j - \tau_k \tag{3}$$

where p_{nijk} is the probability that examinee n receives a ranking of k from rater j on task i, θ_n is the proficiency of examinee n, β_i is the difficulty of task i and α_j is the severity of rater j. τ_k is called the threshold parameter and represents the difficulty of receiving a rating of k relative to $k-1$ (thresholds add up to zero). The MFRM model defined in Equation (3) involves three facets, and the scoring of the tasks shares the same rating scale structure with threshold parameters calibrated jointly across raters, tasks, and examinees. Thus, its name is Three-Facet Rating Scale Model (3 FRSM) [42]. In order to make the model identifiable, some constraints on its parameters must be imposed. In this work, the rater and task facets were centered, i.e., their mean values were constrained to be zero.

In order to evaluate the variability of the measures within each facet, it is possible to use its corresponding separation reliability index (R). In general R indicates how well the model distinguishes between different levels of the facet, thereby supporting its meaningful definition. The index ranges from 0 to 1 and represents the proportion of observed variance in the facet measures that is attributable to true differences rather than to measurement error. The interpretation of R depends on the facet under consideration. At the examinee level, R reflects how effectively the model differentiates examinees according to their proficiency; therefore, higher values, indicating better discrimination, are desirable. At the item level, R describes how well items are spread along the difficulty continuum, with higher values indicating a wider and more informative range of task difficulties. Finally, at the rater (judge) level, R provides information about the degree of similarity or dissimilarity in rater severity or leniency. In this case, lower values are preferable, as they suggest that raters function in a more interchangeable manner.

To investigate the accuracy of the examinees, raters, or tasks measurements, one can use the *Outfit MNSQ* and *Infit MNSQ*, described in Section 3.3.1, and the point-measure correlation (*PtMea Corr*). *PtMea Corr* is computed as the Pearson correlation between the observed scores and the combined measures, where the combined measure for examinee n, rated by the rater j on task i is given by $\hat{\omega}_{nij} = \hat{\theta}_n - \hat{\beta}_i - \hat{\alpha}_j$ ($\hat{\theta}_n$, $\hat{\beta}_i$ and $\hat{\alpha}_j$ are the estimated parameters of Equation (3)). *PtMea Corr* offers insights into how observations align with the model expectations. When computed for the examinees, a positive and high value of the *PtMea Corr* supports the consistency of the observed ratings with the measures estimated by the model, whereas a negative value indicates a substantial departure from model expectations [35].

The model defined in Formula (3) does not account for interactions between facets. However, studying the examinee-by-rater interactions may be informative with respect to the fairness of the judgment process. This analysis allows the assessment of whether raters applied a uniform level of severity across examinees, or whether some raters scored examinees' performances in a manner that was harsher or more lenient than model expectations. This analysis may be performed by estimating the following model, which includes an examinee-by-rater interaction parameter ϕ_{nj}, also known as a bias parameter:

$$\ln\left(\frac{p_{nijk}}{p_{nijk-1}}\right) = \theta_n - \beta_i - \alpha_j - \phi_{nj} - \tau_k. \quad (4)$$

Evaluating the significance of the examinee-by-rater interaction parameter ϕ_{nj} is possible by using a bias statistic $t_{nj} = \hat{\phi}_{nj}/SE_{nj}$, where $\hat{\phi}_{nj}$ was the estimated bias parameter and SE_{nj} its standard error. This statistic is approximately distributed as a student t, under the null hypothesis that there is no bias apart from measurement error [35], i.e., raters behave fairly toward the examinees.

4. Results

In this section we report the results of applying DRIVE-T to our item bank and of selecting items for our pilot study.

4.1. DRIVE-T Steps Applied

In Step 1 of our visualization literacy assessment test, we decided to choose a subset of data viz related to the state-of-the-art assessment tests (see for example the MINI-VLAT test [31]): a stacked area chart (A), a bubble chart (B), a choropleth map (G), a line chart (L), a pie chart (P), a scatter plot (SC), a stacked bar chart (ST) and a tree map (TM). As pointed out in [6], most papers proposing a questionnaire to measure data viz literacy use tasks which require "purely visual operations or mental projections on a graphical representation" [5], such as, for example, identifying a maximum, a minimum, or a quantitative variation. As introduced in Section 3, we adopted a design approach with different aspects of comprehension in mind. These aspects are: the ability to read the data viz and understand what it represents (i.e., syntax), the ability to extract information from it (i.e., semantics), the knowledge of how it is used for serving its purpose (i.e., pragmatics), and knowing its name. We labeled the corresponding tasks as *Represent*, *Content*, *Use*, and *Name*, respectively.

The formulation of each task was data viz specific. In the production phase, we designed 44 task-based items for most of the data viz, as summarised in Table 3. An example of items, one for each task type, is reported in Figure 2 (the items are available at https://osf.io/ngu5q/?view_only=9f4144dcec4e49dabbebc332f539b3c8 (last accessed on 31 January 2026)). The items are in Italian despite an English informal translation is provided as well.

Multiple-choice items with three possible answers (and only one correct) were 54.4% (24 items). Polytomous items on a 3-point Likert scale were 13.6% (6 items). True/false were 11.4% (5 items). One item had a free text answer. The last 8 items (20.6%) were those connected with the task *Name*, and they were free text. All items included the *I don't know* option to prevent guessing.

To evaluate the quality of the items and decide which should compose the final questionnaire, we conducted a performance evaluation of the items and data viz, involving seven expert raters (judges). Raters were asked to assess each of the 44 items based on a six-point Likert scale of difficulty, from very easy to very difficult, shown in Table 2. To facilitate the rating activity and reduce discrepancies among raters, we defined a set of

probability intervals for the probability of correct response, as shown in the table. Each rater was asked to score the difficulty of each item without comparing it to the other items and to each other, as the focus was on determining the level of difficulty of each individual item on the provided scale. At the end of this procedure, we obtained a set of 44 evaluations for each rater.

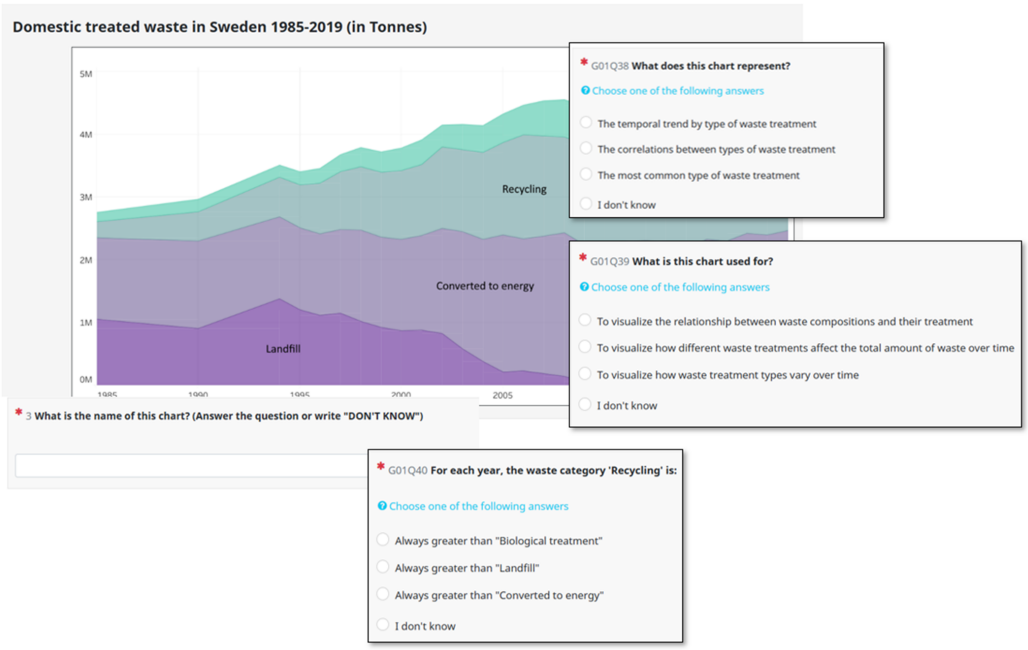


Figure 2. Example of 4 task-based items designed for one of the 8 data viz included in our 44 task-based item bank. Each item in the figure corresponds to a task among: Name (bottom left item), Content (bottom right item), Represent (upper right item), and Use (rightward item).

Table 3. Distribution of item versions by task type and associated data viz format.

Data Viz	Name	Represent	Use	Content
A	1	1	1	2
B	1	1	1	2
G	1	2	2	2
L	1	2	1	2
P	1	1	1	2
SC	1	1	1	1
ST	1	2	2	2
TM	1	1	1	2

4.1.1. The Raters' Score

In this study, the seven raters were recruited among scholars of two universities and among external advisors, with different expertise in data visualization. Two raters are in the research domain of data visualization (R3 and R4), three are in the research domain of statistics (R5, R6, and R7), and two are in the domain of informatics engineering and machine learning (R1 and R2). The single performance of raters concerning their use of the scale and the distribution of scores along the range of values is depicted in the violin plots of Figure 3. Violin plots overlap to a more common box plot a distribution density that may help identify nuances in the use of scores, without looking at the y axes.

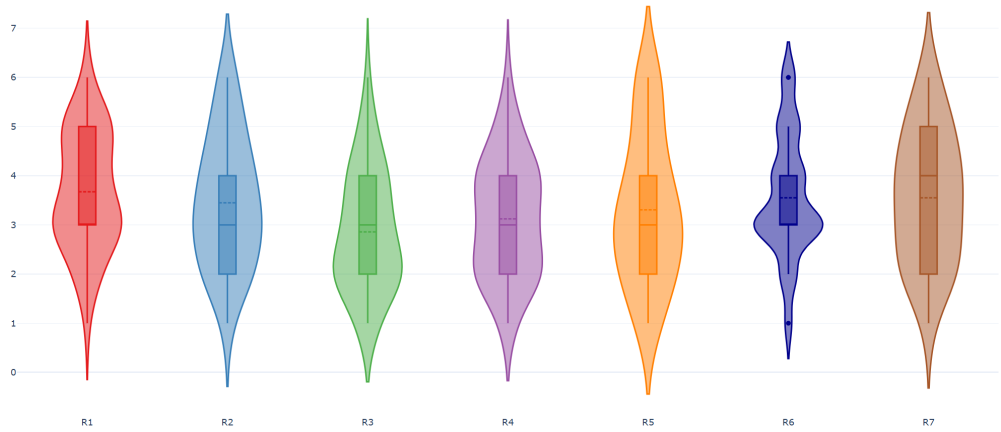


Figure 3. Violin plots of the raters' scoring distributions.

As anticipated in Section 3.2, no irr analysis was conducted on the raters' scores, but an item level analysis of raters' scores absolute differences. The results are reported in Figures 4 and 5, where each heatmap is showing the score difference between ordered pair of raters. The higher the difference the lighter/darker the color of each cell of the heatmap (yellow for max positive differences and dark blue for max negative differences). Considering a difference of agreement of 0 as perfect agreement, of ± 1 as very low disagreement, of ± 2 as low disagreement, of ± 3 as medium disagreement, and ± 4 or ± 5 as high disagreement, the analysis of the score differences among raters also reveals the degree of disagreement, whilst the sign of the difference allows to compare the severity/leniency of raters for each item. In particular, when the pairwise comparison gives a negative score difference, a darker green-to-blue color, this means that the row raters were more severe than the column raters (i.e., the row rater gave a lower score than the column rater); on the contrary, a positive score difference, a lighter green-to-yellow color, means that row raters were more lenient than column raters (i.e., the row rater gave a higher rating to the item, considering it more difficult).

4.1.2. The Many-Facet Rasch Measurement Model of Our Item Bank

In this section, we present the results obtained by estimating the 3 FRSM, as defined in Equation (3), using the joint maximum likelihood estimation method (This analysis was performed using the FACET software, version 4.4.5 [43], a common package mostly used in RM analyses.).

We distilled, from the initial banks of 44 items, 4 task-based items for each of the 8 data viz. We created a set of 29 records, each consisting of the combination of one data viz, one item for the *Name* task, one item for the *Represent* task, one item for the *Use* task, and one item for the *Content* task.

Hereafter, we use the term combination for the byproduct of a data viz and its four tasks, and identify each combination with the Label column of Table 4 (e.g., A1, A2, and the like). Table 4 also reports, on each line, the item number of each task. These 29 labels identify the 29 examinees (data viz combination) of our analysis, rated by the seven raters (judges) on the four tasks.

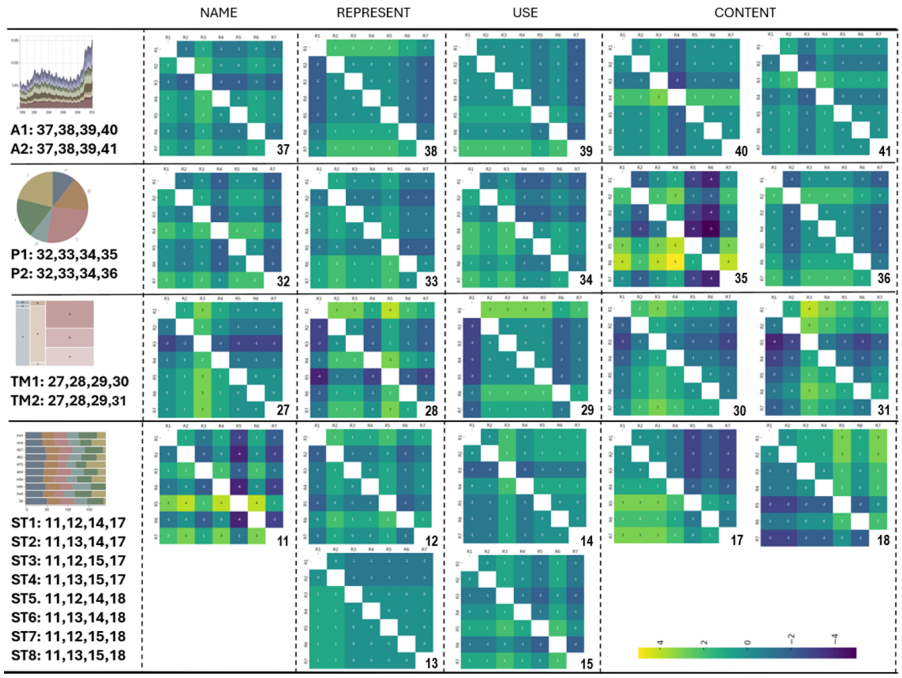


Figure 4. Heatmaps compare data viz task-based items by raters' ordered pairs. Each heatmap shows: at the top left rater R1 and at the bottom left rater R7, as per row disposition; at the top left rater R1 and at the top right rater R7, as per columns disposition. The data viz presented in the figure serve illustrative purpose only; they are of the same kind of those employed in the real test: A, P, TM and ST. The item ID is the number at the bottom right of each heatmap.

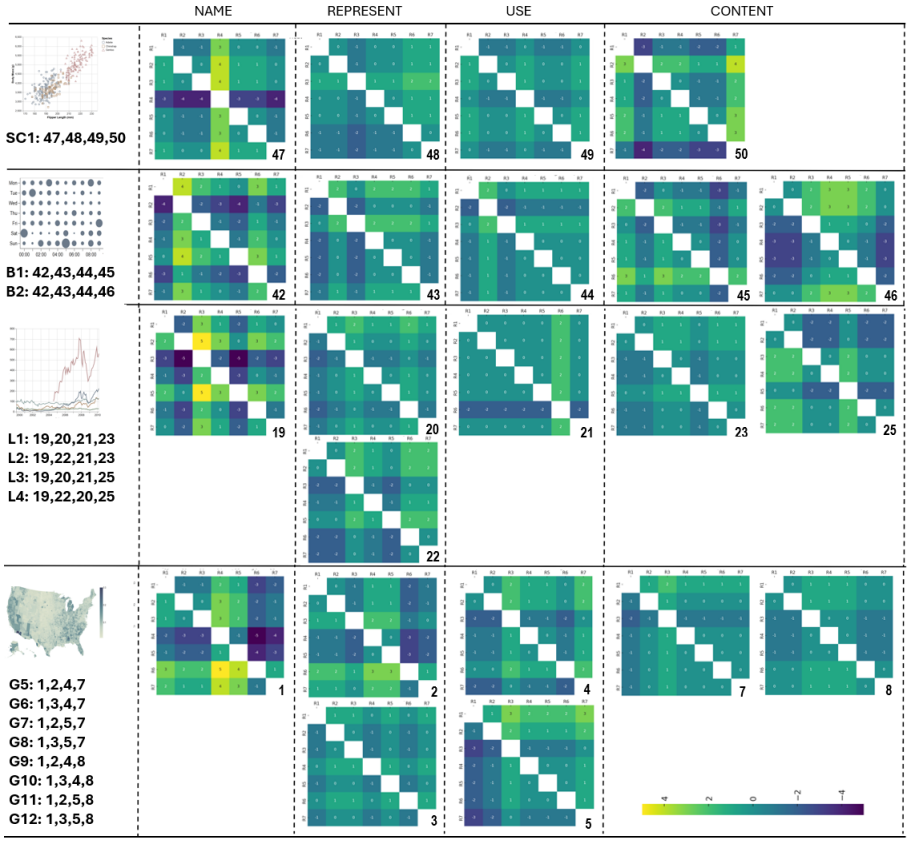


Figure 5. Heatmaps with same arrangement as Figure 4. The data viz kinds here represented are: SC, B, L and G.

Table 4. Combinations of data viz, *Name*, *Represent*, *Use* and *Content* tasks. The numbers refer to the numbers associated with each item.

Data Viz	Label	Name	Represent	Use	Content
A	A1	37	38	39	40
	A2	37	38	39	41
B	B1	42	43	44	45
	B2	42	43	44	46
G	G5	1	2	4	7
	G6	1	3	4	7
	G7	1	2	5	7
	G8	1	3	5	7
	G9	1	2	4	8
	G10	1	3	4	8
	G11	1	2	5	8
	G12	1	3	5	8
L	L1	19	20	21	23
	L2	19	22	21	23
	L3	19	20	21	25
	L4	19	22	21	25
P	P1	32	34	33	35
	P2	32	34	33	36
SC	SC1	47	48	49	50
ST	ST1	11	12	14	17
	ST2	11	13	14	17
	ST3	11	12	15	17
	ST4	11	13	15	17
	ST5	11	12	14	18
	ST6	11	13	14	18
	ST7	11	12	15	18
	ST8	11	13	15	18
TM	TM1	27	28	29	30
	TM2	27	28	29	31

As anticipated in Section 1, in this study the role of the examinees was played by the data viz combinations. The hypothesis is that, given a data viz combination:

- The more it is judged to be difficult, the higher its discriminative power in assessing a person's data viz literacy level;
- The more it is complete and minimal (only one item per task, all the tasks being considered), the more representative it is.

In our model, the data viz combinations were measured positively, i.e., high raw scores meant a high degree of discriminability. The tasks were measured positively too, i.e., high difficulty measures corresponded to high scores. This result implies that, in the Formula (3), the item term was added instead of subtracted. In contrast, raters were modeled such that higher severity corresponded to lower scores assigned to data visualizations.

The estimates of the threshold parameters τ_k resulted as in the following: $\tau_1 = -1.54$, $\tau_2 = -0.33$, $\tau_3 = 0.16$, $\tau_4 = 0.88$, $\tau_5 = 0.83$. The last two thresholds were very close and inverted, causing a disorder in the thresholds progression, which no longer advanced monotonically with the response categories (the ratings). Even if the requirement of the ordered thresholds is not part of the model constraints, we decided to overcome the problem by collapsing the first two and the last two categories, yielding a four-point Likert scale for rating difficulties. The resulting new thresholds resulted as no longer disordered.

A summary of the expected outcome statistics of the 3 MFRM analysis is reported in Table 5.

Table 5. Discrimination abilities and fit statistics of the data viz combinations in the 3 MFRM model.

Data Viz Combination	Measure	Infit MNSQ	Outfit MNSQ	PtMea Corr
B1	0.52	0.93	1.02	0.18
B2	0.40	1.18	1.22	0.22
ST4	0.25	1.09	1.19	0.02
ST8	0.18	1.26	1.35	−0.07
L4	0.14	0.94	0.91	0.31
L3	0.10	0.81	0.79	0.40
ST2	0.07	1.04	1.09	0.04
A1	0.03	0.77	0.77	0.65
ST6	−0.00	1.19	1.24	−0.06
TM2	−0.00	1.04	1.04	0.64
A2	−0.04	0.79	0.78	0.63
ST3	−0.11	1.08	1.05	0.44
TM1	−0.11	1.08	1.07	0.60
G7	−0.18	0.80	0.79	0.53
SC1	−0.18	1.24	1.19	0.55
ST1	−0.18	1.04	1.01	0.74
ST7	−0.18	1.23	1.18	0.35
G8	−0.22	0.76	0.73	0.58
L2	−0.29	1.00	1.01	0.20
L1	−0.33	0.85	0.83	0.28
G5	−0.37	0.80	0.79	0.55
ST5	−0.37	1.09	1.03	0.39
G6	−0.41	0.75	0.71	0.60
G12	−0.52	1.04	1.00	0.44
G10	−0.57	0.93	0.91	0.39
G11	−0.61	0.87	0.81	0.44
P2	−0.69	1.16	1.26	0.01
G9	−0.83	0.77	0.71	0.50
P1	−0.88	1.51	1.67	−0.12

The data viz, raters and tasks facets can be displayed in a single visualization called *Wright map* (Figure 6), as long as all three measurements are expressed in logit [35]. This map is very informative because it facilitates comparisons within and between the various facets. It allows us to inspect whether the task-based items cover a meaningful range of difficulty levels and whether data viz combinations are well distributed along the discriminability contunuum. The *Measr* column displays the measurement scale. The *Data viz Combination* column displays the estimated measure of discrimination ability of each data viz combination, arranged with those at the top having the highest discriminability. The *Raters* column displays the estimated measure of the raters' level of severity; raters appearing at higher positions exhibit greater severity. The *Tasks* column displays the average measure of difficulty of the four tasks considered; tasks located higher in the column are more difficult than those positioned lower. The last column maps the four-category scale to the equal-interval logit scale.

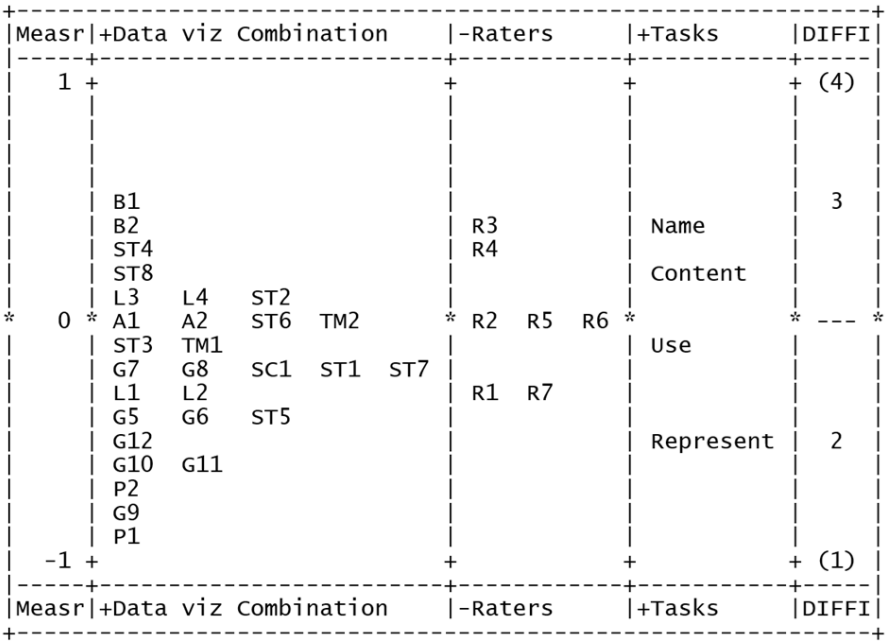


Figure 6. Wright map from the many-facet rating scale analysis. The *Data viz Combination* column displays the estimated measure of discrimination ability of the data viz combinations, the *Raters* column displays the estimated measure of the raters' level of severity, the *Tasks* column displays the average measure of difficulty of the tasks and the *DIFFI* column maps the four-category scale to the equal-interval logit scale.

4.1.3. A Proposed Test to Measure Data Viz Literacy

One of the final outcomes of the DRIVE-T methodology was a more justified and robust selection of a subsample of items from our item bank, based on the rating activity of a group of raters. The analysis of the data performed by estimating the 3 FRSM, came up with the Write map of Figure 6. From the inspection of the *Data viz Combination* column, and on the basis of the 3 FRSM analysis, it was possible to choose one combination of items per data viz, such that the corresponding measures lay on a sort of continuum, with increasing measures. The strategy was, first, to select the most discriminating combinations, namely B1 and ST4. Second, SC1 was added as the only SC combination, and P2 was selected from the two combinations involving the pie chart because the other combination (P1) exhibited problematic fit, as indicated by elevated (though still within acceptable limits) Infit and Outfit MNSQ values and a negative *PtMea Corr*. Third, for A and TM, we selected the combinations with the highest measures not shared with other combinations, namely A1 and TM1. We then selected G12 as the G combination, as it occupied a mid position on the discriminability scale among five other G configurations. Lastly, L3 was selected among the L combinations because, of the two with the highest discriminability power, it showed the highest *PtMea Corr*. The resulting selection should be interpreted as the outcome of a heuristic and iterative procedure, in which informed choices were made by jointly considering discriminability, model fit, and measurement coverage, rather than enforcing a strictly optimal selection rule. Therefore, the final set of items consisted of 17 multiple-choice items, 5 polytomous items, one true/false item and 9 free-text items. The total number of selected items was 32. Table 6 reports, for each data viz, the number associated with the selected items and the corresponding codes that were used in the subsequent Rasch analysis.

Table 6. List of selected items by data viz and tasks. The first rows report the numbers associated with the items, whereas the second rows the codes used in the Rasch analysis.

Data Viz		Name	Represent	Use	Content
A	Item	37	38	39	40
	Code	A	A_Repr	A_Use	A_Cont
B	Item	42	43	44	45
	Code	B	B_Repr	B_Use	B_Cont
G	Item	1	3	5	8
	Code	G	G_Repr	G_Use	G_Cont
L	Item	19	20	21	25
	Code	L	L_Repr	L_Use	L_Cont
P	Item	32	34	33	36
	Code	P	P_Repr	P_Use	P_Cont
SC	Item	47	48	49	50
	Code	SC	SC_Repr	SC_Use	SC_Cont
ST	Item	11	13	15	17
	Code	ST	ST_Repr	ST_Use	ST_Cont
TM	Item	27	28	29	30
	Code	TM	TM_Repr	TM_Use	TM_Cont

4.2. The Pilot Study

In this section, we present the results of applying the final outcome of the DRIVE-T methodology to a group of Italian high-school students. The aim of this pilot study is to provide a proof of concept for the proposed methodology. The items identified in Section 4.1.3 were administered to two high-school students in a private session with one of the authors, for adjusting the wording of the items. The items were then administered in a questionnaire to a sample of 75 high-school students, all attending the 4th or 5th grade classes of an Italian scientific high-school (60), and an Italian technical high-school (15). In particular, the latter students underwent a mini-course about data viz design (10 h) some weeks before taking the test. The former did not undergo any previous specific course about data visualization, but the professor responsible for their vigilance during the test suggested that they may have some literacy about data viz due to their scientific curricula. The sample was equally divided into females and males. The age of the participant was between 18 and 19 years old.

In order to analyse the responses with the most suitable model belonging to the family of Rasch models, the 17 multiple-choice items, the true/false item and the 9 free text items were dichotomised, with code 1 for the correct answer and 0 for any other answer. Therefore, the item bank consisted of 27 dychotomous and 5 polytomous (TM_Use, P_Repr, P_Cont, B_Use and SC_Use) items. Given the nature of the items involved, comprising both dichotomous and polytomous formats, we employed the PCM to analyse the results, as it allows for the estimation of distinct thresholds across items. The mean of item difficulty estimates was set to 0.0 logits and the estimates of the parameters were obtained using the (unconditional) maximum likelihood estimation method (all the analyses were performed using the WINSTEP software, version 4.5.0 [44]).

The inspection of the thresholds for the five polytomous items revealed that the two thresholds of item SC_Use were inverted ($\tau_1 = 1.50$, $\tau_2 = -1.50$), whereas the two thresholds of item B_Use were too close ($\tau_1 = -0.03$, $\tau_2 = 0.03$). So, we decided to collapse, for both items, the first two categories, obtaining two dichotomous items.

As stated before, the successful implementation of PCM necessitates that the assumptions of local independence and unidimensionality are fulfilled, meaning that the items should approximate both local independence and unidimensionality. One tool to detect local independence is the item Pearson correlation, which was computed using residuals

across all subjects who responded to both items. Potentially locally dependent pairs of items should have high positive or negative correlations.

The principal component analysis (PCA) on standardised residuals was used to verify the assumption of unidimensionality (the Rasch method uses the PCA in a counterintuitive, yet valid way: excluding multi-dimensionality in the items.). In the Rasch framework, the purpose of conducting this analysis is not to identify shared factors. The underlying hypothesis is that the data contain only one dimension (the Rasch dimension) caught by the model, so the residuals should not contain other significant dimensions. Thus, the objective was to confirm the unidimensionality assumption [45]. The eigenvalues of the first two dimensions were 3.03 and 2.74, respectively, suggesting the possible presence of multidimensionality. Thus, a further analysis was necessary to understand whether the degree of multidimensionality present in the data was substantial enough to justify partitioning the items into separate tests. In order to investigate this issue, for each dimension, the items were clustered into three clusters according to high, low, and middle values of their factor loadings [44]. Usually, the middle cluster (2) roughly corresponds to the Rasch dimension. Thus, it was important to verify that the correlation between this cluster and the other two was high. Clusters 1 and 3 were both somewhat off-dimension, making the correlation between them somewhat incidental [44]. In order to account for the measurement error, the disattenuated correlation was used. The disattenuated correlation values for the 2-1 and 2-3 clusters were, respectively, 1 and 0.63 for the first dimension and 0.62 and 0.88 for the second dimension. These values were sufficiently high to indicate that the clusters of items were measuring the same thing.

The expected statistical outcomes of applying RM analysis to the pilot study data are available in Table 7.

Table 7. Task-based items difficulty and their fit statistics in the Rasch analysis of the pilot study.

Item	Measure	Infit MNSQ	Outfit MNSQ
A	4.54	0.86	0.35
TM	2.97	1.00	0.66
B	2.23	0.82	0.63
P_Repr	1.74	1.41	2.73
L	1.60	1.03	1.10
G	1.53	1.04	1.34
A_Cont	1.28	1.17	1.48
ST	0.87	1.00	0.94
ST_Cont	0.87	1.07	1.08
B_Cont	0.65	1.18	1.34
G_Repr	0.56	1.03	0.96
SC	0.42	0.89	0.84
TM_Repr	0.28	1.00	0.96
P_Cont	−0.04	1.25	1.24
L_Cont	−0.07	0.88	0.81
L_Repr	−0.44	1.03	1.09
L_Use	−0.52	1.24	1.27
B_Repr	−0.57	1.10	1.20
B_Use	−0.74	1.00	1.03
A_Use	−0.76	0.91	0.86
G_Cont	−0.77	0.69	0.55
SC_Cont	−0.85	0.88	0.80
TM_Use	−0.87	1.00	0.97
ST_Repr	−0.91	0.89	0.82
P_Use	−0.93	0.86	0.74
ST_Use	−1.31	0.71	0.56

Table 7. *Cont.*

Item	Measure	Infit MNSQ	Outfit MNSQ
TM_Cont	−1.36	0.65	0.52
G_Use	−1.41	1.05	0.85
A_Repr	−1.48	0.91	0.74
SC_Repr	−1.57	1.03	0.84
SC_Use	−1.61	0.79	0.70
P	−3.31	0.93	0.48

As done for the 3 FRSM, it is possible to display, in a single visualization called *Wright map*, the distribution of students' levels of data viz literacy and item difficulties, provided that both measures are expressed in logits. (Figure 7). In the map, the symbol "M" on the left-hand side indicates the mean level of students' data visualization literacy, while on the right-hand side it indicates the mean item difficulty, which is constrained to zero. For both data visualization literacy levels and item difficulties, "S" denotes one sample standard deviation from the mean and "T" denotes two sample standard deviations from the mean.

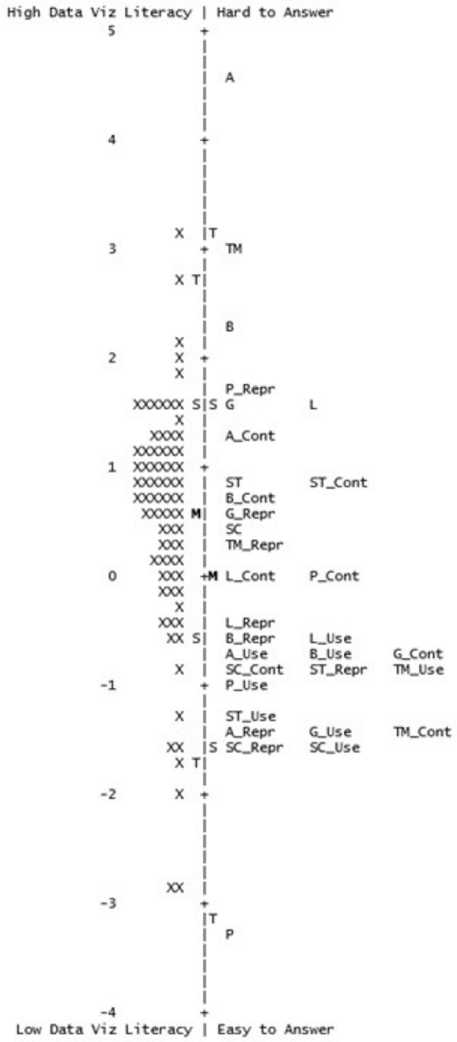


Figure 7. Wright map of the PCM analysis of the pilot study. Labels of each item contain only the data viz abbreviation for the *Name* task, abbreviation and *_Cont*, *_Repr*, and *_Use* for the other three tasks, respectively. "M" denotes the mean mean level of students' data visualization literacy on the left and the mean item difficulty on the right. "S" and "T" indicate one and two sample standard deviations from the mean, respectively.

The students situated at the upper end of the scale exhibited higher levels of data viz literacy, whereas those at the lower end exhibited lower levels. The items at the bottom of the scale were considered easy, meaning that most students could answer them correctly. In contrast, the items at the top of the scale were considered hard, which means that most students were unable to answer them correctly.

5. Discussion

5.1. Extracting Discriminative and Representative Items

Regarding the 3 MFRM model, the application to the original data showed disordered thresholds for higher categories (as reported in Section 4.1.2), and this suggests that the rating scale did not work properly. The original six-category rating scale, being more fine-grained than the four-category scale used in the analysis, assumes that raters are able to distinguish and consistently use the differences between categories, particularly at the extreme ends of the scale. The presence of threshold disordering between the fairly difficult and very difficult categories suggests that raters had difficulty using these two categories. A coarser rating scale, such as the four-category scale, is generally easier to apply than a six-category scale. Nevertheless, the advantages of a more fine-grained scale can be achieved only if raters are able to use it consistently, indicating that appropriate rater training may be required.

Looking at the separation reliability indices for examinees (i.e., data viz combinations), tasks, and raters, we found that

- The value of R for examinees is 0.68, indicating a moderate ability of the model to distinguish data viz combinations according to their discriminability, likely due to the specific type of examinees used in this analysis. For example, the stacked area chart yielded two examinees, A1 and A2, which only differ by one item (*Content*);
- The value of R for tasks is 0.96. This indicates that the tasks are highly distinct from each other;
- The value of R for raters is 0.84 indicating substantial differences in rater severity. This finding is confirmed by the rejection of the null hypothesis of equal severity of the raters, tested by a Wald-statistic (p-value < 0.001) [35], and by the presentation of raters' disagreement on each item, reported in Section 4.1.1.

Concerning the fit statistics, one of the two combinations involving the pie chart, namely P1, presented the highest values of Infit/Outfit MNSQ statistics, though still within the limits of real misfitting. Nevertheless, this finding suggests looking more carefully at P1. Its *PtMeaCorr* is -0.12, indicating a strong departure of the scores awarded by this data viz combination from the model expectation. P1 differs from P2, the other combination that still involves a pie chart, by item 35 (task *Content*). As shown in Figure 4 of Section 4.1.1, this item exhibited a high degree of disagreement among raters, mainly R5 and R6. These two raters appeared to be the ones who gave this item the unexpected score of 4, i.e., fairly or very difficult on the original scale, instead of 1.8, i.e., middle easy, as expected by the estimated 3 MFRM.

Moreover, taking into account the *PtMeaCorr* index, two other data viz combinations showed negative values: ST8, with -0.07, and ST6, with -0.06. These are two of the eight combinations involving the stacked bar chart. Two combinations, namely ST1 and ST3, had the higher *PtMeaCorr* indices instead (respectively 0.74 and 0.44), and what differentiated them from ST8 and ST6 was the *Represent* task (item 12 for ST1/ST3, and item 13 for ST8/ST6), as well as the *Content* task (item 17 for ST1/ST3, and item 18 for ST8/ST6). Looking at the heatmap in Figure 4, items 17 and 18 exhibited a similar degree of disagreement, while item 12 showed a slightly higher level of disagreement than item 13.

Inspection of the fit indices for the raters did not reveal any discrepancy between their observed and expected behaviour. The fit statistics for the tasks highlighted some criticality for task *Name*, which exhibited high values of Infit and Outfit MNSQ statistics (Infit MNSQ = 1.40; Outfit MNSQ = 1.41), although not high enough to indicate misfitting. This finding may suggest that, although the task *Name* is related to the data viz literacy construct, its connection appears weaker than those shown by the other three tasks.

In order to investigate whether the judgment process was fair and whether each rater kept a uniform level of severity among the examinees, we tested the significance of the examinee-by-rater interaction parameters ϕ_{nj} in Formula (4). Our analysis did not reveal any significant parameter, allowing us to conclude that each rater maintained a uniform level of severity across the data viz combinations.

Regarding the Wright Map, bubble chart (B_ combinations) emerged as the data viz with the highest discriminability effect, with both B1 and B2 receiving the highest scores on the difficulty scale. On the other hand, pie chart (P_ combinations) appeared as the data viz that discriminated the least, given that both P1 and P2 received lower scores on the difficulty scale. This result may not be surprising; the pie chart is commonly taught in primary education level curricula, thus the kind of information that can be read from it may be the most intuitive. It is reasonable that the raters judged its tasks very easy.

Most of the eight combinations involving choropleth map (G_) appeared in the lower part of the map, suggesting that this data viz followed the easiest one. Choropleth map is a popular data viz, fairly easy to interpret and widely used for visualising geospatial data across various media, including the news.

Two combinations of L, namely L3 and L4, had a higher than average discriminability effect (-0.19), while the other two combinations, namely L1 and L2, were below this average. This kind of result may be useful when assembling a new test.

Indeed, the inspection of the 3 MFRM model allowed us to infer that both data viz discriminability and representativeness may conflict; thus, DRIVE-T may help make a decision based on a justified trade-off between the two. For example, other combinations can be selected before L, and the best discriminating L combination can be added in the end. By applying DRIVE-T, the hierarchical progression of data viz and tasks representativeness is improved, unwanted interactions are avoided, and the resulting test may be more reliable and effective.

Regarding raters, more severe ones appeared in higher positions, and less severe ones in lower positions. For identifiability reasons, the average measure of raters' severity was constrained to be zero. Raters R2, R5, and R6 had a mean severity. R3 and R4 were the most severe raters, whereas R1 and R7 were the most lenient ones. The variability across raters was low, their measures showing a 0.68-logit spread. This may suggest that, even if there was no agreement among raters, as discussed previously, their level of severity was quite similar. From the visual inspection of the heatmaps in Section 4.1.1, it is observable that the task causing more disagreement among raters is the one related to knowing the *Name* of the data viz presented. The task related to retrieving the *Content* of a data viz is showing disagreements among raters, too. The difficulty of tasks related to what a data viz *Represent*(s) and how a data viz is *Use*(d) are mostly agreed by all the raters. These results reflected some patterns: *Name* and *Content* tasks frequently produced disagreements, whereas *Use* and, to some extent, *Represent* had higher consistency. Raters emerged as isolated contributors to disagreement, highlighting individual interpretive differences among them. However, those of the heatmaps were only providing an explorative analysis that further supports the 3 MFRM results.

Regarding tasks, *Name* emerged as the most challenging task, followed by *Content*, *Use*, and *Represent*. This hierarchy of measures may suggest the presence of a progression

level for the tasks being measured, and this hierarchy may help outline difficulty levels of an underlying construct for data viz literacy.

5.2. The Pilot Study

We note that, according to Figure 6 and the results reported in Section 4.1.2, the item combination selected among the ones modelled with the 3 MFRM, allowed the identification of a hierarchical progression expression of the data viz along the literacy continuum that was better reflected in the test designed to measure it.

The Rasch model used to analyse the pilot study data introduced in Section 3.3.1, yielded a person reliability index of 0.79, sufficiently high to ensure that the questionnaire was sensitive enough to distinguish between students with high and low levels of data viz literacy. The item reliability index was 0.95, sufficiently high to ensure that the sample of students was able to confirm the item difficulty hierarchy of the questionnaire.

The highest positive Pearson correlation value was 0.37 and involved items ST and SC, whereas the highest negative Pearson correlation value was −0.41 and involved the items L_Repr and TM. As the correlation values were relatively small, and no meaningful reasons justify connections among these items, it can be concluded that the items were approximately locally independent.

The results of the assessment of the unidimensionality assumption using PCA provided support for the interpretation that the test primarily measures students' level of data viz literacy.

Looking at the item fit statistics, only item P_Repr showed a high value for Infit (1.41) and Outfit (2.73) MNSQ. This means that this item had criticalities and deserved a rewriting or a substitution.

Regarding the analysis of the Write Map, the easiest item was the one that asked students to indicate the name of a pie chart. As also emerged from the 3 MFRM analysis, given that the pie chart is commonly taught in primary education level curricula, this finding was not surprising. Five out of the eight items asking for the name of the data viz were located to the very upper part of the item scale, indicating that the knowledge of the name of a data viz was, in most cases, a difficult task. This result is in line with the findings from the analysis of the raters' evaluations, where the *Name* task was judged to be the most difficult one.

The average level of data viz literacy observed in the sample of students was 0.51 (SD 1.1), which was greater than the average difficulty of the items (being it zero logits). This discrepancy may suggest that this group of students, on the whole, demonstrated an adequate level of data viz literacy.

5.3. Practical Adoption and Implementation of DRIVE-T

Although DRIVE-T integrates psychometric modeling techniques, its application does not necessarily require large-scale infrastructure or extensive respondent recruitment. The methodology is intentionally designed to be feasible at design time, relying on a limited panel of domain experts and a structured item bank. In its minimal configuration, DRIVE-T can be applied with a small number of raters, analogously to expert-based heuristic evaluation approaches commonly used in HCI, where three to five evaluators are often sufficient to identify critical usability issues [46].

From an implementation perspective, the main operational requirements of DRIVE-T include: (i) the availability of an artifact- and task-centered item bank tagged according to cognitive task dimensions (e.g., semiotics task dimensions); (ii) a short expert rating activity to obtain initial difficulty judgments; and (iii) access to Rasch-based software tools for modeling rater severity, task difficulty, and artifact discriminability. In this study, the

Many-Facet Rasch Measurement analysis was performed using the FACETS software [43], while the pilot study calibration was conducted with WINSTEPS [44]. However, equivalent Rasch modeling environments are increasingly available (e.g., in the Python *R* environment, version 4.5.2), and the methodological contribution of DRIVE-T remains independent of a specific software package.

Noticeably, DRIVE-T is modular and can be adopted at different levels of complexity depending on practitioners' goals and resources. Designers or educational practitioners may benefit from applying only Step 1 (task-based tagging) and Step 2 (expert difficulty rating) to qualitatively improve item design and ensure coverage of key construct traits. Step 3 (MFRM modeling) can then be incorporated in later validation phases when more rigorous psychometric calibration is required. This progressive adoption pathway lowers the application threshold for practitioners while preserving the possibility of full construct modeling and item selection grounded in discriminability and representativeness criteria. Overall, by supporting both lightweight expert-based screening and full Rasch-based modeling, DRIVE-T aims to enhance the generalizability and usability of construct-driven assessment design for HCI quality evaluation and literacy measurement contexts.

5.4. Limitations

Limitations of the work could be raised regarding the number of raters involved in the activity of scoring items. We recall that the main purpose of the paper was to introduce the DRIVE-T methodology, whose intent was to provide a quick and cheap way to qualitatively identify the more discriminative yet task-representative items among a set of ready-to-use or created-from-scratch items for each data viz. For this reason, the methodology should be proved robust enough to be applied with few raters, who were experts in the field. The activity of raters is analogous to what is done during heuristics evaluation approaches, where a panel of experts, usually three to five people [47], are asked to evaluate qualitatively the usability of application interfaces within their realm of expertise. Critiques may be moved to the number of items exploited in this study. Again, what was evaluated through the methodology was not the numerosity of items, but rather their expressivity in terms of cognitive traits related to the ability to master the data viz language. For this reason, the items should be qualitatively discriminating and representing all the traits of the measurement construct, as well as the majority of data viz difficulty levels. To reach this goal, it was not the quantity, but rather the quality of items that was prioritized. Another limitation may lay in the use of eight data viz, compared with the wider spectrum of available alternatives. As remarked in the introduction, the complexity of the measurement construct and of the item modelling required to focus on fewer, yet mostly used and qualitatively robust data viz. Relying on a set of them that is representative and common in data viz literacy assessment tests also guarantees future comparability with the other approaches. The occurrence of disordered thresholds and the moderate separation reliability for raters may suggest that the original six-point difficulty scale may have exceeded the discriminative capacity of expert judgments. In particular, adjacent categories such as fairly difficult and very difficult may not have been meaningfully distinguishable in practice. While collapsing categories improved model interpretability, future applications of DRIVE-T may benefit from coarser rating scales or from more explicit calibration examples provided to raters during briefing. Structured anchor items and short calibration exercises could further reduce rater variability without enforcing consensus, preserving the advantages of rater modeling within MFRM.

A final remark could be done on the use of a specific group of participants (high-school students) for the administration of the test with selected items from DRIVE-T. This should

be regarded as a proof of concept of the DRIVE-T methodology. Full validation of the latent construct would require a broader and more diverse population.

6. Conclusions

In this paper, we showed the potential and limitations of a methodology for the unprecedented rapid prototyping of items expressing latent traits of a data viz literacy measurement construct explicitly hypothesised but still blurred, and before administering items to respondents. DRIVE-T was able to better outline such construct (RQ1), to let emerge constructs from practice, and provide the best potentials of items, before administering the items to respondents (RQ2). Furthermore, modelling the raters' scoring with a 3 MFRM, we were able to select the best items (under the criteria of discriminability and representativeness) to create an assessment test (RQ3). This test was then administered to a sample in a pilot study. This pilot study confirmed the quality of the items being designed, rated, and selected, proving their discriminability and representativeness in measuring levels of data viz literacy. This work paved the way for an unprecedented method for measurement construct design, and the identification of its hierarchical and progressive levels, with a minimum effort (time and cost) towards high-quality tests. Future work will regard the administration to a broader population of 1000 participants at the same survey, which will provide a potential for a full validation of the approach. Constructs-foregrounding is possible when the most discriminative and representative items are selected, which may help the community of test designers scrutinize, discuss, and adjust items when they want to design or make sense of existing data visualization literacy assessment tests, whenever no data visualization literacy measurement construct was explicitly or completely outlined beforehand.

Author Contributions: Conceptualization, all; methodology, A.L. and S.G.; software, all; validation, all; formal analysis, all; investigation, all; resources, A.L.; data curation, S.G.; writing—original draft preparation, A.L. and S.G.; writing—review and editing, all; supervision, A.L.; project administration, A.L.; funding acquisition, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: Research funded by the European Union – Next-Generation, Mission 4 Component 1 CUP D53D23008690006, project name: "Characterizing and Measuring Visual Information Literacy" ID 2022JJ3PA5, CUP D53D23008690006.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Universitá degli Studi di Brescia, protocol code 36/2025, date of approval 15 September 2025.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All data are available on request. All items are made available as a link in the paper text to an external repository.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Barricelli, B.R.; Fogli, D.; Locoro, A. EUDability: A new construct at the intersection of end-user development and computational thinking. *J. Syst. Softw.* **2023**, *195*, 111516. [CrossRef]
2. Islam, M.N.; Khan, N.I.; Inan, T.T.; Sarker, I.H. Designing user interfaces for illiterate and semi-literate users: A systematic review and future research agenda. *Sage Open* **2023**, *13*, 21582440231172741. [CrossRef]
3. Ge, L.W.; Cabouat, A.F.; Bonilla, K.; Cui, Y.; Ding, Y.; Rakotondravony, N.; Creamer, M.M.; Otto, J.T.; Hedayati, M.; Kwon, B.C.; et al. An Autoethnography on Visualization Literacy: A Wicked Measurement Problem . *IEEE Trans. Vis. Comput. Graph.* **2025**, 1–11. [CrossRef] [PubMed]

4. Burns, A.; Xiong, C.; Franconeri, S.; Cairo, A.; Mahyar, N. How to evaluate data visualizations across different levels of understanding. *arXiv* **2020**, arXiv:2009.01747. [CrossRef]
5. Boy, J.; Rensink, R.A.; Bertini, E.; Fekete, J.D. A principled way of assessing visualization literacy. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1963–1972. [CrossRef]
6. Beschi, S.; Falessi, D.; Golia, S.; Locoro, A. Characterizing Data Visualization Literacy for Standardization: A Systematic Literature Review. *IEEE Access* **2025**, *13*, 65704–65725. [CrossRef]
7. Barricelli, B.R.; Fogli, D.; Gargioni, L.; Locoro, A.; Valtolina, S. Towards the Unification of Computational Thinking and EUDability: Two Cases from Healthcare. In Proceedings of the 2024 International Conference on Advanced Visual Interfaces, Arenzano, Italy, 3–7 June 2024; pp. 1–9.
8. Mari, L.; Wilson, M.; Maul, A. *Measurement Across the Sciences: Developing a Shared Concept System for Measurement*; Springer Nature: Berlin/Heidelberg, Germany, 2023.
9. Wilson, M. *Constructing Measures: An Item Response Modeling Approach*; Routledge: London, UK, 2004.
10. Cabitza, F.; Locoro, A. "Made with Knowledge"—Disentangling the IT Knowledge Artifact by a Qualitative Literature Review. In Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Rome, Italy, 21–24 October 2014; pp. 64–75. Available online: https://api.semanticscholar.org/CorpusID:28985187 (accessed on 10 January 2026).
11. Eco, U. *A Theory of Semiotics*; Indiana University Press: Bloomington, IN, USA, 1979. Available online: https://books.google.it/books?id=BoXO4ItsuaMC (accessed on 20 December 2025).
12. De Souza, C.S. *The Semiotic Engineering of Human-Computer Interaction*; MIT Press: Cambridge, MA, USA, 2005.
13. Stagg, B.C.; Schlechter, C.R.; Del Fiol, G. The Digital Divide in Eye Health Care Delivery. *JAMA Ophthalmol.* **2024**, *142*, 452–453. [CrossRef]
14. Star, S.L.; Griesemer, J.R. Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Soc. Stud. Sci.* **1989**, *19*, 387–420. [CrossRef]
15. Otto, J.T.; Davidoff, S. Visualization Artifacts are Boundary Objects. In Proceedings of the 2024 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV), St. Pete Beach, FL, USA, 14 October 2024; pp. 81–88.
16. Tsai, M.-J.; Liang, J.-C.; Lee, S.W.-Y.; Hsu, C.-Y. Structural Validation for the Developmental Model of Computational Thinking. *J. Educ. Comput. Res.* **2021**, *60*, 56–73. [CrossRef]
17. Schneider, W.J.; McGrew, K.S. *The Cattell-Horn-Carroll Model of Intelligence*, 3rd ed.; Guilford Press: New York, NY, USA, 2012.
18. Aoyama, K.; Stephens, M. Graph interpretation aspects of statistical literacy: A Japanese perspective. *Math. Educ. Res. J.* **2003**, *15*, 207–225. [CrossRef]
19. Galesic, M.; Garcia-Retamero, R. Graph Literacy: A Cross-Cultural Comparison. *Med. Decis. Mak.* **2011**, *31*, 444–457. [CrossRef] [PubMed]
20. Merbitz, C.; Morris, J.; Grip, J.C. Ordinal scales and foundations of misinference. *Arch. Phys. Med. Rehabil.* **1989**, *70*, 308–312.
21. Lee, S.; Kim, S.H.; Kwon, B.C. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 551 – 560. [CrossRef] [PubMed]
22. Krejci, S.E.; Ramroop-Butts, S.; Torres, H.N.; Isokpehi, R.D. Visual literacy intervention for improving undergraduate student critical thinking of global sustainability issues. *Sustainability* **2020**, *12*, 10209. [CrossRef]
23. Locoro, A.; Fisher, W.P.; Mari, L. Visual information literacy: Definition, construct modeling and assessment. *IEEE Access* **2021**, *9*, 71053–71071. [CrossRef]
24. Yang, L.; Xiong, C.; Wong, J.K.; Wu, A.; Qu, H. Explaining with examples: Lessons learned from crowdsourced introductory description of information visualizations. *IEEE Trans. Vis. Comput. Graph.* **2021**, *29*, 1638–1650. [CrossRef]
25. Camba, J.D.; Company, P.; Byrd, V.L. Identifying deception as a critical component of visualization literacy. *IEEE Comput. Graph. Appl.* **2022**, *42*, 116–122. [CrossRef]
26. Firat, E.E.; Denisova, A.; Wilson, M.L.; Laramee, R.S. P-lite: A study of parallel coordinate plot literacy. *Vis. Inform.* **2022**, *6*, 81–99. [CrossRef]
27. Ge, L.W.; Cui, Y.; Kay, M. CALVI: Critical Thinking Assessment for Literacy in Visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI'23), Hamburg, Germany, 23–28 April 2023*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–18. [CrossRef]
28. Davis, R.; Pu, X.; Ding, Y.; Hall, B.D.; Bonilla, K.; Feng, M.; Kay, M.; Harrison, L. The risks of ranking: Revisiting graphical perception to model individual differences in visualization performance. *IEEE Trans. Vis. Comput. Graph.* **2024**, *30*, 1756–1771. [CrossRef]
29. Cui, Y.; Ge, L.W.; Ding, Y.; Harrison, L.; Yang, F.; Kay, M. Promises and Pitfalls: Using Large Language Models to Generate Visualization Items. *IEEE Trans. Vis. Comput. Graph.* **2024**, *31*, 1094–1104. [CrossRef]
30. Cabitza, F.; Locoro, A.; Batini, C. Making open data more personal through a social value perspective: A methodological approach. *Inf. Syst. Front.* **2020**, *22*, 131–148. [CrossRef]

31. Pandey, S.; Ottley, A. Mini-VLAT: A Short and Effective Measure of Visualization Literacy. *Comput. Graph. Forum* **2023**, *42*, 1–11. [CrossRef]
32. Linacre, J.M. *Many-Facet Rasch Measurement*; MESA Press: Chicago, IL, USA, 1989.
33. Eckes, T. Many-facet Rasch measurement. In *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning Teaching, Assessment*; Council of Europe/Language Policy Division: Strasbourg, France, 2009; p. Section H.
34. Myford, C.; Wolfe, E.W. Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *J. Appl. Meas.* **2003**, *4*, 386–422. [PubMed]
35. Eckes, T. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*, 2nd rev. and updated ed.; Peter Lang Edition: Lausanne, Switzerland, 2015.
36. Marasini, D.; Quatto, P.; Ripamonti, E. Assessing the inter-rater agreement for ordinal data through weighted indexes. *Stat. Methods Med. Res.* **2016**, *25*, 2611–2633. [CrossRef]
37. Andrich, D. *Rasch Models for Measurement*; Sage Publications: Thousand Oaks, CA, USA, 1988.
38. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; University of Chicago Press: Chicago, IL, USA, 1960.
39. Andrich, D. A rating formulation for ordered response categories. *Psychometrika* **1978**, *43*, 561–573. [CrossRef]
40. Masters, G.N. A rasch model for partial credit scoring. *Psychometrika* **1982**, *47*, 149–174. [CrossRef]
41. Linacre, J.M. What do infit and outfit, mean-square and standardized means? *Rasch Meas. Trans.* **2002**, *16*, 878.
42. Linacre, J.M.; Wright, B.D. Construction of measures from many-facet data. *J. Appl. Meas.* **2002**, *3*, 484–509.
43. Linacre, J.M. Facets Computer Program for Many-Facet Rasch Measurement. 2025. Available online: https://www.winsteps.com/facets.htm (accessed on 31 January 2026).
44. Linacre, J.M. Winsteps® Rasch Measurement Computer Program (Version 5.4.0). 2023. Available online: https://www.winsteps.com (accessed on 31 Januart 2026).
45. Bond, T.G.; Fox, C.M. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3rd ed.; Routledge: New York, NY, USA, 2015.
46. Nielsen, J. Finding usability problems through heuristic evaluation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Monterey, CA, USA, 3–7 May 1992; pp. 373–380.
47. Nielsen, J. *The Theory Behind Heuristic Evaluations*; Nielsen Normal Group: Dover, DE, USA, 1994.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.