EasyChair Preprint

№ 4425

An Empirical Methodological Study of Evaluation Methods Applied to Educational Timetabling Visualizations

Wanderley de Souza Alencar, Walid Abdala Rfaei Jradi, Hugo Alexandre Dantas Do Nascimento, Juliana Paula Félix and Fabrizzio Alphonsus Alves de Melo Nunes Soares

October 19, 2020

An Empirical Methodological Study of Evaluation Methods Applied to Educational Timetabling Visualizations

Wanderley de Souza Alencar[1 [0000−0002−3785−9527]],
Walid Abdala Rfaei Jradi[1 [0000−0003−3251−3604]],
Hugo Alexandre Dantas do Nascimento[1 [0000−0003−1690−1201]],
Juliana Paula Felix[1 [0000−0003−4095−1639]], and
Fabrízzio Alphonsus Alves de Melo Nunes Soares[2,1 [0000−0003−1598−1377]]

[1] Instituto de Informática, Universidade Federal de Goiás, Goiânia/GO, Brazil,
http://www.inf.ufg.br
wanderleyalencar@ufg.br, walid.jradi@gmail.com,
hadn@inf.ufg.br, jufelix16@gmail.com
[2] Department of Computer Science, Southern Oregon University, Ashland/OR, USA,
http://www.sou.edu
soaresf@sou.edu

Abstract. The conception, and usage, of methods designed to evaluate information visualizations is a challenge that goes along with the development of these visualizations. In the scientific literature there is a myriad of proposals for such methods. However, none of them was able to pacify the field or establish itself as a *de facto* standard, due to difficulties like: (a) the complexity of its usage; (b) high financial and time costs; and (c) the need of a large number of raters to guarantee the reliability of the results. One way to circumvent such adversities is the usage of *Heuristic Evaluation* given its simplicity, low cost of application and the quality of reached results. This article intends to conduct an *empirical methodological study* about the use of three of such methods (Zuk *et al.*, Forsell & Johansson and *Wall et al.*) for evaluation of visualizations in the context of Educational Timetabling Problems. Five different visualizations were evaluated using the original methods and versions modified by the current authors (where an *importance factor* was assigned to each statement being evaluated, as well as the rater's *level of confidence*) in order to improve their efficiency when measuring the quality of visualizations. The experimental results demonstrated that for the two first heuristics, only the modification on the *importance* of the statements proved to be (statistically) relevant. For the third one, both factors did not induce different results.

Keywords: Information Visualization Evaluation · Heuristic Evaluation · Educational Timetabling Problems · Combinatorial Problems · NP-Hard · User Interface

1 Introduction

It is well known, and accepted, by the Information Visualization (IV) scientific community that assessing the *quality* and/or *usefulness* of a visualization is not a trivial task, as showed by several works [23,20,14,4]. Some of the difficulties involved includes determining whether a given visualization is able to: (1) positively contribute to the understanding and analysis of the data [20,14]; (2) induce significant *insights*[3]; and (3) be expressive enough, easy to *memorize* and *aesthetically* appropriate [10]

Considering the evaluation of interactive visualizations, Xiaozhou Zhou *et al.* [24] list three main difficulties: (1) the semantics meaning significantly differs between distinct types of databases and to build an uniform evaluation system is not trivial; (2) the subjective influence of cognitive process of each individual is remarkable. This includes person's knowledge structure and the familiarity on the field and technology. In the course of cognition, even the psychological and physiological state of an individual has a significant influence on cognition performance; (3) the number of visual elements in the interactive visualization is excessive and a high linkage relationship is present, which leads to an increase in the visual complexity. Therefore, the creation and evaluation of a visualization interface that comprises a separate set of visual elements is impossible.

There are currently two major lines, or general approaches, when evaluating a visualization: (1) **Generalist**: applies assessment instruments that are *independent* of the problem domain, evaluating sufficiently generic and high-level characteristics of the visualization. Some studies [8,5] are examples; (2) **Specific Problem Domain**: as the name says, this kind of evaluation is tightly coupled to the problem's characteristics, not being applicable in other contexts [12].

Following the second line, a common approach is the so called *task-based*. It consists of recording the accuracy and the time consumed to carry out a well-defined list of *tasks*, conducted by a carefully selected group of users, as shown by [18,2]. This approach aims to answer the following general question: "*Are users able to use the visualization for the understanding of the underlying data/information, and correctly and efficiently perform the proposed tasks?*".

Based on the critical analysis of several contemporary methodologies for evaluating visualizations, the current paper performs an *empirical methodological study* of evaluation methods when applied to *Educational Timetabling Problems* (Ed-TTPs) [15]. Such problems arrive from the need of conceiving and presenting timetables for educational institutions (schools, universities, etc.). Hamed Babaei *et al.* [3] point out that this *big* class of problems can be divided in: 1. High-School Timetabling Problem (HS-TTP); 2. University Timetabling Problems (U-TTP), including Course (UC-TTP) and Examination (UE-TTP).

The remainder of this paper is organized as follows: Section 2 presents the main works in information visualization evaluation field and provides a critical

[3] Chris North [17] clarifies that the ability to measure whether a given visualization can induce insights, or not, is subjective and individual. Also, it is known that insights are characterized by being *complex*, *deep*, *qualitative*, *unexpected* and *relevant*. In another words, an insight can not simply be directly extracted from the visualization.

analysis on them. Section 3 details the empirical methodological study aimed to compare three different evaluation methods when applied in the context of Ed-TTPs. Section 4 reports the experimentation process and records the conclusions from the evaluation of three visualizations extracted from the scientific literature of the area. Finally, Section 6 synthesizes the conclusions obtained by this work and points out to potential future research.

2 Literature Review

This section presents a brief review of the literature on scientific works related to the topic under analysis, with an emphasis on publications from the last twenty years. Even with this restriction, the number of studies is vast and, therefore, only those with high affinity with the approach here proposed are discussed.

In 2012, Heidi Lam *et al.* [14] examined 850 selected publications from 1995–2010 and identified seven scenarios that defined the practical application of information visualization evaluations (IV-Eval) in that period. The most frequent types of evaluation are aimed at measuring people's task performance (also known as *task-based* evaluation), user experience and quality/performance of algorithms. Expanding this previous study, Tobias Isenberg *et al.* [13] reviewed 581 papers to analyze the practice in the context of evaluation visualizations (data and information). They concluded that, in general, the level of evaluation reporting is low. Some found pitfalls were: (1) the goals of the evaluation are not explicit; (2) participants do not belong to the target audience; (3) the strategy and the method of analysis are not appropriate; (4) the level of rigor is low. The authors of the present study observe that, until the current days, these pitfalls are still noticed in the reports of IV evaluation.

Steven R. Gomez *et al.* [9] devised an evaluation method that combines *insight-based* and *task-based* methodologies, called LITE (*Layered Insight- and Task-based Evaluation*). Experimentation was carried out on a *visual analytics* system that required the user to carry out both research on the data set as well as the analysis of this data in order to identify broader patterns. As *general guidelines* for the visualization designers, the authors suggested that: (1) low-level tasks must be chosen in order to not steer participants toward insights; (2) ordering effects must be mitigated by counterbalancing the ordering of visualizations in the *insight* component of LITE; (3) it must be considered the complexity of the data and participant expertise when choosing insight characteristics to measure; (4) the details of the process of coding insights have to be reported: who are the coders, how well did they agree, and how were disagreements resolved into one score.

Right after these studies and going beyond the *task-based* approach, John Stasko [20] conceived the idea that a visualization should be evaluated by its *value*, a metric to measure the broader benefits that a visualization can generate. He says that a "*Visualization should ideally provide broader, more holistic benefits to a person about a data set, giving a* ***big picture*** *understanding of the data and spurring insights beyond specific data case values.*" (emphasis added).

Thus, the *value* goes beyond visualizations' ability to answer questions about the data. It is at the heart of visualizations' the aptitude to allow a true understanding of the data, the creation of a holistic scope and an innate sense of context, evidencing the importance of data in forming a general overview. Stasko defines that the *value* (V) of a visualization is expressed by the Equation (1):

$$V = I + C + E + T \tag{1}$$

where, I (*insights*) measures the discovery of *insights* or *insightful questions* about the data, C (*confidence*) represents the level of conviction, trust and knowledge about the data, its domain and the context, E (*essence*) conveys the general essence or perception of the data, and T (*time*) indicates the total time required to answer a wide variety of questions about the data.

The aforementioned study uses a qualitative analysis bias, as there is no specification about a measurement method (a quantitative approach) related to the previous four components. Based on it, Emilly Wall *et al.* [23] propose a *Heuristic Evaluation* (HE) [16] – a method in which experts employ experimental-based rules to assess the usability of user interfaces in independent steps and report issues – which goal is to evaluate interactive visualizations, seeking a methodology that: (1) is low cost in terms of time and resources required for its usage; (2) allows measuring the usefulness of the visualization in addition to that provided by a *task-based* approach; (3) is practical and relatively easy to use; and (4) admits comparison between different visualization applications.

To make the proposed methodology (called ICE-T) viable, the components were decomposed into *guidelines* and these were defined by low-level heuristics expressed by a set of statements, that follows the Likert scale, from *Strongly Disagree* to *Strongly Agree*. After experimentation involving fifteen researchers and three visualizations, the authors concluded that the methodology was promising for the evaluation of interactive visualizations, highlighting that they obtained results consistent with a qualitative assessment carried out previously. They also call attention to the fact that only five raters would be enough to obtain the same results, demonstrating its applicability to real-world problems.

The ability of experts in an HE to identify usability problems in a visualization application – when compared to using a group of non-experts in the same task – was subject of a study by Beatriz Santos *et al.* [19]. They concluded that the use of experts employing HE, such as those proposed by Camilla Forsell *et al.* [6,7], Torre Zuk *et al.* [25] and Jakob Nielsen [16], was able to identify most of the problems later reported by a larger amount of non-experts, which subsidized the choice made by [23]. Following another path, M. Tory & T. Möller [22] argue that *expert feedback* can be a complement to HE and a mechanism that helps understand high-level cognitive tasks.

Also focusing on (dynamic) interactive visualizations (DIV) interfaces, Xiauzhou *et al.* [24] present a new quantitative method based on eye-tracking (visual momentum – VM). The authors argue that the performed experiments proved there are a positive effect on reducing cognitive load in DIVs and, therefore, VM can be used, with reliability and convenience to evaluate a visual interface. M. A. Hearst *et al.* [11] argue that there is a mutual influence between

conducting assessments using HE and *Query-Based Scoring* and that the usage of both can contribute to improve the quality of the visualization evaluation process.

2.1 Critical Analysis of Previous Methodologies

The work of Emilly Wall *et al.* [23], based on Stasko's seminal idea [20], signals strength for applicability in other real-world contexts, considering that it requires a small number of evaluators to obtain results equivalent to those obtained by the application of a qualitative evaluation method. It is characterized, therefore, as a low cost application methodology. However, as highlighted by the authors of the article, the methodology does not serve as a panacea, and should be used as a complementary approach to the application of other assessment techniques. Another drawback is that its validation was done using only three visualizations, thus lacking a large scale experimentation that could involve other problem domains in order to reinforce its validity.

The work of Steven R. Gomez *et al.* [9] is relevant for identifying and presenting four general guidelines applicable to visualization assessments, as well as for emphasizing the usefulness of combining *insight* and *task-based* evaluation approaches. Just like the previous work, it also needs a broader application, with more case studies and different designs.

Despite its highlighted qualities for the evaluation of (dynamic) interactive visualizations, the approach proposed by M. A. Hearst *et al.* [11] has, as a drawback, the (current) high cost of the equipment involved in the experiments, ultimately making extremely difficult its adoption in large-scale in real-world scenarios.

The evidence presented in the studies of Beatriz Santos *et al.* [19] and M. Tory & T. Möller [22] reinforces the idea that the proper use of strategies based on heuristic evaluation, subsidized by the guidelines of Steven R. Gomez *et al.* [9], induces and allows the conduction of empirical studies to compare different visualization assessment methods. The present work compares three heuristic-based methods available in the literature to measure their effectiveness to estimate the quality of visualizations when applied to *Educational Timetabling Problems*. The next section details how the empirical study was coined.

3 Proposed Empirical Methodological Study

Inspired by the research listed at Section 2.1, the present paper proposes and performs an empirical study to compare three heuristic evaluation methods: (1) Zuk *et al.* [25], with thirteen statements to measure the quality of the visualization; (2) Forsell & Johansson [7], which uses ten statements; and (3) Wall *et al.* [23] employ twenty one statements.

The three methods were chosen because they are well known and, despite sharing some common concepts, investigate complementary aspects for evaluating information visualization applications. The study is structured in three phases, as shown in Fig. 1.

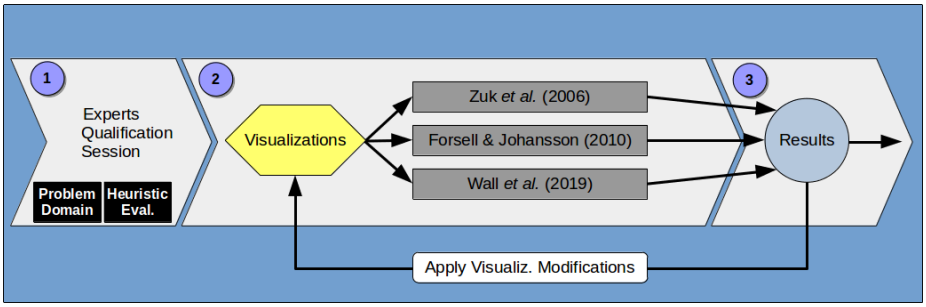


Fig. 1. Proposed empirical methodological study structure.

1. **Expert Qualification Session**: In order to uniform the knowledge of the evaluating specialists (raters), in this phase a session is held in which are provided the main theoretical-practical foundations about *Heuristic Evaluation* (HE) and the target problem domain *Educational Timetabling Problems* (Ed-TTPs). In addition, its opened space for the exchange of experiences between the participants. The authors of this article conduct the meeting and make notes to plan, and improve, future experiments;
2. **Individual Heuristic Evaluation Session**: Using as keystone the three HEs revisited in the previous phase, the specialists assign a score for each statement, based in a natural numeric scale. The scores can vary from 1 (minimum) to 5 (five), being possible to indicate that the statement is *Not Applicable* (N/A) and, therefore, no score is attributed to it. This was done to make the evaluation simpler, more natural, homogeneous and, later, to make feasible better statistical treatment;
3. **Results and Analysis**: Finally, after compiling the assigned scores, the final results of each visualization are generated. This is followed by a statistical analysis aimed at identifying the level of confidence in the result.

It is important to note that in Phase 2, in the original heuristics, all statements were considered to be of equal importance. However, this approach make it difficult for factors resulting from the particularities of the problem domain to be observed during the visualization evaluation process. Thus, to address this issue, the authors of this article (specialists in educational timetabling problems), in a phase prior to those mentioned, defined the relative importance (weight) of each statement of all heuristics used when applied in the context of Ed-TTPs. In the experiments, the weights can vary in the range 0 (minimum) to 6 (maximum). They were employed in Scenarios 3 and 4, as depicted by Table 1. These scenarios are described at the beginning of Section 5.

The current authors also define that for all statements in the Phase 2, each rater must record his/her confidence in the score, expressed by one value in the set $\mathcal{S}_C = \{1, 2, 4\}$, where 1 means *small*, 2 *medium* and 4 *total* confidence.

Table 1. Statement weights per heuristic adopted in Scenarios 3 and 4.

Heuristic Method	Statement Number																				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Zuk *et al.*	4	3	3	3	3	3	3	6	6	4	4	6	4	-	-	-	-	-	-	-	-
Forsell & Johansson	6	2	2	2	6	6	6	2	4	4	-	-	-	-	-	-	-	-	-	-	-
Wall *et al.*	4	2	2	2	2	2	2	2	4	4	4	4	4	6	6	4	4	6	6	4	4

Rater's confidence is employed to weight his/her evaluation using the factors $\frac{1}{4}$, $\frac{1}{2}$ and 1, respectively, to multiply the rater's score.

Typically the phases 1–3 are applied in this sequence and only once each, since it is considered in this scenario that the visualization being evaluated is in use and cannot be changed, that is, it is a process of evaluation carried out after the definitive visualization implementation.

However, according to the purpose of the assessment, phases 2 and 3 can be applied repetitively, during the process of developing/evaluating a new visualization. In this case, after each application cycle, the visualization can be improved to eliminate the identified problems and implement suggestions provided by the raters. At the end of this process the results are considered the definitive ones.

The next section details how the empirical study was carried out according to these guidelines.

4 Experiment Description

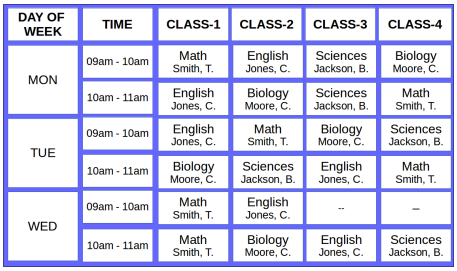
This section describes the application of the empirical methodological study presented in Section 3, aiming to check if the approach:

1. identifies differences between the three heuristic-based methods, regarding the measure of the quality of the evaluated visualizations;
2. whether the modifications applied by the authors of this article benefited the original heuristics, improving their efficiency when measuring the quality of visualizations in *Educational Timetabling Problems* (Ed-TTPs) context;
3. is capable of assisting decision-making by professionals involved in the problem domain.

Initially, an explanatory text was sent to fifteen specialists, all them involved in Information Visualization (IV) research projects, explaining the motivations for the carried out study and inviting them to take part in the planned experimentation. Ten accepted the invitation: seven men and three women, aged between 30 and 52 years. Of these, five participants are professors in higher education, three are PhD candidates and two have a master degree in Computer Science. Among the participants, four are also experts in the problem domain (Ed-TTPs) and the others have good knowledge in this field.

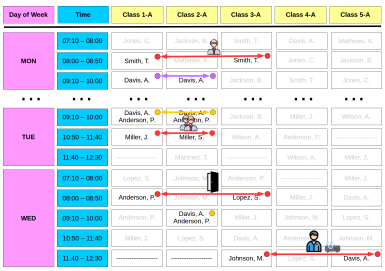
The visualizations used in the experiments were obtained from the following sources in the literature:

VisT2D – *Traditional 2D-Table*: Shown in Fig. 2(a), this type of visualization is widely used in the area of Ed-TTPs. It displays a column to represent the days of the week, another for the possible event times and one column per group (a class) of involved students.

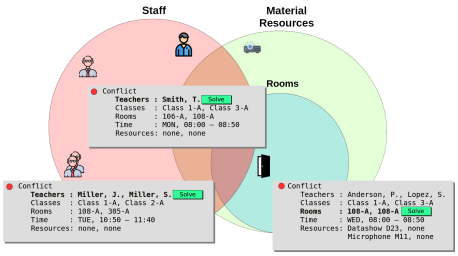


DAY OF WEEK	TIME	CLASS-1	CLASS-2	CLASS-3	CLASS-4
MON	09am - 10am	Math Smith, T.	English Jones, C.	Sciences Jackson, B.	Biology Moore, C.
	10am - 11am	English Jones, C.	Biology Moore, C.	Sciences Jackson, B.	Math Smith, T.
TUE	09am - 10am	English Jones, C.	Math Smith, T.	Biology Moore, C.	Sciences Jackson, B.
	10am - 11am	Biology Moore, C.	Sciences Jackson, B.	English Jones, C.	Math Smith, T.
WED	09am - 10am	Math Smith, T.	English Jones, C.	--	–
	10am - 11am	Math Smith, T.	Biology Moore, C.	English Jones, C.	Sciences Jackson, B.

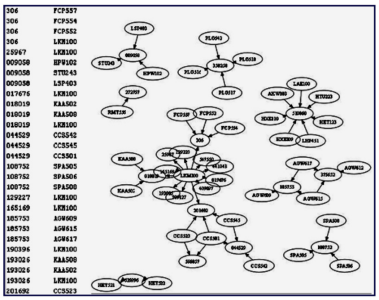
(a) VisT2D.



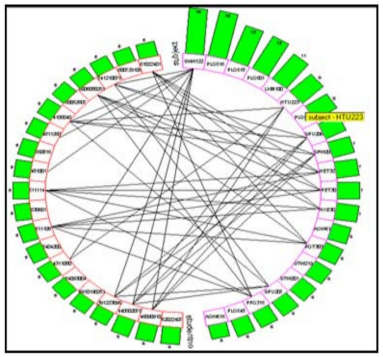
(b) VisETV [1].



(c) VisMDV [1].



(d) VisG [21].



(e) VisDC [21].

Fig. 2. Visualization of Educational Timetabling used in this study.

VisETV – *Enhanced Tabular Visualization* (ETV): Extracted from [1] and depicted in Fig. 2(b). It is based on the VisT2D approach, but extends it to highlight the elements of the problem involved in the *conflicts* (or *clashes*). A *cell* contains information about one lecture. If it's in light gray, this indicates the ab-

sence of any conflict. Two cells can be connected through a colored pair of points and a bidirectional arc, pointing to a *real*/*apparent* conflict between them. The color of the points/arc makes the distinction. A *teacher*, a *door*, a *technical staff member* and a *datashow* represent, respectively, a conflict in: (1) the teacher's timetable; (2) a room, being used simultaneously by two or more classes; (3) a technical staff member's timetable; and (4) a material resource.

VisMDV – *Multilayer Diagram Visualization*: Also presented in [1] and illustrated in Fig. 2(c), it resembles the *Venn Diagram*, although it is distinct because it considers that the sets involved are in *different layers* in the visualization, since each one is associated to one of the categories of resources involved in a conflict (staff and material resources). As a result, the intersection of sets is reinterpreted: instead of indicating elements in common to two sets, it means that a conflict involves elements associated with those two categories. The rooms are considered a special subcategory of the material resources that deserve differentiated treatment and, therefore, appear as a separated subset. The user, clicking a conflict icon, opens an correspondent pop-up window, which is an artistic representation that, in order to avoid visual pollution and due to space constraints, omits the *technical staff member* and *material resource* conflict pop-up windows.

VisCG – *Cluster Graph View*: Registered in [21] and presented in Fig. 2(d). It shows a cluster graph whose *nodes* represent students or exams of some subject. An *arc* connecting a subject to a student indicates that he(she) must take that exam, showing possible conflicts.

VisDC – *Daisy Chart View*: Also registered in [21] and presented in the Fig. 2(e). In it, the boxes represent students or subjects and the *edges* denote the relationship that the student must take the exam of that subject. The histograms show the number of existing associations for each element (student or subject) of the semicircle on which the user is *focusing* at that moment.

5 Experimental Results

Due to COVID-19 pandemic, Phase 1 was performed in a remote session. In Phase 2, as aforementioned, the evaluations were conducted providing a questionnaire in a numeric scale format, from 1 to 5, and 0 when *Not Applicable*, and answers collected through an individually shared spreadsheet. The raters were not required to provide answers in a specific order. Although we estimated five hours to perform evaluation process, we provided one week to all raters.

In Phase 3, four different scenarios were used to confront the evaluations: (*Scenario 1*) Ignore statement's weights and rater's self-declaration confidence; (*Scenario 2*) Ignore statement's weights and take in account rater's confidence; (*Scenario 3*) Take in account statement's weight and ignore rater's confidence; (*Scenario 4*) Take in account statement's weight and rater's confidence.

The Fig. 3 presents our results for all three heuristics, in scenarios 1 to 4, in which we compare all five visualizations. Although our result is an average result of all raters, since each heuristic has a different amount of statements to evaluate, we normalized our results to values between 0 (zero) and 10 (ten) to

allow a comparison of results in the same scale. Equations 2 and 3 present how scores are computed.

$$Score_r = \frac{\sum_{s=1}^{|S|} (weight_s \times confidence_s \times score_s)}{\mid S \mid} \times 10, \tag{2}$$

$$Score_v = \frac{\sum_{r=1}^{|R|} Score_r}{\mid R \mid}, \tag{3}$$

where $Score_r$, $Score_s$ and $Score_v$ are raters (r), statements (s) and visualizations (v) scores, respectively. The weights of the statements and raters confidence score are $weight_s$ and $confidence_s$, in this order. The values $\mid R \mid$ and $\mid S \mid$ are the number of raters and heuristic statements.

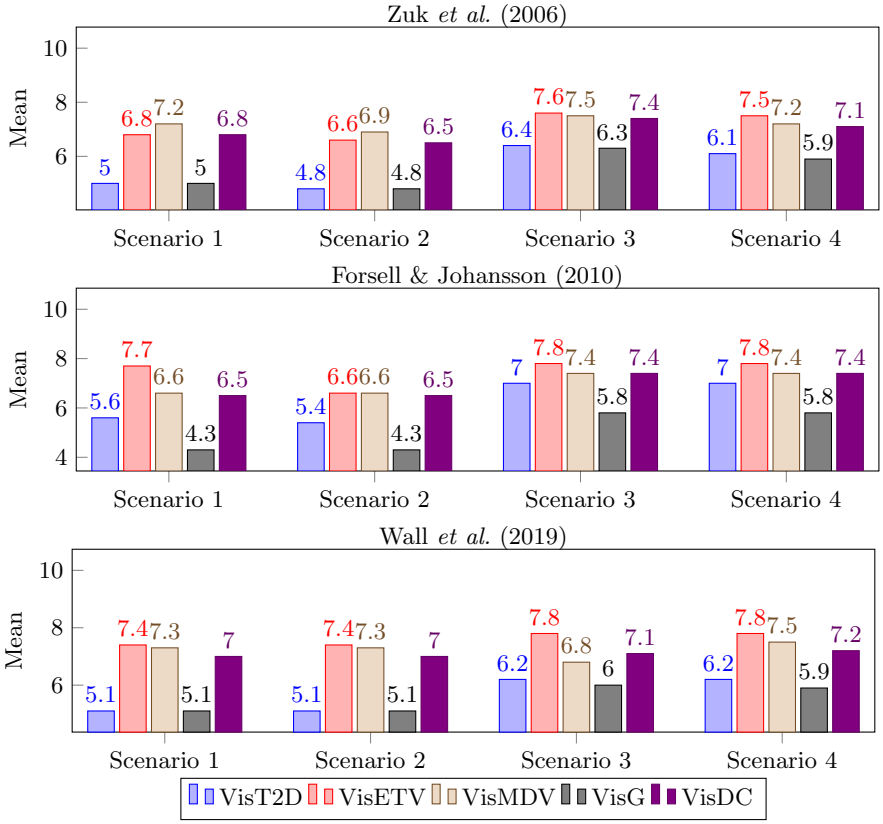


Fig. 3. Comparison of the evaluation of five visualizations in four different scenarios and three heuristics (Zuk *et al.*, Forsell & Johansson *et al.* e Wall *et al.*).

In the application of Zuk *et al.* [25], it was possible to conclude that, regardless the visualization considered, the insertion of the raters' confidence did not

generate statistical difference in the results of the evaluations. Nevertheless, the insertion of the statement weights in the problem domain generated statistical difference and, therefore, was relevant in the final results of the evaluations.

When analyzing the application of Forsell & Johansson [7], despite the slight numerical differences regarding the previous heuristic, the same conclusion is obtained: the raters' confidence did not generate statistical difference, but the insertion of the statements weights performed by the authors did.

Applying Wall *et al.* [23], it was possible to infer that, in any scenario, neither the raters' confidence nor the weighting of the statements were able to induce statistical difference in the results evaluations. Therefore, in the scope of this study, Scenario 4 proved to be unnecessary. Another unique fact is that, in any scenario, the VisETV visualization is the one with the best evaluation, as well as the VisMDV and VisDC occupied the second or third places.

From the comparison of these facts, it was possible to realize that for the first two heuristics, only the weights of the statements proved to be statistically relevant. This was expected by the present authors, since invited raters are researchers in IV, which contributed to a high confidence in their answers.

In order to support our conclusions, we conducted an hypothesis test, in which heuristic methods and scenarios can be considered independent variables, while raters score is the dependent variable. However, heuristic and scenarios were divided in experiments to avoid a multivariate analysis. Initially we applied a *normality test* to each experiment group (heuristic) via Lilliefors test and Shapiro-Wilk tests and both reported that our results follow a normal distribution. Therefore we conducted an ANOVA (Analysis of Variance) to verify statistical difference between scenarios considering 0.05 of significance. Table 2 presents ANOVA results for each comparison.

Table 2. Hypothesis Test results (ANOVA).

Heuristic Method	**Scenario 1 vs 2**		**Scenario 1 vs 3**		**Scenario 1 vs 4**	
	F-Stat	P-Value	F-Stat	P-Value	F-Stat	P-Value
Zuk *et al.* (2006)	0.39	0.5331	8.93	**0.0036**	5.09	**0.0265**
Forsel & Johanssen (2010)	0.32	0.5734	8.29	**0.0050**	8.22	**0.0052**
Wall *et al.* (2019)	0.01	0.9417	1.33	0.2513	3.04	0.0849

As we can see in Table 2, Scenario 1 vs. 2 has show no statistical difference for all heuristic methods. This show that the level of confidence that a rater has in each question did not provide any improvement in our results. This was also expected, since we invited a group of researchers in IV to evaluate.

Scenario 1 vs. 3 show statistical difference for Zuk and Forsel heuristics, meaning that the weights assigned by rater performs a significant change in the result and can be helpful when selecting a visualization for timetabling. And although scenarios 1 vs. 4 also show statistical difference for Zuk and Forsell, this is a consequence of Scenario 3, since Scenario 4 is a combination of Scenarios 2 and 3. Finally, no statistical difference was presented for scenarios 1 vs. 3

and 1 vs. 4. In this sense the timetabling specialist weights did not provide any improvement to the evaluation process. Wall *et al.* has about a double of statements to evaluate, which probably buffered the effect of weights. In this sense, this heuristic is more robust and require a strong variation of the specialist defined weight.

At the end of the evaluations, the authors informally asked the raters to express their perceptions about the process. The most frequent observations, although not majority, were the following: (1) In Zuk *et al.* it was hardest to identify, in certain visualizations, whether the assessment for some of the statements should be marked with *Not Applicable*, e.g., *Color perception varies with size of colored item* in the VisG visualization; (2) Forsell & Johansson's statements were considered more *abstract* than those of other heuristics and, therefore, took more time for evaluation and score assignment, as it was necessary to revisit the visualization several times; (3) Due to the number of statements, Wall *et al.* is the most expensive to fulfill and, consequently, the most time consuming.

6 Conclusions and Future Works

This article designed and conducted an empirical methodological study aimed to compare results of the application of three heuristic methods for the evaluation of visualizations available in the scientific literature in the area of Information Visualization, namely: Zuk *et al.*, Forsell & Johansson, Wall *et al.* The three methods were applied to evaluate five different views, also extracted from scientific articles [1,21], related to *Educational Timetabling Problems* (Ed-TTPs) context. The defined goals were the following:

1. to identify differences between the three heuristic-based methods, regarding the measure of the quality of the evaluated visualizations;
2. to check if the modifications applied by the authors of this article benefited the original heuristics, improving their efficiency when measuring the quality of visualizations in *Educational Timetabling Problems* (Ed-TTPs) context;
3. to verify whether the empirical methodological study is capable of assisting decision-making by professionals involved in Ed-TTPs.

Going beyond the simple application of the heuristic methods as they were conceived, the present authors introduced changes that aimed to provide greater adherence of the evaluation process to the particular characteristics of Ed-TTPs. This was accomplished through the definition of the relative importance between each of the statements to be evaluated, expressed by means of an *weight* for the statement. Another modification, similar to that employed by Wall *et al.*, was to collect the raters' declaration of confidence for each statement of heuristics.

From the carried out experiments, it was realized that only the association of weights with the statements, performed by the authors of this study – who are Ed-TTPs specialists – had statistical relevance during the application of the heuristics of Zuk *et al.* and Forsell & Johansson. The raters' self-declaration of confidence for each statement was irrelevant and, therefore, can be ignored. This

phenomenon was already expected, given the experience of the raters in IV field (a high confidence in the answers). When applying Wall *et al.*, it was possible to infer that, in any scenario, neither the rater's confidence nor the weighting of the statements were able to induce statistical difference in the evaluations. Hence, the scenario four was proved unnecessary. A relevant fact is that, in any scenario, the VisETV visualization is the one best evaluated, with VisMDV and VisDC occupying the second and third places, alternately.

As future works, we intend to: (1) perform a similar study involving a greater number of users, including non-specialists in IV and Ed-TTP; (2) design and apply a heuristic evaluation method focusing on Ed-TTP visualizations.

Acknowledgements

The first author is a PhD candidate and thanks the Brazilian research supporting agency FAPEG for scholarships. The others authors thanks CAPES.

References

1. Alencar, W.d.S., do Nascimento, H.A.D., Jradi, W.A.R., Soares, F.A.A.M.N., Felix, J.P.: Information visualization for highlighting conflicts in educational timetabling problems. In: Adv. in Visual Comput. pp. 275–288. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33720-9_21
2. Amar, R., Stasko, J.: A knowledge task-based framework for design and evaluation of information visualizations. In: IEEE Symp. on Inform. Vis. pp. 143–150. IEEE (10 2004). https://doi.org/10.1109/INFVIS.2004.10
3. Babaei, H., Karimpour, J., Hadidi, A.: A survey of approaches for university course timetabling problem. Comput. & Indust. Eng. **86**, 43–59 (2015). https://doi.org/10.1016/j.cie.2014.11.010
4. Carpendale, S.: Evaluating information visualizations, pp. 19–45. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-70956-5_2
5. Elmqvist, N., Yi, J.S.: Patterns for visualization evaluation. Inform. Vis. **14**(3), 250–269 (2015). https://doi.org/10.1177/1473871613513228
6. Forsell, C.: A guide to scientific evaluation in information visualization. In: Proc. of the 2010 14th Int. Conf. Inform. Vis. pp. 162–169. IV '10, IEEE Computer Society, USA (07 2010). https://doi.org/10.1109/IV.2010.33
7. Forsell, C., Johansson, J.: An heuristic set for evaluation in information visualization. In: Proc. of the Int. Conf. on Adv. Visual Interf. pp. 199–206. AVI '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1842993.1843029
8. Fu, X., Wang, Y., Dong, H., Cui, W., Zhang, H.: Visualization assessment: A machine learning approach. In: Proc. of the 2019 IEEE Vis. Conf. (VIS 19). pp. 126–130. IEEE (10 2019). https://doi.org/10.1109/VISUAL.2019.8933570
9. Gomez, S.R., Guo, H., Ziemkiewicz, C., Laidlaw, D.H.: An insight and task-based methodology for evaluating spatiotemporal visual analytics. In: Proc. of the IEEE Symp. on Visual Analytics Sci. and Techn. 2014. pp. 9–14. IEEE, IEEE (11 2014). https://doi.org/10.1109/VAST.2014.7042482
10. Harrison, L., Reinecke, K., Chang, R.: Infographic aesthetics: designing for the first impression. In: Proc. of the ACM Conf. on Human Factors in Comput. Syst. pp. 1186–1190. ACM (04 2015). https://doi.org/10.1145/2702123.2702545

11. Hearst, M.A., Laskowski, P., Silva, L.: Evaluating information visualization via the interplay of heuristic evaluation and question-based scoring. In: Proc. of the 2016 CHI Conf. on Human Factors in Comput. Systems. pp. 5028–5033. CHI '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858280
12. Hermawati, S., Lawson, G.: Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? Applied Ergonomics **56**, 34–51 (2016). https://doi.org/10.1016/j.apergo.2015.11.016
13. Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., Möller, T.: A systematic review on the practice of evaluating visualization. IEEE Trans. on Vis. and Comp. Graphics **19**, 2818–2827 (12 2013). https://doi.org/10.1109/TVCG.2013.126
14. Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical studies in information visualization: seven scenarios. IEEE Trans. on Vis. and Comp. Graphics **18**, 1520–1536 (11 2012). https://doi.org/10.1109/TVCG.2011.279
15. Mühlenthaler, M.: Fairness in academic course timetabling. Lecture Notes in Economics and Mathematical Systems No. 678, Springer (2015). https://doi.org/10.1007/978-3-319-12799-6
16. Nielsen, J.: Finding usability problems through heuristic evaluation. In: Proc. of the SIGCHI Conf. on Human Factors in Comput. Systems. pp. 37–380. CHI '92, ACM (1992). https://doi.org/10.1145/142750.142834
17. North, C.: Toward measuring visualization insight. IEEE Comp Graphics and Appl. **26**(3), 6–9 (2006). https://doi.org/10.1109/MCG.2006.70
18. Saket, B., Endert, A., Stasko, J.: Beyond usability and performance: a review of user experience-focused evaluations in visualization. In: Proc. of the Sixth Workshop on Beyond Time and Errors on Novel Eval. Methods for Vis. pp. 133–142. BELIV '16, ACM (2016). https://doi.org/10.1145/2993901.2993903
19. Santos, B.S., Silva, S.S., Dias, P.: Heuristic evaluation in visualization: An empirical study (position paper). 2018 IEEE Eval. and Beyond - Methodol. Approaches for Vis. pp. 78–85 (2018). https://doi.org/10.1109/beliv.2018.8634108
20. Stasko, J.: Value-driven evaluation of visualizations. In: Lam, H., Isenberg, P. (eds.) Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualizations. pp. 46–53. BELIV 2014, ACM, New York, USA (11 2014). https://doi.org/10.1145/2669557.2669579
21. Thomas, J.J., Khader, A.T., Belaton, B.: Visualization techniques on the examination timetabling pre-processing data. In: 2009 Sixth Int. Conf. on Comp. Graphics, Imaging and Vis. pp. 454–458. IEEE (2009). https://doi.org/10.1109/CGIV.2009.23
22. Tory, M., Moller, T.: Evaluating visualizations: do expert reviews work? IEEE Comp. Graphics and Appl. **25**(5), 8–11 (2005). https://doi.org/10.1109/MCG.2005.102
23. Wall, E., Agnihotri, M., Matzen, L., Divis, K., Haass, M., Endert, A., Stasko, J.: A heuristic approach to value-driven evaluation of visualizations. IEEE Trans. on Vis. and Comp. Graphics **25**, 491–500 (2019). https://doi.org/10.1109/TVCG.2018.2865146
24. Zhou, X., Xue, C., Zhou, L., Yafeng, N.: An evaluation method of visualization using visual momentum based on eye-tracking data. International Journal of Pattern Recognition and Artificial Intelligence **32**(5), 1850016 (2018). https://doi.org/10.1142/S0218001418500167
25. Zuk, T., Schlesier, L., Neumann, P., Hancock, M.S., Carpendale, S.: Heuristics for information visualization evaluation. In: Proc. of the 2006 AVI Workshop on BEyond Time and Errors. pp. 1–6. BELIV '06, ACM (2006). https://doi.org/10.1145/1168149.1168162