Textual-based DSL for Conceptual Database Modeling: A Controlled Experiment

Jonnathan Lopes[1], Maicon Bernardino[1], Fábio Basso[1] and Elder Rodrigues[1]

[1]Laboratory of Empirical Studies in Software Engineering (LESSE)
Postgraduate Program in Software Engineering (PPGES)
Federal University of Pampa (UNIPAMPA)
Av. Tiarajú, 810 - Bairro Ibirapuitã - Alegrete, RS, Brazil

{jonnathan.riquelmo, fabiopbasso, eldermr}@gmail.com, bernardino@acm.org

Abstract. *The variety of database system technologies that have become available in recent years makes it difficult to select tools for entity-relationship modeling (ER) in the teaching-learning context. This paper reports a replicated controlled experiment carried out with 33 subjects in order to compare effort spent (time) and quality, using the harmonic average between precision and recall, of the models produced with two different approaches. The models were produced in a proposed tool (ERtext) with a textual-based DSL and in another tool with a graphical approach (brModelo). Briefly, the data obtained indicate: i) both approaches present similar performance in relation to associated effort, and; ii) that there is a statistically significant difference in relation to the quality of the generated models, with a slightly advantage for the textual approach. Therefore, we conclude that the use of a textual-based DSL is feasible and our proposal is an acceptable solution in the context of conceptual database modeling.*

1. Introduction

Designing a model for a relational database is not a trivial task. This is due to the fact that the definition process is composed of several stages, with different levels of complexity. Software Engineering (SE) itself is intrinsically related to persistence and data manipulation. A database is a collection of stored operational data, used by systems of certain organizations in specific contexts. Nowadays, databases can be considered part of the most important assets of organizations. With this exposed scenario, it is evident the need for continuous improvement of the academy regards to the teaching-learning process in the training of professionals who master the process of building and maintaining databases. Higher education institutions often approach a database area with specific courses that converge in their programs.

Assuming that there is a growing search for instruments that support the teaching-learning process in the academy, our motivation is to provide a textual alternative for students to perform conceptual modeling given the demands of automation (DevOps) while benefiting from advanced features that the solution can offer e.g. SQL code generators for different database technologies.

It is noteworthy that the cognitive model of some students is visual and others textual, hence the importance of having different solutions to solve the same problems. Furthermore, the proposed tool aims to provide an alternative option for a smoother tran-

sition in the teaching-learning process among conceptual and logical modeling activities (both mostly visual) and physical modeling (primarily textual).

Therefore, this study aims to offer a software product for the conceptual database modeling. This software makes use of the textual approach, being built on a grammar designed for easy understanding and using. Thus, the main goal of this work is to report the replication of an experimental protocol [Lopes et al. 2021] for the evaluation our proposal ERtext, a modeling tool based on a Domain-Specific Language (DSL) [Kelly and Tolvanen 2008] for database design and modeling at the conceptual level.

This paper is organized as it follows. Section 2 presents the related work. Section 3 provides a wide sight of the tool evaluated in the experiment. Section 4 describes an overview of the replicated experimental protocol planning, conducting and results obtained. Section 5 presents the validity threats. Finally, Section 6 concludes this paper.

2. Related Work

According to Brambilla [Brambilla et al. 2017], since the beginning developers have used text to specify software products. Programming languages increase the level of abstraction in a similar way to models. Therefore, as a logical consequence, this results in textual modeling languages. They are usually processed by mechanisms that transform the information expressed in textual format for models. The execution of these mechanisms are based on the syntactic structure of a textual modeling language, which is formalized in a grammar. A grammar defines keywords in a language, the nesting of its elements and also the notation of its properties.

Hence, it can be inferred that textual models can bring some benefits [Obeo and TypeFox 2020]: (i) **Transmission of many detail**s: when it comes to elements with numerous properties, the textual approach often stands out in relation to graphics. (ii) **Increase model cohesion**: a textual model usually specifies the elements entirely in one place. While this can be a disadvantage for high-level display, on the other hand it can make it easier to find out low-level property definitions. (iii) **Perform a quick edit**: during the creation and editing of textual models there is no need for a recurring switch between keyboard and mouse. Therefore, it is likely that less time will be spent formatting textual models *e.g.* refining the position, connections or even the edges of elements in diagrams; (iv) **Use generic editors**: this is not a mandatory requirement for a specific tool to create or modify textual models. However, when we have larger tasks it is better to have some support for modeling language. Hence, this work includes the integration of a language with an Eclipse editor, thus providing a high level of assistance for writing.

Complementary to aforementioned work, our proposal involves new findings from an experimental evaluation of a tool that implements a textual DSL. After an extensive literature study, composed by a systematic literature mapping and by a multivocal review [Lopes 2019] we selected proposals and tools whose approaches are closest to the ERtext, discussed as follows.

Both studies [Celikovic et al. 2014] and [Dimitrieski et al. 2015] present a tool called Multi-Paradigm Information System Modeling Tool (**MIST**). This tool uses a DSL called EERDSL, a language based on the improved Extended Entity-Relationship (EER). MIST presents a bidirectional (graphical and textual) approach for database modeling.

The authors argue that this decision is based on the understanding that the preference for the adopted modeling approach may depend on the problem domain, knowledge and personal preferences of a database developer. They also present a previous experience, where a modeling tool was built with a forms-based approach. From the results obtained in this experiment, the idea of MIST was conceived. The purpose of the tool is to apply it both to the professional market and for teaching database design and modeling in academia. MIST was developed with the help of Xtext and Eugene frameworks, afterwards Eugene was replaced by Sirius framework. Besides, MIST supports the generation of SQL code.

The **dbdiagram.io** is a free web-based tool for ER diagram design, with a textual approach implementing its own DSL. This DSL uses a model very close to the logical data models. The tool's differential is a fast learning curve and the presentation of a graphical representation. The presentation of the diagram elements can be freely organized by the user in real time. However, it is important to note that all the modeling is in fact done textually. Furthermore, the tool also offers automatic generation of SQL code.

Likewise, **QuickDBD** is a Web-based tool with similar operational mode as dbdiagram.io, also implementing its own textual DSL for modeling databases. However, it is a proprietary tool with a clear focus on the industry. Both tools are also very similar in terms of the generation of graphic representations and present several attributes for their adoption, such as the quick DSL understanding, the perspective of carrying out fluid works, the access of any platform and the sharing of models with other users.

Finally, we can mention the free Web-based tool **RelaX** [Kessler et al. 2019]. It is a tool aimed at teaching relational algebra by performing operations on databases. RelaX uses a modeling approach already at a physical data model level, *e.g.* data definition and data manipulation languages. Despite its feature, RelaX is not characterized as a database designing and modeling tool and their use is restricted for teaching within the academia.

3. The Modeling Tool

This section presents an overview of the tool, with the DSL implemented in the final version of the plugin for the proposed solution[1]. This plugin can either be integrated with the Eclipse Rich Client Platform (RCP), or it can be used as a standalone application. The difference is that when used as an Eclipse plugin the editor can provide, in addition to the grammar features, support for other languages, *e.g.* Java, PHP. On the other hand, a standalone product provides the entire infrastructure focused solely on the developed language and can be distributed as an open-source tool as long as the EPL-2.0 software license guidelines are followed.

Our DSL is defined in three main blocks: domain, entities and relationships. This blocks are like objects that represents a kind of container, this being indicated through the assignment operators. The domain block, as the name implies, identifies the domain that the data model represents. The entity block contains the entities, which contains attributes with data type. In addition, optional representation of inheritance is supported. The relations are described, already inside the relationships block, with an optional declaration of their identification. They consist of a description with two sides, right and left, which indicates the entities that are related and their associated cardinalities. Representation

[1]Solution Repository: `https://github.com/ProjetoDSL/ERDSL`

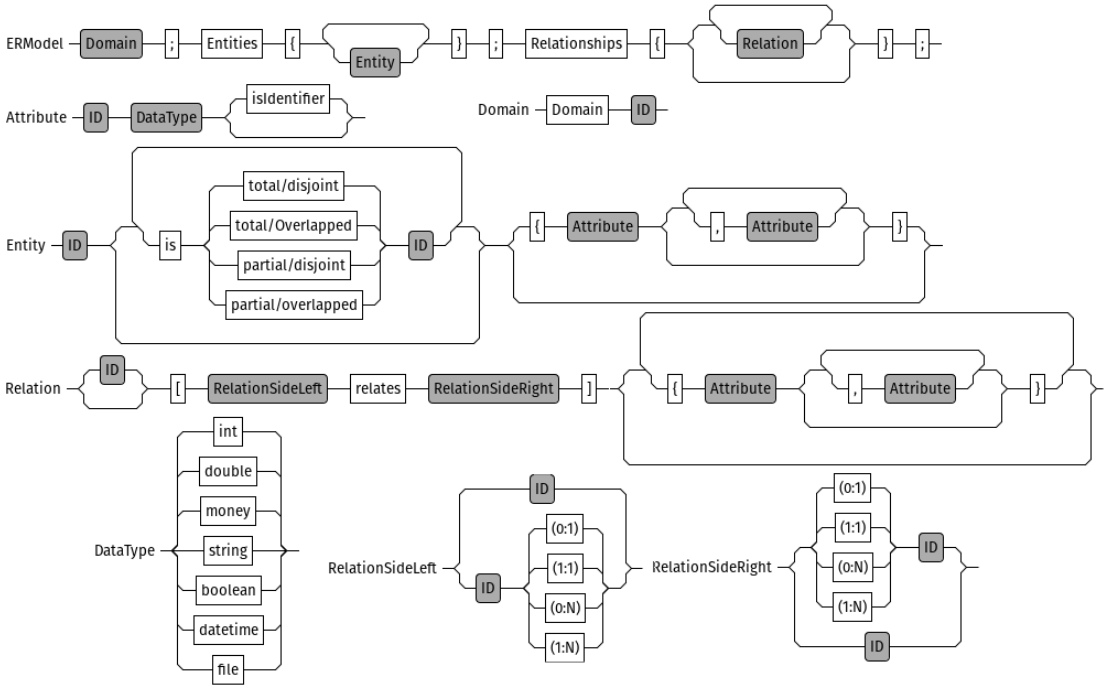


Figure 1. ERtext syntax diagram.

of ternary/associative entities is also supported, in addition to attributes associated with relationships. Figure 1 shows the DSL syntax diagram.

The software architecture was built with the help of Xtext, an open-source framework for the development of textual programming languages. Hence, from the final grammar most of the editor's infrastructure was generated, the parser and Ecore model. For understanding, Ecore is a representation in memory at runtime of the model created using DSL. Then, there were some necessary manual adjustments and the addition of the writing of a generator for the conversion of conceptual models to logical models.

4. Controlled Experiment

In this section we presented the research questions and hypotheses investigated, besides the experimental design adopted and its conduction. The results obtained as well as their discussion analysis are also exposed.

4.1. Planning

The replication of this experimental protocol aims to obtain evidence from the comparison of two approaches for modeling relational databases, one graphical and the other textual. The treatments identified were: (i) Control treatment: the brModelo tool (graphical approach), and; (ii) Experimental treatment: the ERtext tool (textual approach). The purpose of this replication is to evaluate the feasibility of using a textual approach to support the teaching-learning process of conceptual modeling of relational databases.

Context: The context of the experiment is characterized according to four dimensions: (i) Process: an *in-vitro* approach was used, since the tasks were performed under controlled conditions. (ii) Subjects: undergrad students in Computer Science (CS) and SE programs. (iii) Reality: the experiment addressed a real problem, *i.e.* the difference in the

effort spent of subjects in the conceptual modeling of relational databases. (iv) Generality: this evaluation is inserted in a specific context, involving database modeling students. However, the general ideas of this experiment can be replicated in another set of subjects, approaches or DSLs that support database designing.

Research Questions (RQs): For the discussion of the controlled experiment results, we decided to formulate four RQs that were related to the activities performed in the protocol execution. **RQ1.** Which approach requires the most effort spent on average during the modeling activity? **RQ2.** What is the quality level of the models produced using the graphical and textual approaches? **RQ3.** What is the subjects perception regarding the Perceived Ease of Use (PEoU) and Perceived Usefulness (PU) of the proposed DSL? **RQ4.** What is the subjects assessment in relation to the representation of the ER modeling builders supported in the proposed DSL?

Hypotheses Formulation: The first two RQs were taken into account. Regarding to **RQ1.** we defined a two-sided hypotheses that measure the average effort spent between textual and graphical approaches during conceptual modeling. State the null (no difference) $H_0 : \mu Time_T = \mu Time_G$ and alternative (significant difference) $H_a : \mu Time_T \neq Time_G$ hypotheses. Regarding to **RQ2.** in the same way we stated a two-sided hypotheses that measure the modeling effectiveness between textual and graphical approaches during conceptual modeling. The null (no difference) and alternative (significant difference) hypotheses are, respectively: $H_0 : \mu Effectiveness_T = \mu Effectiveness_G$ and $H_a : \mu Effectiveness_T \neq \mu Effectiveness_G$.

Statistical Methods: Unlike the first experiment, this replication included a change in the statistical methods adopted. Previously, the Shapiro-Wilk normality test and the paired T-test for dependent samples were used. This was because the sample was smaller (27) than 30 elements. However, for samples equal to or greater than this quantity, alternative tests are recommended [Triola 2018].

For the effort we chose the Kolmogorov-Smirnov test to verify normality, and the Wilcoxon Signed-Rank test for paired samples to investigate the hypotheses using the time spent metric. For the effectiveness tests, the same statistical methods were adopted, but instead of using the time spent metric, another measure was necessary. Thus, the F1 calculations were performed, which is derived from harmonic mean of *Precision* and *Recall* metrics, for each of the models produced in both approaches. The F1 calculation [Derczynski 2016] takes into account variables known as *True Positives*, *False Positives* and *False Negatives*. ***True Positives (TP)***: Amount of elements correctly modeled using the approach. ***False Positives (FP)***: Amount of elements incorrectly modeled using the approach. ***False Negatives (FN)***: Amount of elements not modeled using the approach. From the variables identification it is then possible to calculate the *Precision*, *Recall* and *F1* of each model according to these formulas:

Precision (PR): $\frac{TP}{TP + FP}$ ***Recall (RE)***: $\frac{TP}{TP + FN}$ ***F1-Score (F1)***: $\frac{2 * (PR * RE)}{PR + RE}$

Experiment Design: Finally, the design of the controlled experiment performed is presented in Figure 2. We followed the design of one factor with two treatments, where we blocking, balancing and randomizing the subjects, which carried out both treatments, featuring a paired comparison design.

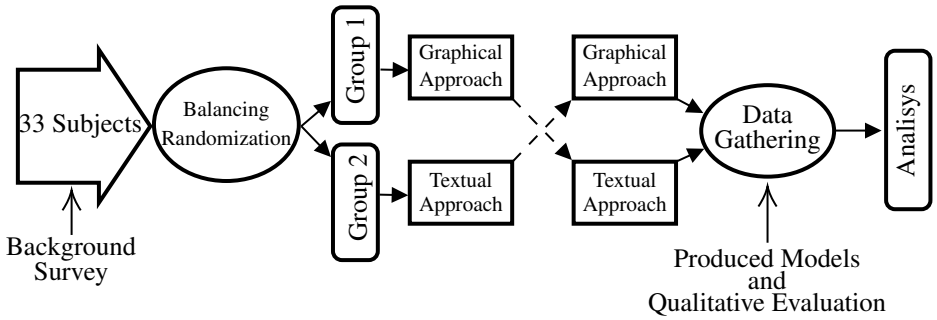


Figure 2. Experiment design.

4.2. Conduction

Preparation: Initially, remote meetings were held between the researchers involved to define the planning and the mode of operation that should be adopted, in response to the current scenario of exception due to the worldwide pandemic. As a result, activities were defined that should be adapted in relation to the first experiment, which was conducted in person. In order to capture a significant sample for the object of study, it was decided to contact the professors responsible for teaching two courses of different undergraduate courses: Database (SE) and Database I (CS) in the first half of 2021. With the initial objectives aligned, the collaborating teachers made the disclosure of profile questionnaires (Google Forms) to the subjects.

The four (4) instruments from the original experiment were reused. The first two were modeling problems with similar levels of complexity, while the last two were of qualitative assessment. It was decided that the activities would be carried out remotely, respecting the health security protocols required (social distancing). For this purpose, a virtual machine was prepared with the tools installed, as well as the instruments and supporting materials. This virtual machine should be accessed on the university's computers by the subjects using their institutional credentials.

Execution: The profile form also served as a term of participation, since the presence in the experiment was voluntary. With this information, the subjects were then randomized to define the groups. We found that there were no major discrepancies between the levels of knowledge of the subjects, demonstrating that there was a homogeneous sample in general. On the experiment day, the first activity carried out was a brief initial presentation. Then, the training phase of the participants began. During this phase, the two database modeling tools that would be used were presented, providing an overview of the operation and answering possible questions that arose. The training included the display of videos demonstrating the tools, brModelo and our proposal, respectively.

Then, we start the modeling phase of the proposed problems. All subjects accessed the virtual machines with the problems provided in PDF documents. When starting the Instrument 1, all participants were informed with which tool they should develop the solution, thus respecting the groups to which they were part. We asked the subjects to write down the start and end times of the tasks for each instrument they performed. We stipulated no time limit for completion and, according the subjects completed the modeling task, they were asked to comply with the guidelines included in the support material for

saving the generated artifacts. With the models saved, we informed the subjects that they should move on to the next task described in Instrument 2, although it was necessary to use the reverse approach to the one they had initially used.

At the end of the instruments that contained the modeling problems, we delivered the qualitative assessment instruments. As the subjects had completed the tasks, then we had thanked and released them. With the conclusion of the experiment by 33 subjects, we close the evaluation and we performed the stage of result analysis.

4.3. Results Analysis

All Kolmogorov-Smirnov and Wilcoxon Signed-Rank calculations were performed with the support of the R language and the Gnumeric software, in parallel with the validation of a specialist in the field of statistics and the aid of literature [Triola 2018].

Effort: To answer **RQ1.** the execution times were extracted from the instruments. From the gross amount of the execution times, we calculated the difference in order to be able to perform the Kolmogorov-Smirnov normality test. Because it is a statistical test, this technique has the product of measuring the p-value. For this test, we adopted a significance level of $\alpha = 5\%$. This means that the p-value is less than 5% ($p < 0.05$), a hypothesis that the distribution is normal should be rejected.

After calculations with the set of time differences, we reached a p-value of 0.26218. As p-value $> \alpha$, we do not reject the null hypothesis, thus concluding that the data is normally distributed. In other words, the difference between the data sample and the normal distribution is not large enough to be statistically significant. It is important to note that the higher the p-value, the more it supports a null hypothesis. In the case of the result obtained, the chance of type 1 error (rejecting a null hypothesis that is correct) is very high, and can be translated into 26.21% (0.26218). Once we performed the normality tests on the sample, we carried out the hypothesis test of the average effort regarding to **RQ1.** In the Wilcoxon Signed-Rank test for dependent samples, we used a significance level of $\alpha = 5\%$, with which we reached a measure of 0.77948 for the p-value. Because it is a two-tailed test, *i.e.* it includes equality in its null hypothesis, this p-value shows no enough evidence to guarantee the rejection of the statement of $H_0 : \mu Time_G = \mu Time_T$. Therefore, we do not reject null hypothesis that the approaches has no difference in average efforts, once according to the test this difference is not statistically significant. Figure 3a displays a box-plot with the variation of data observed through these data.

Effectiveness: To answer **RQ2.** regarding the effectiveness of the use of approaches, we evaluated the artifacts produced by the subjects according to the established reference models[2]. In this evaluation, we used F1 from the area of pattern recognition and information retrieval. F1 represents the combination of the observed accuracy and recallability of a result in relation to a reference. By definition, this combination refers to Precision and Recall metrics, where Precision is the proportion of recovered instances that are relevant and Recall is the proportion of relevant instances that are recovered.

In addiction, we performed the Kolmogorov-Smirnov normality test to F1 for each model. After calculations with the set of differences in F1 for each model, we reached a p-value of 0.45459. With this test result, the chance of type 1 error (rejecting a null

[2]Available at: `https://doi.org/10.5281/zenodo.5454378`

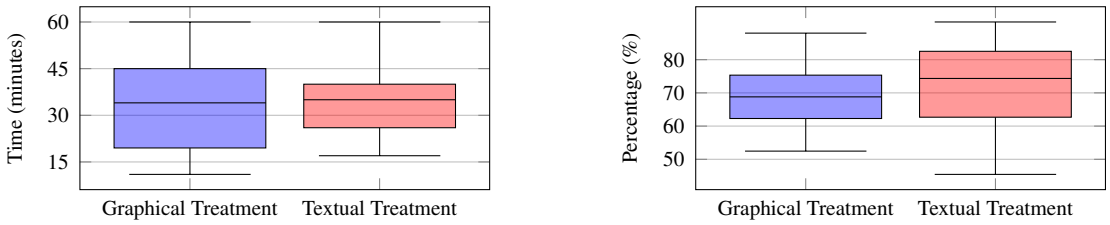
hypothesis that is correct) can be very high, and can be translated into 45.45% (0.45459). As the p-value > α, we do not reject the null hypothesis, thus realizing that the data is normally distributed After the sample was tested for normality, we tested the second hypothesis defined in this experiment. This time, in the Wilcoxon Signed-Rank test for dependent samples, we used again a significance level of $\alpha = 5\%$, with which we reached a measure of 0.00197 for the p-value. By the original statement including an equality, also characterizing this test as two-tailed, it was concluded that the calculated p-value demonstrates that there is enough evidence to guarantee the rejection of the statement of the original null hypothesis, denoted as $H_0 : \mu Effectiveness_G = \mu Effectiveness_T$. Therefore, we reject the null hypothesis that the approaches have equal effectiveness, because according to the statistical test, the average difference of F1 between treatments is statistically significant. Table 1 shows average measures of the evaluated values, and also provides the possibility to carry out a dispersion analysis. Based on these data it was possible to verify that the textual approach has an advantage on average.

Table 1. Measures of the conceptual data models produced in the experiment.

	Graphical Treatment					Textual Treatment				
Measure	**MI**	**RI**	**Precision(%)**	**Recall(%)**	**F1(%)**	**MI**	**RI**	**Precision(%)**	**Recall(%)**	**F1(%)**
Maximum	47.00	36.00	96.67	92.31	88.00	56.00	46.00	97.22	97.87	91.36
3° Quartile	31.00	28.00	92.31	63.04	76.32	34.00	31.00	94.74	75.00	82.86
Median	26.00	24.00	88.89	56.41	68.85	30.00	29.00	90.63	65.96	74.63
Average	27.52	24.12	87.69	57.96	69.13	30.88	27.45	89.49	63.65	73.16
1° Quartile	23.00	20.00	84.21	50.00	62.50	26.00	23.00	87.88	51.06	63.01
Minimum	18.00	16.00	72.73	41.03	52.46	19.00	15.00	72.73	31.91	45.45
Variance	35.58	28.65	42.87	143.35	78.93	63.32	39.76	42.59	259.85	133.58
Standard Deviation	5.97	5.35	6.55	11.97	8.88	7.96	6.31	6.53	16.12	11.56

Legend: MI = Modeled Items; RI = Relevant Items; F1 = F1-Score.

Figure 3b box-plot graph displays of the F1-Score for each treatment applied. Based on this graph, it is possible to verify the result obtained in the hypothesis test because the data dispersion does present a significant difference between the approaches.



(a) Box-plot - Effort per treatments.

(b) Box-plot - F1 per treatments.

Figure 3. Box-plot graphs per treatments.

Qualitative Evaluation: We took place with the analysis of the two instruments applied after the modeling tasks. The first was used to respond to **RQ3.** regarding the PEoU and PU of treatments, according to the TAM model [Davis 1989, Persico et al. 2014]. This occurred through the selection of quality attributes described in ISO/IEC 25010. For this, we established a Likert scale from one to six points to measure the level of agreement of the subjects in the face of the statements exposed in the form. We chosen an even number of alternatives to avoid possible neutral responses. Thus, the 7 quality attributes are grouped in 3 categories being defined as follows: 1. **Functionality** - *Conformity*: ability level to which the software to achieve specified goals with functional

completeness, correctness and appropriateness related to their functionalities. 2. **Usability** - *Understandability*: ability level to which users can recognize whether a software is appropriate for their needs; *Learnability*: ability level to which the software enables the user to learn how to use it with effectiveness, efficiency in emergency situations; *Operability*: ability level to which the software is easy to operate, control and appropriate to use. 3. **Quality in Use** - *Quality in Use*: ability level to which the software to achieve specified goals with effectiveness and efficiency with their users in specific contexts of use; *Productivity*: ability level to which the software to achieve specified goals with time-behavior, resources utilization and capacity, when performing its functions, meet requirements; *Satisfaction*: ability level to which the software to achieve specified goals with usefulness, trust, pleasure and comfort with their users in specific contexts of use.

After summarizing the results, we observed a good acceptance by the subjects for the ERtext tool, developed in this work. Figure 4 synthesizes the responses received for each quality attributes, showing a certain degree of similarity in the subjects perception during the treatments application. A point that can be emphasized is the set of positive responses in relation to the *Productivity* and *Operability*, since in the hypothesis test related to the effort, the treatments demonstrated a similar need for execution time. According to the evaluations received, the disadvantages of ERtext that are most evident are manifested mainly with regard to *Understandability* and *Learnability* quality attributes.

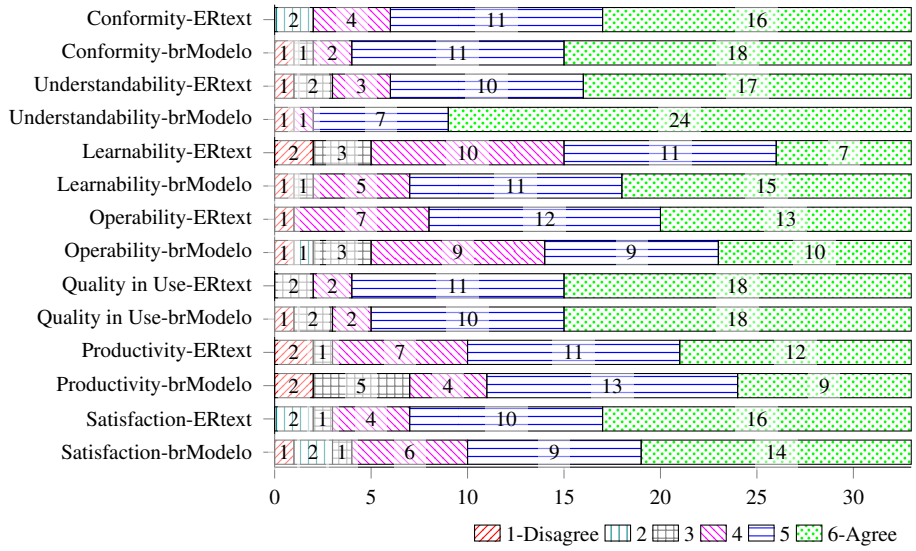


Figure 4. Quality attributes per treatments.

With regard to **RQ4.** on the assessment of DSL designers, we analyzed the artifacts of the 2nd qualitative assessment instrument. This instrument listed the 8 ER modeling builders covered by DSL, arranged with a Likert scale from one to six points. Again, an even number was chosen on the scale to avoid neutral responses that could lead to a more subjective bias. Figure 5 compiles all the responses received. The builders related to Entities, Referential Attributes, Descriptive Attributes and Cardinality were the best evaluated. In contrast, all builders obtained at least one disagreement, highlighting the most disagreeing evaluations related with the current representations of Ternary Relationship and Self-relationship, unfortunately.

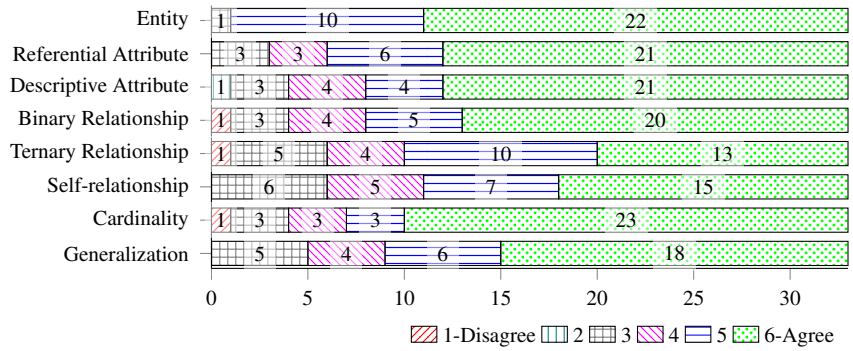


Figure 5. Evaluation of DSL designers.

All data collected and used for statistical tests can be accessed in a public repository available at Zenodo[3].

5. Threats to Validity

In empirical studies it is necessary to analyze and discuss the threats to validity, as well as the strategies used to mitigate them. For the list of possible threats, we adopted the classification scheme published by Cook and Campbell [Cook and Campbell 1979]. These threats followed the proposed pattern and were divided into four (4) categories:

Construct Validity (i) The fact that the experiment did not have the objective of the artifacts sufficiently defined before translation into measures or treatments. To mitigate this threat the effort of each approach was compared, as well as their effectiveness carried out according to the Precision and Recall metrics. (ii) If the subjects involved in one more study, the controls of the different studies, can change and reverberate in the final results. To avoid this, we followed a paired design and, therefore, all subjects performed both treatments. However, learning issues among the execution of activities were not observed. This can be verified through the analyzed distributions normality of the both samples: effort and effectiveness, demonstrating that the results remained similar as a whole with a low variation. (iii) We not informed the subjects about further details of the experiment to mitigate the bias the behavior, that can affected positively or negatively, depending on the anticipated hypothesis.

Internal Validity (i) There is a risk when a specific time period influences the experiment performance. Due to the coronavirus pandemic we conducted the entire experiment in a controlled environment providing remote access for the subjects. (ii) Sometimes the subjects could be affected negatively (tiredness or boredom) or positively (learning) during the experiment execution. In order to alleviate this threat, we informed subjects from the beginning that they could terminate their participation at any time, without any penalty. (iii) All artifacts were minimally adapted due to the design of the experiment (remote), and previously verified and validated in meetings between the researchers involved in this work to avoid that the instrumentation of artifacts can be affect the execution.

External Validity (i) The experiment was carried out with undergrad students of SE and CS programs, and soon, inserted in the context of using the conceptual database

[3]Available at: `https://doi.org/10.5281/zenodo.5454378`

modeling seeking to mitigate the selection subjects from a significant group for an study area. (ii) The participants were asked to answer the questionnaires post-experiment and, before thanking them, we checked the submission of the instruments to avoid that the subjects interaction with the evaluation artifacts can be affected the experimental results. (iii) We used the documentation based on templates and traditional models found in database teaching material to soften an unrepresentative configuration and material.

Conclusion Validity (i) Some statistical methods were adopted, such as the Kolmogorov-Smirnov normality test, the Wilcoxon Signed-Rank Test as a hypothesis test for dependent samples to try mitigating the low statistical power of our experimental results. (ii) We adopted objective measurements that did not depend on subjective judgment (effort spent, measured in time, and F1) to mitigate the measurements reliability threat. On the other hand, the metrics used for the qualitative evaluation still served as a complementary input in the discussion of the results obtained, alongside with the indication of possible points of improvement in our proposal. (iii) We created copies of a virtual machine that were accessed by all subjects to guarantee that the experiment was carried out in a controlled environment.

6. Final Remarks

This study presented a controlled experiment replication with a sample of 33 subjects evaluating ERtext, a proposed textual DSL tool for database conceptual modeling. ERtext is compared with the brModelo, a graphical DSL well-known in Software Engineering ER lectures. With the results obtained it was possible to answer the four RQs of the experiment, as well as the two associated hypotheses. In this sense, we can say that three main aspects of our DSL are investigated: effort, effectiveness and quality in use.

From the analysis it is possible to highlight the following aspects: (i) **Effort:** the computed average difference states that there is no differences between the approaches, *i.e.*, one approach is not better than the other. (ii) **Effectiveness:** through the statistical test we reject the null hypothesis that the approaches are equally effective. When comparing the boxplot graphs produced, it is possible to observe a slight advantage for the textual approach. In the first experiment, the result was different, since the tests indicated that there was no significant difference. We believe that the biggest reason for this event is the fact that all participants in this replication are undergraduate students at the very beginning in the academy program, basically without anymore previous experience with graphic modeling tools. In the first experiment, there were, in addition to these, master's and doctoral students. We also do not discard possible influences caused by the fact that the experiment was conducted remotely in the pandemic context. (iii) **Qualitative comparison between treatments:** we observed a certain balance between treatments, but with a positive evaluation for ERtext regarding the "Productivity" and "Operability" attribute. Because it was the first time that the subjects had contact with our grammar, and also considering a first release of our DSL and the negative feedback also collected, we conclude that ERtext is on the rails for achieving better productivity indexes.

We also collected qualitative feedback from subjects. As a result, there are some improvements regarding the DSL design that need to be revised, in particular to the ternary relationships and self-relationships. Since the continuity of DSL development is foreseen, the execution of a possible refactoring, and also the implementation of new ER builders,

is a natural step for its software evolution. From the experimental results we conclude that there is feasibility and good perspectives for the motivated context, *i.e.*, as a tool for teaching ER modeling with the differential of adopting a textual approach for conceptual database modeling in classrooms instead of a graphical notation.

References

Brambilla, M., Cabot, J., and Wimmer, M. (2017). *Model-Driven Software Engineering in Practice, Second Edition*. Synthesis Lectures on Software Engineering. Morgan & Claypool.

Celikovic, M., Dimitrieski, V., Aleksic, S., Ristic, S., and Lukovic, I. (2014). A DSL for EER data model specification. In *23rd Int. Conf. on Information Systems Development*, pages 290–297, Varaždin, Croatia. Springer.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin, Chicago, IL, USA.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *Management Inf. Systems Quarterly*, 13(3):319–340.

Derczynski, L. (2016). Complementarity, f-score, and NLP evaluation. In *10th Int. Conf. on Language Resources and Evaluation*, pages 261–266. ELRA.

Dimitrieski, V., Čeliković, M., Aleksić, S., Ristić, S., Alargt, A., and Luković, I. (2015). Concepts and evaluation of the extended entity-relationship approach to database design in a multi-paradigm information system modeling tool. *Comput. Lang. Syst. Struct.*, 44:299–318.

Kelly, S. and Tolvanen, J.-P. (2008). *Domain Specific Modeling: Enabling Full Code Generation*. John Wiley & Sons.

Kessler, J., Tschuggnall, M., and Specht, G. (2019). Relax: A webbased execution and learning tool for relational algebra. In *Datenbanksysteme für Business, Technologie und Web*, pages 503–506. Gesellschaft für Informatik.

Lopes, J., Bernardino, M., Basso, F., and Rodrigues, E. (2021). Empirical evaluation of a textual approach to database design in a modeling tool. In *Proc. of the 23th International Conference on Enterprise Information Systems (ICEIS)*, page 8. SciTePress.

Lopes, J. R. (2019). Ertext: uma linguagem específica de domínio para a representação de modelos conceituais de bancos de dados relacionais (in portuguese).

Obeo and TypeFox (2020). Xtext/sirius - integration the main use-cases. Technical report, Obeo and TypeFox.

Persico, D., Manca, S., and Pozzi, F. (2014). Adapting the technology acceptance model to evaluate the innovative potential of e-learning systems. *Computers in Human Behavior*, 30:614–622.

Triola, M. (2018). *Elementary Statistics*. Pearson.