

Empirical Evaluation of a Textual Approach to Database Design in a Modeling Tool

Jonnathan Lopes^a, Maicon Bernardino^b, Fábio Basso^c and Elder Rodrigues^d

Postgraduate Program in Software Engineering (PPGES),
Laboratory of Empirical Studies in Software Engineering (LESSE),
Federal University of Pampa (UNIPAMPA), Av. Tiarajú, 810, Ibirapuitã, CEP 97546-550, Alegrete, RS, Brazil

Keywords: Domain Specific Language, Conceptual Data Model, Conceptual Model, Database Modeling.

Abstract: This article reports an empirical assessment study conducted with 27 subjects intended to verify the effort (time spent), precision, recall, *F-Measure* of a proposed tool based on a textual approach (ERtext), thereby a study developed for comparative reasons of ERtext with a tool based on graphical approach: thebrModelo. The results are: 1) less effort is associated with the graphical approach and that ERtext, and 2) regarding the model quality, brModelo have similar performances in both approaches. Since the results shows no considerable statistical differences among the two design approaches, we conclude that the usage of a textual approach is feasible, thus ERtext is a good tool alternative in the context of conceptual database modeling.


1 INTRODUCTION


Software Engineering is not feasible without persistence and data manipulation. A database is a collection of stored operational data, used by the application systems of a given organization. For Elmasri and Navathe (2011), a database can be defined as an abstraction from the real world, also called the mini-world, since it represents aspects that, together, carry an implicit meaning.


Analyzing more objectively, databases can be considered the most important organizational assets today. This is due to the fact that they do not only store trivial information but, *e.g.* also billing data, and other aspects assisting in decision making. However, its importance is not only focused on organizations, it is also possible to attest that the use of database can play a critical role in the lives of end users when they are analyzed individually. In this scenario, it is notable that there is an increasing effort of the academia to provide a good level of preparation for future professionals who will enter an increasingly demanding industry. Higher education institutions often approach the database area with specific courses converging in their programs. The database teaching area is an es-


sential part of computer professional training. The focus on database teaching is generally divided into four stages: design and modeling, database management systems, comparative studies between these systems and the development of applications (Al-Dmour, 2010). Assuming that there is a growing search for instruments supporting the teaching-learning process in academia, we present a study focusing on the first stage. In general, the teaching of database design and modeling is conducted with the presentation of essential topics and the subsequent introduction to the use of modeling tools using generally graphic approaches. Hence, this paper is motivated to offer a software product for the conceptual modeling of relational database. This software makes use of the textual approach, built on a grammar perceived in the proposed study as “ease to use” and as “useful”. Thus, the main objective of this paper is to report the evaluation of ERtext, a modeling tool based on a Domain-Specific Language (DSL) (Kelly and Tolvanen, 2008) for database designing and modeling with the scope in conceptual level modeling.

This paper is organized as it follows. Section 2 presents the related work. Section 3 describes an overview of the ERtext modeling tool. Section 4 provides the detailed planning and execution of the experiment. Section 5 briefly summarizes the threats to validity. Finally, Section 6 concludes this paper.

^a  <https://orcid.org/0000-0001-9402-801X>

^b  <https://orcid.org/0000-0003-2776-8020>

^c  <https://orcid.org/0000-0003-4275-0638>

^d  <https://orcid.org/0000-0003-3431-2814>

2 RELATED WORK

According to 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. A textual modeling language is usually processed by mechanisms that transform the information expressed in textual format for models.

Hence, it can be inferred that textual models can bring some benefits Obeo and TypeFox (2020): (i) Transmission of many details: 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; (iv) Use generic editors: there is not necessarily a requirement for a specific tool to create or modify textual models. For simple changes it is possible to use any generic text editor. However, for 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.

Complementary to aforementioned work, our proposal involves new findings from an 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, we selected proposals and tools whose approaches are closest to the ERtext, discussed as follows.

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 to database modeling. 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, a project that has been discontinued, for its graphical version. As a result, Eugene was replaced by Sirius framework. Besides, MIST also supports the generation of SQL code.

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 (Relational Algebra Calculator) Kessler et al. (2019). It is a tool aimed at teaching relational algebra by performing operations on relational databases. It has a textual approach, using a DSL called RelAlg, and even presenting two operation perspectives: RelAlg instructions and SQL statements. Relax uses a modeling approach already at a physical data model level, e.g. data definition and data manipulation languages. Despite its functionality, Relax is not characterized as a database designing and modeling tool and their use is restricted to teaching within the academia.

3 ERTEXT MODELING TOOL

3.1 Software Requirements

Our focus is on the teaching process, so it is essential to tag ERText as an open-source license, allowing the evolution and collaborative maintenance with the involvement of other developers.

In the following, we listed the Software Requirements (SR) that were defined based on the surveyed literature: **SR1.** DSL must be made available under an open-source license. **SR2.** The DSL should allow for the textual representation of conceptual data models. **SR3.** Conceptual data models should support the definition of fundamental domain concepts such as entities, attributes, relationships, and their cardinalities. **SR4.** Conceptual data models should support the definition of attributes, identifiers, generalization and specialization, self-relationships, and ternary relationships. **SR5.** DSL implementation must transform from the conceptual to the logical model displaying the result generated to the user.

3.2 The Language

In the current phase of the work, the language at the conceptual level is not fully finalized. There are topics related to scope validation, as in the case of the treatment of unwanted cross-references and other restrictions inherent to the ER model that must be analyzed and then implemented. The definition of the created DSL is displayed in Figure 1.

```

grammar org.xtext.unipampa.lesse.ertext.ERtext
with org.eclipse.xtext.common.Terminals
generate ERtext "xtext.org/unipampa/lesse/ertext"
ERModel:
    domain=Domain ';'
    ('Entities' '(' entities+=Entity+ (' ' ';')
    ('Relationships' '(' relations+=RelationA(' ' ';')
    );
Domain:
    'Domain' name=ID;
Attribute:
    name=ID type=DataType (isKey?='isIdentifier'?);
Entity:
    name=ID ('is' is+=[Entity])A
    (' ' attributes+=Attribute
    (' ' attributes+=Attribute)A ' ')?;
Relation:
    (name=ID)? (' ' leftEnding=RelationSide
    'relates'
    rightEnding=RelationSide ' ')?
    (' ' attributes+=Attribute
    (' ' attributes+=Attribute)A ' ')?A;
RelationSide:
    cardinality=('(' (0:1)' | '(1:1)' | '(0:N)' | '(1:N)
    ')')
    target=[Entity] | target=[Relation];
enum DataType:
    INT='int' | DOUBLE='double' |
    MONEY='money' | STRING='string' |
    BOOLEAN='boolean' | DATETIME='datetime' |
    BLOB='file';

```

Figure 1: Grammar definition of the DSL.

4 EMPIRICAL EVALUATION

This section exposes the controlled experiment that was performed in order to verify the feasibility of using the textual approach (ERtext) developed in this work in the context of teaching entity-relationship modeling. We followed the guidances and recommendations raised in Wohlin et al. (2012).

4.1 Planning

The experiment aims to obtain evidence from the comparison of two approaches to the modeling of relational databases, one in a graphical and another in a textual way. The titled treatments, were: (i) Control treatment: the brModelo tool, with a graphical approach, and; (ii) Experimental treatment: the ERtext tool, with a textual approach. The purpose of this experiment is to assess the feasibility of using a textual approach to support the teaching-learning process of conceptual modeling of relational databases.

Research Questions (RQs): For the discussion of the experiment results, we decided to formulate four RQs that were related to the activities performed.

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?

Context: The experiment context is characterized according to four dimensions: **(i) Process:** An *in-vitro* approach was used, since the tasks were performed in the lab under controlled conditions and without online activities. **(ii) Subjects:** Undergrad, master and doctoral students in Computer Science and Software Engineering. **(iii) Reality:** The experiment addressed a real problem, that is, the difference in the effort spent of subjects in the conceptual modeling of relational databases, the artifacts quality produced and the subjects perceived usefulness (PU) using both approaches. **(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 modeling.

Hypotheses Formulation: the first two RQs were taken into account. Regarding to **RQ1**, the average effort spent required using each approach, our scientific hypotheses are as follows:

Null Hypothesis: $H_0 : \mu Time_G = \mu Time_T$: There is no difference in average effort spent measure between textual and graphical approaches during conceptual modeling of relational databases.

Alternative Hypothesis: $H_1 : \mu Time_T \neq Time_G$: There is a significant difference in average effort spent measure between textual and graphical approaches during conceptual modeling of relational databases.

Regarding to **RQ2**, the modeling effectiveness using each approach, our hypotheses are as follows:

Null Hypothesis: $H_0 : \mu Effectiveness_G = \mu Effectiveness_T$: There is no difference in effectiveness measure between textual and graphical approaches during conceptual modeling.

Alternative Hypothesis: $H_1 : \mu Effectiveness_T \neq \mu Effectiveness_G$: There is a significant difference in effectiveness measure between textual and graphical approaches during conceptual modeling.

For the evaluation related to the effort measure, the Shapiro-Wilk normality test and the paired T-test for dependent samples were used, in which the times collected during the execution of the modeling activi-

ties of the experiment were taken into account.

For the effectiveness tests, the same statistical methods were adopted, but instead of using the time metric, another quantity was necessary. Thus, the *F-Measure* calculations were performed, which is derived from harmonic mean of *Precision* and *Recall* metrics, for each of the models produced in the approaches. The *F-Measure* (Derczynski, 2016) calculation takes into account variables known as *True Positives*, *False Positives* and *False Negatives*. From the variables identification it is then possible to calculate the *Precision*, *Recall* and *F-Measure* of each model.

Selection of Subjects: The subjects were selected by non-probabilistic sampling, indicated for exploratory studies. This type of sampling is characterized by the deliberate choice of subjects with one or more characteristics that interest to the study object.

In this context, 27 undergrads (Computer Science and Software Engineering) and post-grads (Software Engineering) students from our university participated in this experiment. Among the subjects there were 14 students enrolled in database course. After identifying the potential subjects, the execution date of the controlled experiment was defined, which included training in both approaches. In addition, before the training, subjects would be asked to complete a profile questionnaire for leveling. From the data extracted of questionnaires, the subjects were randomly distributed into two groups composed of 13 and 14 subjects. The reason for that it is because the total number of subjects was odd, and we also tried to maintain a balanced level of skills among the groups.

Experiment Design: According to Wohlin et al. (2012), a controlled experiment must meet some fundamental concepts: (i) Standard Design Type: There are a few possible types for the standard design in an experiment, and this study adopted One Factor with Two Treatments. The Factor was the modeling of relational databases, and the Treatments were the two approaches used (graphical and textual). (ii) Blocking: This item refers to the fact that the subjects of the controlled experiment may have different experience levels in database modeling and design. As a result, a profile questionnaire was applied to level the subjects. (iii) Balancing: Subjects were separated into two groups with similar background levels. In this way, both approaches were carried out by homogeneous groups. (iv) Randomization: The subjects were randomly allocated to each group and approach. They performed both treatments, featuring a paired comparison design. The execution sequence of the treatments for each group was also randomly defined.

After all the activities performed by the subjects using the treatments, we collect the data results. This

stage consists of the qualitative assessment made by the subjects and the saving of the models produced by the treatments application. These models serve for a qualitative assessment. Finally, the analysis stage is performed, where the result data are compiled and analyzed with the aim of drawing conclusions.

Instrumentation: As participation in this experiment was voluntary, a Free, Prior and Informed Consent (FPIC) was prepared to record the agreement of all them in carrying out the activities. Profile questionnaires were created and applied to balance the groups. Beyond, a glossary was also developed for concepts that were used in the opening presentation of the controlled experiment and during training. In order to provide support to the evaluation subjects, instruments were provided describing the step by step with use examples of both tools (brModelo and ER-text) used in this experiment. In addition, training was conducted that included videos with tutorials on how to use the approaches in each tool. The videos showed how to start modeling, problem examples of the builders foreseen that they would solve and recommendations for saving the artifacts produced.

We elaborated two instruments that contained a problem each, with similar levels of complexity, which should be modeled and also noted the start and end times of the activity. For further subjects evaluation, we prepared two other instruments. The first instrument presented 7 quality attributes based on ISO/IEC 25010, a standard for software product quality, and served to evaluate the approaches of the two tools from the point of view of the subjects who performed the ER modeling activities. The second instrument was used to evaluate the ER modeling builders representation of the textual approach solution evaluated in this study. The data generated by these instruments served for us to answer the qualitative RQs defined for this experiment. In addition, we also created an equal working environment for all experiment subjects. To this end, a Virtual Machine (VM) was created using the Xubuntu OS, in this environment, the support materials previously mentioned were made available, as well as the ER modeling tools. To avoid possible external influences, the VMs did not have access to the Internet, thus ensuring that the subjects could not consult external content. Thus, it was possible to ensure that all subjects performed the same tasks, and under the same conditions.

4.2 Conduction

Preparation: Initially, we took place meetings among the researchers involved to define the planning and the mode of operation that should be adopted.

Seeking to capture a significant sample for the study object, we decided to contact the lecturer responsible for teaching two courses for different undergrad programs: Database (Software Engineering) and Database I (Computer Science) in the second semester from 2019. With the initial objectives aligned, the lecturer who collaborated made the dissemination of the profile questionnaires via learning management system (Moodle) to the participants.

After the elaboration of all the artifacts that would be used in the experiment, they were analyzed and validated jointly by the researchers involved in this study, and there it was still a need to adapt some instruments along the way regarding the suggestions and possible corrections necessary.

Execution: On the experiment day, the first activity carried out was a brief initial presentation, where we informed that the experiment was of an unavailable character. With that clarified, the FPIC was then made available to all subjects. After signing, we distributed the profile questionnaires to the subjects who have not yet been completed previously. We found that there were no strong discrepancies among the subjects' levels of knowledge, thus demonstrating that there was a homogeneous sample in general.

Before the random division of the groups, we carried out the training phase. During this phase, both database modeling tools that would be used were presented, providing an overview of operation and answering possible questions that might arise. Then, we began the modeling phase of the proposed problems. All subjects received Instrument 1 and were informed with which tool they should develop the solution. We asked for each subject to write down in the instrument their identification and the start time of the task. We no stipulated 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 collected and moved the instruments 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 then we had thanked and released them.

4.3 Results and Data Analysis

Effort: To answer **RQ1**, regarding the effort to use the approaches, 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 Shapiro-Wilk 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 $\alpha = 5\%$.

After calculations with the set of time differences, we reached a p -value of 0.606530. As p -value $> \alpha$, we accepted the null hypothesis, thus concluding that the data is normally distributed, *i.e.*, 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 60.65% (0.606530). Still in relation to the normality test, the calculated value of W was 0.970178, being within the accepted range of the confidence level of 95% (0.9242: 1.0000). This means that there is a 95% chance that the sample comes from a normal population.

Once we performed the normality tests on the sample, we carried out the hypothesis test of the average effort regarding to **RQ1**. In the paired T-test for dependent samples, we used a significance level of $\alpha = 5\%$, with which we reached a measure of 0.000962084 for the p -value. Because it is a two-tailed test, *i.e.* it includes equality in its null hypothesis, this p -value shows enough evidence to guarantee the rejection of the statement of $H_0 : \mu Time_G = \mu Time_T$. Therefore, we accepted the alternative hypothesis that the approaches have different efforts, once according to the test this difference is statistically significant. Figure 2 displays a box-plot with the variation of data observed through these data. Based on these data it was possible to verify that the graphical approach has an advantage on average.

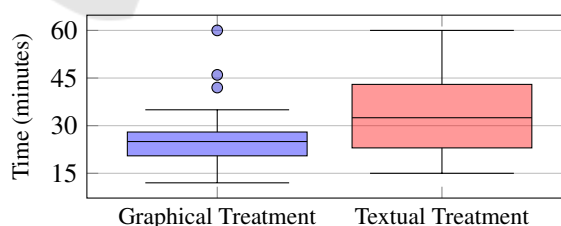


Figure 2: Box-plot - Effort per treatments.

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. F-Measure 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 fraction of recovered in-

stances that are relevant and Recall is the fraction of relevant instances that are recovered.

In addition, we performed the Shapiro-Wilk normality test to F-Measure for each model. After calculations with the set of differences in F-Measure for each model, we reached a p -value of 0.404455. With this test result, the chance of type 1 error (rejecting a null hypothesis that is correct) can be very high, and can be translated into 40.45% (0.404455). As the p -value $> \alpha$, we accepted the null hypothesis, thus realizing that the data is normally distributed, *i.e.* the difference between the data sample and the normal distribution is not large enough to be statistically significant. After the sample was tested for normality, we tested the second hypothesis defined in this experiment. This time, in the paired sample T-test, we used again a significance level of $\alpha = 5\%$, with which we reached a measure of 0.396468 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 not 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 accepted the null hypothesis that the approaches have equal effectiveness, because according to the statistical test, the average difference of F-Measure between treatments is not statistically significant. Table 1 shows average measures of the evaluated values, and also provides the possibility to carry out a dispersion analysis. Figure 3 box-plot graph displays of the F-Measure 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 not present much difference between the approaches.

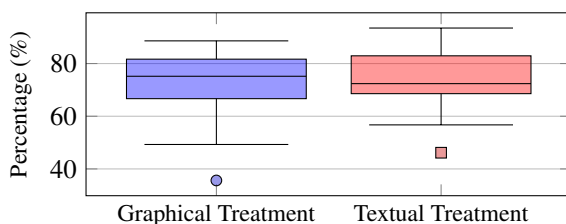


Figure 3: Box-plot - F-Measure per treatments.

Qualitative Evaluation: took place with the analysis of the two instruments applied after the modeling tasks. The first was used to respond to **RQ3**, regarding the Perceived Ease Of Use (PEOU) and Perceived Usefulness (PU) of treatments, according to the TAM model (Davis, 1989). 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. This scale served 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 seven quality attributes are grouped in three categories, being defined as follows:

Functionality: *Conformity:* ability level to which the software to achieve specified goals with functional completeness, correctness and appropriateness related to their functionalities.

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.

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 quality attribute, since in the hypothesis test related to the effort, the treatment using the brModelo demonstrated a lesser need for execution time.

In analyzing the open comments on this evaluation form, which asked subjects to indicate positive and negative points of the tools, it was explicitly reported that the code completion feature provided a sense of agility in the modeling database process.

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 and Descriptive Attributes were the best evaluated, with all 27 agreeing with their current representation. In contrast, all the other 6

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

Measure	Graphical Treatment					Textual Treatment				
	MI	RI	Precision(%)	Recall(%)	F-Measure(%)	MI	RI	Precision(%)	Recall(%)	F-Measure(%)
Maximum	53.00	35.00	96.00	89.74	88.61	67.00	43.00	95.56	97.50	93.48
3° Quartile	39.50	33.00	87.50	82.05	81.66	44.50	37.00	89.44	80.43	82.93
Median	36.00	29.00	82.93	71.74	77.11	38.00	31.00	83.78	75.00	72.46
Average	35.70	28.26	79.61	68.57	72.79	38.89	31.30	81.50	70.88	74.73
1° Quartile	32.00	25.00	73.80	59.03	67.10	32.50	26.00	73.30	60.85	69.72
Minimum	23.00	13.00	38.24	33.33	35.62	20.00	18.00	58.06	38.30	46.15
SD	6.33	5.63	11.94	15.39	12.16	9.97	7.30	10.44	15.36	10.94

Legend: SD = Standard Deviation; MI = Modeled Items; RI = Relevant Items.

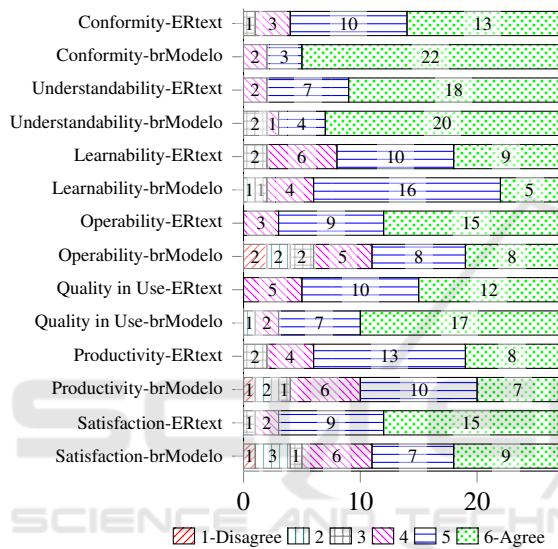


Figure 4: Quality attributes per treatments.

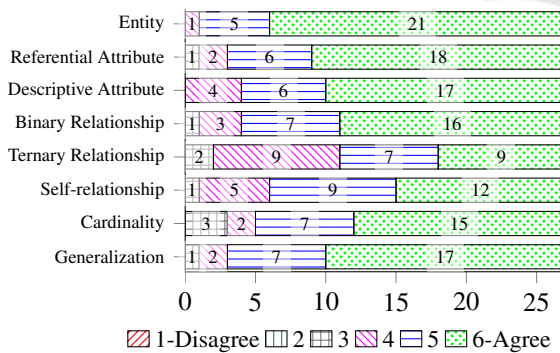


Figure 5: Evaluation of DSL designers.

obtained at least one disagreement. In this sense, the contributors of Ternary Relationship and Cardinality stand out, with 2 and 3 evaluations disagreeing with their current representations respectively.

5 THREATS TO VALIDITY

In this section we discuss the main threats to validity of our study and present the strategies we used to mitigate them (Cook and Campbell, 1979).

Construct Validity - Inappropriate Pre-operational Explanation: 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. **Interaction of Different Treatments:** We followed a paired design, all subjects are executed 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, *i.e.* low standard deviation indicates that the data points tend to be very close to the mean.

Internal Validity - History: To soothe this threat, we carried out the experiment in an academic environment, and because we conducted the entire process in August, when in general students are not necessarily overwhelmed with academic activities. **Maturation:** In order to alleviate this threat, we informed subjects from the beginning that they could terminate their participation at any time, without any penalty.

External Validity - Experiment Subjects: Seeking to mitigate this threat, the experiment was carried out with undergrad students of Software Engineering and Computer Science programs, and soon, inserted in the context of using the conceptual modeling of relational databases. However, the fact that the sample has less than 30 subjects is a statistical threat in the analyzed area, and it was not possible to mitigate this fact. **Subjects Interaction with the Evaluation Artifacts:** Depending on the moment this can affect the experimental results. For instance, if a questionnaire is answered a few days after the execution experiment, people tend to answer differently than they would do moments after the activities.

Conclusion Validity - Low Statistical Power: To try mitigating this threat, some statistical methods were adopted, such as the Shapiro-Wilk normality test, the paired T-test as a hypothesis test for dependent samples, and the F-Measure for qualitative analysis of the models produced. *Reliability of Measurements:* To mitigate this threat, it was adopted objective measurements that did not depend on subjective judgment (effort, measured in time spent, and F-Measure). On the other hand, the metrics used for the qualitative evaluation still served as a complementary input in the discussion of the results obtained. *Experimental Environment:* In order to mitigate this possible threat, we instructed subjects that conversations could not take place during the entire activities execution, or leave the environment or access electronic devices.

6 CONCLUSION

This study presented a controlled experiment evaluating ERText, a proposed textual DSL for database conceptual modeling. ERText is compared with the brModelo, a graphical DSL well-known in Software Engineering ER lectures.

From the analysis it is possible to highlight the following aspects: (i) Effort: the graphical approach to ER modeling requires less effort to perform the evaluated tasks. However, we considered that this difference is small and it can be reduced with future improvements in the proposed DSL. (ii) Effectiveness: the computed average difference states that there is no differences between the approaches, *i.e.*, one approach is not better than the other. However, we observed that there is a need to carry out tests involving problems of greater complexities for better assessment. (iii) Qualitative comparison between treatments: We observed a certain balance between treatments, but with a positive evaluation for ERtext regarding the “Productivity” attribute. Because it was the first time that the subjects had contact with our grammar, and also considering a first release of our DSL, we conclude that ERText is on the rails for achieving better productivity indexes.

We also collected qualitative feedback from participants. As a result, there are some improvements regarding the language design that need to be revised, in particular to the cardinalities and ternary relationships. From the experimental results we conclude that there is feasibility and good perspectives for the motivated context, *i.e.*, as a tool for teaching entity-relationship modeling with the differential of adopting a textual approach for conceptual database modeling in classrooms instead of a graphical notation.

This conclusion is sustained by the results, which did not obtained expressive differences with regard to the evaluated quality attributes between the tools.

ACKNOWLEDGEMENTS

This study was partially funded by PROPESQ through AGP, and by FAPERGS, through the ARD project N^o19/2551-0001268-3.

REFERENCES

- Al-Dmour, A. (2010). A cognitive apprenticeship based approach to teaching relational database analysis and design. *Inf. & Comp. Science*, 7(12):2495–2502.
- Brambilla, M., Cabot, J., and Wimmer, M. (2017). *Model-Driven Software Engineering in Practice, Second Edition*. Synthesis Lectures on Software Engineering. Morgan & Claypool, San Rafael, CA, USA.
- 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., Alarç, 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.
- Elmasri, R. and Navathe, S. (2011). *Sistemas de Banco de Dados*. Pearson Universidades, Franca, SP, Brasil.
- 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.
- Obeo and TypeFox (2020). Xtext/sirius - integration the main use-cases. Technical report, Obeo and TypeFox.
- Wohlin, C., Runeson, P., Hst, M., Ohlsson, M. C., Regnell, B., and Wessln, A. (2012). *Experimentation in Software Engineering*. Springer, London, England.