Quality Evaluation of Mobile Applications

Vitor Maia, Taisa G. Gonçalves, Ana Regina C. da Rocha

Federal University of Rio de Janeiro – PESC/COPPE, Rio de Janeiro, Brazil

{vmaia, taisa}@cos.ufrj.br, darocha.anaregina@gmail.com

Abstract. *Mobile applications (apps) hold a well-established market. These systems are impacted by context variations, usage of mobile device's sensors, usability and a plentiful of other features, which makes them different from other types of software. App stores provide quality guidelines with generic publishing recommendations, but these do not cover all apps' needs and usages. Quality models for quality assessment of software product are general-purpose and made to be applied to any kind of software product, although they may be particularized for specific contexts. This dissertation proposed a set of context-specific quality characteristics for apps, based on the particularization of quality models defined in international standards. This set was used to extend the ISO/IEC 25010 quality models. The identification was performed by conducting a systematic mapping and a survey with app users. Once the quality characteristics were identified, an evaluation procedure for apps was also proposed, adapted from QPS software evaluation model. The feasibility of using this procedure was verified through the evaluation of a mobile banking application, already on the market.*

Resumo. *Aplicações móveis (apps) possuem um mercado bem estabelecido. Estes sistemas são impactados por variações de contexto, pelo uso de sensores presentes nos dispositivos, pela usabilidade, além de outras particularidades que as diferenciam de outros tipos de software. Lojas de aplicativos possuem diretrizes de qualidade para publicação com recomendações genéricas, que não abrangem todas as necessidades e formas de uso. Modelos de qualidade para avaliação da qualidade em produtos de software são genéricos, apresentam características de qualidade tendo em vista qualquer tipo de aplicação. Porém, estes modelos podem ser particularizados para uso em contextos específicos. Este trabalho propõe um conjunto de características de qualidade específicas para o contexto de apps a partir da particularização dos modelos de qualidade definidos em normas internacionais. Este conjunto foi utilizado para estender os modelos de qualidade da ISO/IEC 25010. A identificação foi realizada através da condução de um mapeamento sistemático e de um survey com usuários de apps. Com o conjunto de características de qualidade identificado, é também proposto um procedimento de avaliação para apps, adaptado do modelo QPS de avaliação de produtos de software. A viabilidade de uso deste modelo foi verificada através da avaliação de uma aplicação móvel bancária, disponível no mercado.*

1. Introduction

Quality Models for evaluating software quality are general-purpose. However, different applications and stakeholders demand context-specific quality requirements. It becomes crucial to identify specific quality requirements for these types of application, as well as stakeholders with whom draw up quality characteristics, sub-characteristics, and evaluation procedures. Mobile applications (apps) hold a robust and established market, and they became popular since the creation of the app stores' distribution model [Cortimiglia et al. 2011]. Both the quantity of apps and the variety of their functionalities highly increased ever since. The biggest app stores are for Android and iOS systems. Both platforms provide publishing quality guidelines, but they are not broad enough to grant the final users' expectations and apps quality.

The Android app store guidelines list several quality criteria. Testing can be done through alpha and beta publications, visible to a private team of testers. However, the store does not check the quality criteria during publication phase. Applications may be published in minutes, even if they do not provide minimum quality requirements. The quality guidelines include recommendations for user interaction, functionality, compatibility, performance, security, publishing, and testing.[1] The iOS publishing procedure is more reliable. It provides an intermediary environment called TestFlight, which might be used for testing and homologation. There is an automatic quality check before deploying to TestFlight. Publishing to the iOS app store goes through a manual check by an Apple employee, who checks the correct implementation of quality guidelines for Safety, Performance, Business, Design and Legal Requirements.[2]

An initiative called App Quality Alliance (AQuA) proposed a more comprehensive set of guidelines. The guidelines are deeply concerned with Resource Utilization (e.g. care with battery life), Fault Tolerance (e.g. handling exceptions when connection in unavailable), Data Persistence (e.g. implement pause, suspend and resume capabilities), Functional Correctness (e.g. grant the correct implementation of calculations), Confidentiality (e.g. do not store sensitive data), among others.[3]

1.1. Related Work

The ISO/IEC 25000 family, also known as SQuaRE, contains a set of standards relative to software quality. These standards define management guidelines, quality models, measures, requirements, and software evaluation processes. ISO/IEC 25010 [ISO/IEC 2011] defines two quality models: **Quality in Use**, with quality characteristics related to the outcome of interaction when a product is used in a context of use; and **Product Quality**, with characteristics related to static and dynamic properties of software. This standard defines the quality of a system as the degree to which the system satisfies the stated and implied needs of its various stakeholders, and thus provides value. The growth of app market, in contrast to the seeming lack of comprehensive quality guidelines, makes room for new studies about context-specific quality concerning apps.

---

[1] https://developer.android.com/docs/quality-guidelines/core-app-quality

[2] https://developer.apple.com/app-store/review/guidelines/

[3] https://www.appqualityalliance.org/

Several studies propose sets of quality characteristics for apps: [Corral et al. 2014] uses several app store guidelines as basis to obtain a list of quality characteristics; [Springer and Spriestersbach 2004] lists common challenges of apps; and [Idri et al. 2016b] focus on the quality of pregnancy monitoring apps. Meanwhile, none of these studies presents a generic set of quality characteristics for apps.

This dissertation presents both, main and secondary objectives. The main objective is to identify a set of **essential** quality characteristics which should be considered when appraising apps. This set shall be used to adapt the ISO/IEC 25010 quality models for the context of apps. The secondary objective is to apply the adapted quality models in the development of an appraisal procedure.

1.2. Methodology

Figure 1 describes the methodology, divided in eight steps. The appraisal procedure cited in steps seven and eight is based on the QPS (*Qualidade de Produto de Software*) [Rocha et al. 2017], Brazilian reference model, for appraising software products.

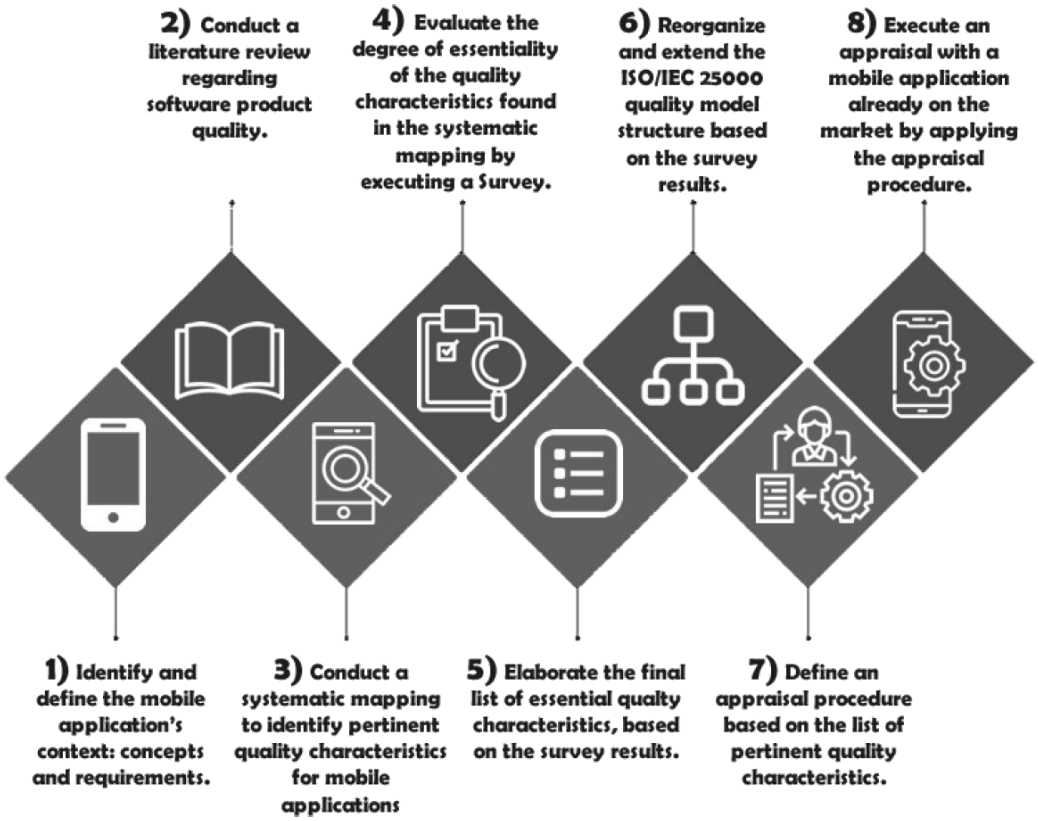

Figure 1. Methodology

2. Systematic Mapping

A systematic mapping study [Maia et al. 2020] was conducted in order to identify which quality characteristics are most pertinent in the context of apps. Literature was analyzed in search of occurrences of ISO/IEC 25010 quality characteristics and sub-characteristics, and in search of attributes not previewed by ISO/IEC 25010. The number of occurrences was a deciding factor to define if a certain characteristic is

pertinent or not. Notice that the mapping intends to identify pertinent quality characteristics, even though the main objective of the dissertation is to propose a set of essential quality characteristics. The adjectives **pertinent** and **essential** are not treated as synonyms. The pertinent quality characteristics are those somehow related to the context of mobile applications, but not necessarily essential. Afterwards, a survey was conducted to identify which pertinent quality characteristics are also essential.

2.1. Planning

The goal of the systematic mapping is defined using the GQM paradigm [Basili et al. 1994]: **Analyze** quality characteristics **for the purpose of** characterizing **with respect to** pertinence **from the point of view of** software engineering researchers **in the context of** mobile applications.

ISO/IEC 25010 contains two quality models, each with a hierarchy of characteristics and sub-characteristics. At first, we did not intend to bind the analysis of the results to these models, because we were not sure if the papers in the literature would follow the standard's definitions. Furthermore, we expected to find attributes beyond those present in the standards. Surprisingly, the preliminary execution of the search string returned several papers about the quality of apps that indeed used ISO/IEC 9126 [ISO/IEC 2001] or ISO/IEC 25010, even though the search string did not explicitly include them. Due to this observation, two research questions were developed:

- **RQ1**: Which quality characteristics from ISO/IEC 25010 are identified as pertinent in the context of mobile applications?
- **RQ2**: Which quality characteristics are not present in ISO/IEC 25010, but are identified as pertinent in the context of mobile applications?

The selection of papers was carried out through a search string, applied to five search engines: Scopus, IEEE, Web of Science, Engineering Village and ACM. The string followed the PICO process [Pai et al. 2004]. Our population are mobile applications, our intervention are software quality models, quality in use and quality requirements, and our outcome are characteristics, metrics, measures, evaluation criteria and attributes. The final string was **((“mobile app*”) AND (“software quality” OR “quality model” OR “quality in use” OR “quality requirement”) AND (“characteristic” OR “metric” OR “measure” OR “evaluation criteria” OR “attribute”))**. The execution of the search string in the search engines returned several papers, but not all of them might be suitable for the purpose of this study. Therefore, inclusion and exclusion criteria were developed to refine the initial list of papers. Exclusion criteria are presented in Table 1. Inclusion criteria are the opposite of exclusion criteria E1-E4.

Table 1. Exclusion Criteria

E1	Was not published in conferences or journals.
E2	Is not available in the internet.
E3	Is not in English.
E4	Is not about quality in mobile applications.
E5	Is nearly equal to another paper by the same authors.
E6	Is about design patterns, source code or quality of services.

2.2. Execution

The exclusion criteria E1, E2 and E3 were immediately applied to the initial set of papers. The application of the other exclusion criteria depended on further interpretation, so the authors read the abstracts of the remaining papers, then decided together on more exclusions, based on criteria E4, E5 and E6. Some papers could only be safely excluded after the full text was read. Table 2 shows the number of results before and after the application of exclusion criteria. In total, 35 papers were selected using the search string and 18 using snowballing. At all, 53 papers were selected.

Table 2. Search Results

Search Engine	Initial Set	After E1, E2 & E3	After E4, E5 & E6
Scopus	57	45	29
IEEE	27	26	13
Engineering Village	27	19	13
Web of Science	30	23	13
ACM	7	7	3

2.3. Analysis

The selected papers were analyzed by means of the research questions. The extracted data were collected with the assistance of a form. To answer RQ1, the form contained tables to store information on quality characteristics and sub-characteristics of both quality models from ISO/IEC 25010. To answer RQ2, the form contained a table to retain information on characteristics apparently not related to those proposed by ISO/IEC 25010. RQ1 and RQ2 were answered by following the processes described in Figure 2. Part of the processes was based on the methodology present in [Marinho and Resende 2012a]. By analyzing RQ2, nine quality attributes were identified: **Information Quality**, **Navigation**, **Interface Visibility**, **Use of Clear and Minimized Forms**, **Use of Hierarchical Menus**, **Data Persistence**, **Sense of Community**, **Continuous Communication** and **Memorability**. They will be referred just as **attributes** to distinguish from the quality characteristics and sub-characteristics in ISO/IEC 25010.

3. Survey

A survey [Maia et al. 2019] with mobile users was conducted to confirm some of the systematic mapping results. Literature papers were analyzed in search of occurrences of ISO/IEC 25010 quality characteristics and sub-characteristics, and in search of occurrences of additional attributes. The **number of occurrences** was used as a deciding factor to define if a certain sub-characteristic would be considered **essential**.

Due to the generality of characteristics, only sub-characteristics were taken into consideration. Three intervals were defined based on the range of occurrences, which were comprised between zero and fifteen. The sub-characteristics with **less than five occurrences** were dismissed as not essential, and those with **eight or more occurrences** were directly considered essential. The remaining ones, with an **intermediary number of occurrences (five to seven)**, were added to the survey for further investigation. These sub-characteristics were **Functional Correctness**, **Interoperability**, **Appropriateness**

Recognisability, **User Error Protection**, **User Interface Aesthetics**, **Accessibility**, **Fault Tolerance**, **Confidentiality**, **Usefulness** and **Pleasure**. The attributes identified with RQ2 were directly added to the survey regardless of their number of occurrences.

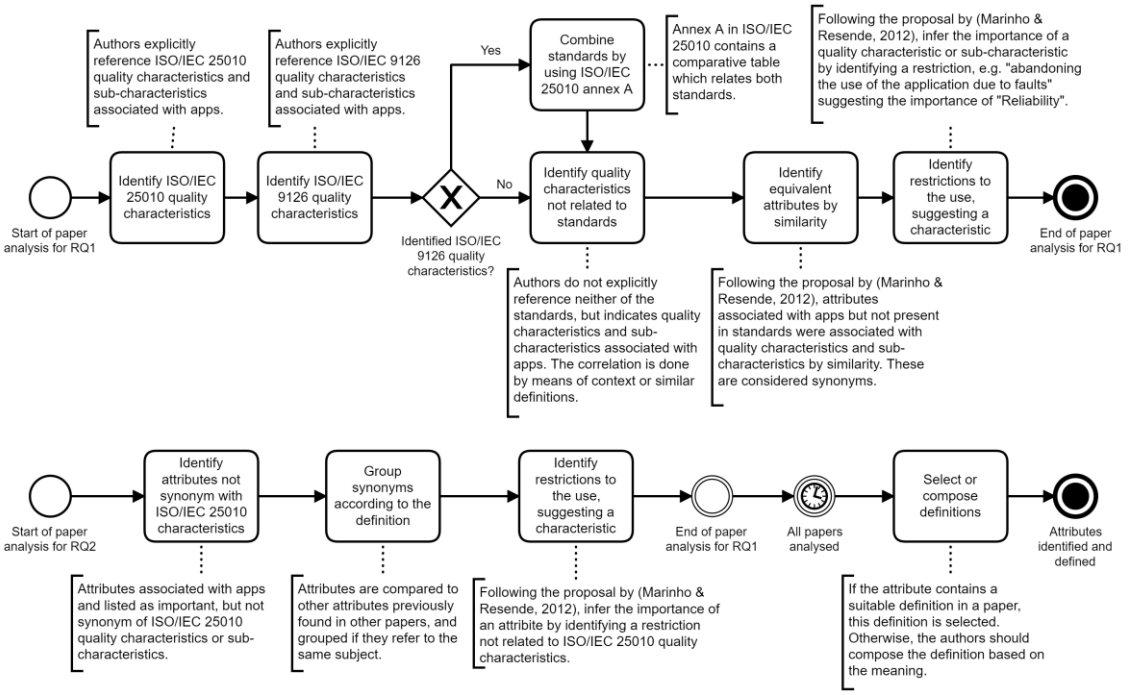

Figure 2. Processes for answering the Research Questions

The respondents should opine over the importance of a set of 19 quality characteristics in the context of apps. The elaboration of the questions' titles was tricky because they could not be neither the name of the characteristic nor its technical definition. The simpler it was, the more answers the survey would obtain, so the chosen strategy was to develop a small and informal questions.

The survey was composed of two parts. The first part collected demographic data. The second part collected substantive data. One question inquired about the respondent's favorite app category, given five possibilities: mobility, food delivery, tourism, messaging, and banking. The respondents were guided to answer questions considering only the specific selected app category. Each question contained a title, a VAS [Wewers and Lowe 1990] and a multiple-choice with two options: 1) if the respondent did not know how to answer or 2) were not sure if the question was related to the selected app category.

3.1. Pilot Test and Execution

The survey was sent to five people as part of a pilot test. The age frequency was well-balanced. Every respondent selected the messaging category. Their answers in the VAS questions were well-balanced and made sense as a whole. Their average responding time was of five minutes. Once prepared, the survey request was shared in social networks. Except for the minimum age limit of 18, the survey could be answered by any mobile user. It was carried out for 9 days and had 500 answers. For every answer, we

donated R$1.00 to Pro Criança Cardiaca[4], a Brazilian project which assists children with heart diseases. This initiative not only helped the institution, but also created empathy for the survey.

3.2. Data Analysis

Regarding the characterization questions: **Gender**: 58.4% were female, 41.6% were male. **Level of Education**: 74.6% finished college, 13.8% did not finish college, 6.8% finished high school, 2.2% did not finish high school, 2.2% did finish primary education and 0.4% did not finish primary school. **Age**: 29% aged between 18 and 30, 27.8% between 31 and 40, 14.6% between 41 and 50, 17.4% between 51 and 60, and 11.2% aged above 60. Regarding the **favorite app category** question, 70.2% selected messaging apps, expressively higher than the other categories, presumably due to the popularity of WhatsApp in Brazil[5].

Five different results derived from the VAS questions, one for each app category. For each answer, respondents indirectly selected a value between zero and ten in the VAS, and averages were calculated. Independently of the app type or number of answers, **Interoperability** and **Sense of Community** were evaluated with overall averages of 4.73 and 5.36 respectively. **User Error Protection** and **Fault Tolerance** had the greatest averages in the banking category. Even though messaging apps do not have forms, the average for **Use of Clear and Minimized Forms** was considerably high for this category. **User Interface Aesthetics** had the second worst average in ISO/IEC 25010, while **Functional Correctness** had the best average.

The resulting average of every question for every app category altogether was 8.3. It was decided to use the rounded value 8.0 as a cutoff score, to define what would be essential or not given the results of the survey. By applying this decision, Interoperability, Sense of Community, Memorability and User Interface Aesthetics were dismissed and **not considered essential**. The remaining sub-characteristics and attributes, along with the sub-characteristics with more than eight occurrences in the systematic mapping, comprises the final list of essential quality characteristics for mobile applications. The final list is presented in Table 3 and Table 4.

4. Quality evaluation of Mobile Applications

The results obtained by the systematic mapping and the survey should now be used in a real software evaluation to verify its applicability. QPS is a Brazilian reference model used for evaluating software products. It assesses software products considering four dimensions: organizational, software engineering, services, and product quality. It delivers results in a three-level ranking system: gold, silver, and bronze. An appraisal session starts with a diagnostic analysis of the product as a whole, whereby the product is assessed considering gold level requirements. Based on the diagnostic analysis results, the company in charge of the product determines the **final assessment** level, which is held after a period of adjustments. Considering that at gold level the model evaluates

[4] http://www.procrianca.org.br/

[5] https://www.messengerpeople.com/pt-br/whatsapp-no-brasil/, visited in May 2020

context-specific quality characteristics, it was decided to use QPS as a basis for assessing the quality of apps.

Table 3. Final list of essential sub-characteristics

Quality Sub-Characteristics
Functional Correctness
Time Behaviour
Resources Utilisation
Learnability
Operability
Appropriateness Recognisability
User Error Protection
Accessibility
Fault Tolerance
Confidentiality
Adaptability
Usefulness
Pleasure
Effectiveness
Efficiency
Context Completeness
Flexibility

Table 4. Final list of essential attributes

Attributes
Navigation
Interface Visibility
Use of Clear and Minimized Forms
Use of Hierarchical Menus
Information Quality
Data Persistence
Continuous Communication

Mobile applications may be appraised in two cases: 1) the appraisal is commissioned by the company in charge of the product. 2) The appraisal is based on an app available in an app store, where no documentation is available. In this setting, only the product quality dimension can be appraised. Considering that the company in charge of the product does not intervene with the appraisal process and consequently does not perform adjustments, only the **final assessment** is conducted. A list of quality sub-characteristics for appraising apps in the product quality dimension was drawn up, adding up quality sub-characteristics already present in QPS for any software product and the sub-characteristics assembled by the end of the survey. The apps are evaluated by means of questions related to each quality sub-characteristic.

An appraisal was conducted with a Brazilian mobile banking application, to illustrate the feasibility of the proposal. This app is available in the app store and the company was not involved in the process. Consequently, this describes the second appraisal case. Two certified QPS evaluators were selected to make up the evaluation team and one of them was assigned as the leader appraiser. The appraisal planning was conducted by the local coordinator and by the leader appraiser. The first activity was the kickoff meeting, in which the local coordinator explained the evaluation instructions and the questions to the appraisal team. As soon as they stated to have understood the questions and procedure, the appraisal started. It lasted two hours. The level assignment rules were then applied. The banking application was awarded the **bronze** level. By the end, the lead appraiser produced the evaluation report. The evaluation satisfactorily showed the possibility of using the proposal of this dissertation as a basis to conduct a mobile application's appraisal.

5. Limitations

The systematic mapping is limited to its threats to validity: 1) some papers provided differing definitions from ISO/IEC 25010, which demanded a deep interpretation of what the authors meant. 2) The search string may not have captured several relevant papers. In addition, the quantity of papers may have been relatively small, possibly due to a too restrictive search string. 3) The interpretation of the abstracts and the manual selection of papers via snowballing may have biased the results. 4) The research questions were subjective and demanded effort to both understand what to be extracted and how to interpret the extracted data.

The survey is also limited to its threats to validity: 1) the survey does not contain hypothesis and was developed based on the goal of the study itself. 2) The answers permitted the inferring of some results, but none of them can be proven due to the much greater size of the population in comparison with the sample size. 3) The VAS questions were developed as simple as possible, yet some of them might have been too subjective. 4) Some respondents might have misunderstood the meaning of what was being inquired. Interoperability, for example, presented unexpected results.

Some decisions were made by the authors during the development of this work, and they might have biased the results: 1) not including quality characteristics in the survey. 2) Only adding the quality sub-characteristics with occurrences between five and seven in the survey. 3) Using mean 8.0 as a cutoff score, given the results of the survey.

6. Conclusion

Mobile applications are different from other software products, given its unique features and hardware. The aim of the systematic mapping was to identify quality characteristics pertinent in the context of mobile applications, both those previewed in ISO/IEC 25010 and those not previewed. The amount of returned studies was considerably large, and the themes of each study was quite diverse. However, it only allowed the ranking of quality characteristics by occurrences. A survey was proven necessary in order to question mobile users about their opinion over a set of sub-characteristics. With the result, it was possible to compose an adapted version of ISO/IEC 25010 quality models.

An evaluation procedure was defined, based on QPS model and on the adapted quality models. The evaluation of the mobile banking application confirmed the feasibility of evaluating mobile applications by using the previously adapted quality models. The evaluation was quite simple, and the results were reasonable for what was intended, indicating that the appraisal of mobile applications can be conducted by using the identified characteristics and the provided appraisal procedure.

Acknowledgments. We acknowledge Guilherme Horta Travassos and Gleison Santos for their important comments and suggestions; the QPS evaluators Elaine Nunes and Jhonatan Boarim for participating in the quality evaluation; and all the respondents of the survey.

References

Basili, V. R., Caldiera, G. and Rombach, D. H. (1994). The Goal Question Metric Approach. Encyclopedia of Software Engineering. John Wiley & Sons. v. Ip. 528–532.

Corral, L., Sillitti, A. and Succi, G. (2014). Defining Relevant Software Quality Characteristics from Publishing Policies of Mobile App Stores. Mobile Web Information Systems. Cham: Springer International Publishing. v. 8640p. 205–217.

Cortimiglia, M., Ghezzi, A. and Renga, F. (2011). Mobile Applications and Their Delivery Platforms. IT Professional, v. vol 13, issue 5, p. 51–56.

Idri, A., Bachiri, M. and Fernández-Alemán, J. L. (2016a). A Framework for Evaluating the Software Product Quality of Pregnancy Monitoring Mobile Personal Health Records. Journal of Medical Systems, v. vol 40, issue 3, p. paper 50.

ISO/IEC (2001). Software engineering – Product quality – Part 1: Quality model - 9126-1. Geneva, Switzerland, 2001.

ISO/IEC (2011). Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models - 25010. Geneva, Switzerland, 2011.

Maia, V., Gonçalves, T. G. and Da Rocha, A. R. C. (2019). Quality Characteristics of Mobile Applications: A Survey in Brazilian Context. In Proceedings of the XVIII Brazilian Symposium on Software Quality (SBQS'19).

Maia, V., Gonçalves, T. G. and Da Rocha, A. R. C. (2020). Identification of quality characteristics in mobile applications. In Proceedings of the XXIII Iberoamerican Conference on Software Engineering (CIbSE'20) (to appear).

Marinho, E. H. and Resende, R. F. (2012b). Quality Factors in Development Best Practices for Mobile Applications. Computational Science and Its Applications – ICCSA 2012. Berlin, Heidelberg: Springer Berlin Heidelberg. v. 7336p. 632–645.

Pai, M., McCulloch, M., Gorman, J. D., et al. (2004). Systematic reviews and meta-analyses: an illustrated, step-by-step guide. The National Medical Journal of India, v. vol 17, issue 2, p. 86–95.

Rocha, A., Travassos, G., Santos, G. and Reinehr, S. (2017). QPS-Modelo para Avaliação da Qualidade de Produtos de Software: Resultados Iniciais. In Proceedings of the Brazilian Software Quality Symposium (SBQS'17).

Springer, T. and Spriestersbach, A. (2004). Quality Attributes in Mobile Web Application Development. Product Focused Software Process Improvement. Berlin, Heidelberg: Springer Berlin Heidelberg. v. 3009p. 10.

Wewers, M. E. and Lowe, N. K. (aug 1990). A critical review of visual analogue scales in the measurement of clinical phenomena. Research in Nursing & Health, v. 13, n. 4, p. 227–236.