International Journal of Cooperative Information Systems
Vol. 22, No. 1 (2013) 1350003 (35 pages)

DOI: 10.1142/S0218843013500032

TOWARD A QUALITY FRAMEWORK FOR BUSINESS PROCESS MODELS

LAURA SÁNCHEZ-GONZÁLEZ*, FÉLIX GARCÍA†,
FRANCISCO RUIZ‡ and MARIO PIATTINI§

Alarcos Research Group, TSI Department
University of Castilla La Mancha, España
**Laura.Sanchez@uclm.es*
†Felix.Garcia@uclm.es
‡Francisco.RuizG@uclm.es
§Mario.Piattini@uclm.es

Received 14 January 2012
Accepted 8 July 2012
Published 15 April 2013

Business process modeling is recognized as a key part of the business process lifecycle. It is during this stage that a conceptual model is produced by collecting business process requirements and representing them with a specific business process notation. While there has been much research into process modeling techniques, little has taken place with regard to the characteristics that should be considered for an effective assessment of the models' quality. This paper presents a synthesis of quality characteristics for business process models, based on a systematic review of the relevant literature. It then goes on to describe a reference model for the quality assessment of business process models, and to relate the aforementioned quality characteristics to existing relevant process model measures. These relations may help organizations to guide the improvement of their business process models according to their chosen quality characteristics.

Keywords: Quality framework; business process model; measurement.

1. Introduction

Business process (BP) modeling is a key activity in the BP lifecycle.[1] This activity produces blueprints of organizational issues which are called process models or conceptual models and can be used to make decisions about where, how, and why changes to the processes should be enacted to warrant improved operational efficiency.[2] In addition, it is known that more than half the errors that occur during process developments are requirements errors[3]: post-implementation errors cost a 100 times more than errors which occur during the design and analysis stage.[4] A poor-quality conceptual model may increase the development effort (as a consequence of detecting and correcting defects) or result in a system that does not satisfy users (as a consequence of not detecting or not correcting defects).[5]

A process model is a cognitive design tool to explore possible connections between people and machines.[6] Since the main purpose of process models is to enable communication between stakeholders (IT experts versus business analysts), there is no consensus about what makes a good model. No standard or reference document has yet been published; therefore, researchers attempt to define a group of characteristics which are potentially useful, based on their experience. This research implies some difficulties, owing to the fact that conceptual models are products of the mind or ideas embodied in models, so quality assessment is recognized as being a complex task. In addition, BP models are not only instruments that collect business requirements but have also recently been used as an important part of system development. For example, in model-driven engineering,[7] models are used as instruments with which to construct software systems because of their automatic code generation. More specifically, when applying model-driven architecture, BP models can be used as computation independent model.[8] The way in which process models are used in new paradigms leads to the need for a group of quality characteristics, in order to cover all their perspectives.

The evaluation of BP conceptual models by means of suitable measures is fundamental for us to know to what extent the model satisfies a specific quality attribute. Since measurement provides objective information about quality, some authors have attempted to associate quality characteristics with measures through an empirical analysis.[9–12] However, authors have typically focused on certain specific characteristics but not within the scope of a quality reference model. The lack of measurement associated with quality characteristics is a major weakness in the literature related to conceptual model quality.[5] Finally, all of the lacks that have just been pointed out are a sign of an immature field.

In this work, we have tackled the limitations mentioned by providing the following contributions:

(i) We have carried out a systematic literature review (SLR) aimed at analyzing relevant proposals concerning quality characteristics for BP models. In this case, quality of models is analyzed from a general perspective, regardless of the model notation and the context, which could be internal documentation, staff planning, automation, etc. The significant findings obtained are used to propose a reference quality model for BP models which includes a group of relevant external quality characteristics (EQC) found in the literature, along with others from international standards.

(ii) We have described a relationship between measures and external quality, to enhance the practical utility of quality models. We also introduce an algorithm to help in the empirical validation of measures and we illustrate its usage by means of an example, which focuses on measures for understandability and modifiability. The other EQC are not comprehensively addressed in this paper, but we have begun to tackle the issue in an approach which can serve as a starting point for future work.

To sum up, the main contribution of this paper is to collect a group of quality characteristics for BP models and their relationship with measures (structural measures, mainly) as a first step toward the definition of a complete quality framework for BP models. The remainder of this paper is organized as follows: Section 2 includes the background of this research, principally the quality of conceptual models and measures for BP models (details are in the appendix section). Section 3 describes the SLR carried out to obtain insights into the quality characteristics that have been researched for BP models. The papers selected through the application of a SLR are set out in the appendix. The results obtained from the SLR are used to propose a quality model for BP models. Section 4 presents the set of supporting measures collected from related literature which can be used to assess the quality characteristics of the proposed quality model. This section also provides an analysis of the influence of internal quality on external quality, along with a validation of a group of measures through a meta-analysis (other details in Appendix B). Then, in Sec. 5, a practical example is described to illustrate the application of the quality model and measures. Finally, our conclusions and future work are presented.

2. Background

This section discusses the background of our research. First, we introduce quality for process models and second, we summarize different measures for process models, as found in literature.

2.1. *Quality for process models*

Several proposals concerning quality characteristics exist in the Software Engineering field. Software quality is described in international standards such as ISO 9126[13] or its subsequent version, ISO 25010.[14] These standards define a *software product quality model* that is composed of eight characteristics and a *system quality in use model* composed of three characteristics. However, these international standards are not only used for software product quality but also for conceptual model quality. This is due to the fact that a conceptual model could be treated as a piece of software, owing to the similarities between them. Osterweil[15] states that, "software processes are software too", and this can be extrapolated to BP models as they can be managed as software artifacts. Since some authors argue that models are a kind of software artifacts, several proposals concerning the quality characteristics for conceptual models of processes are also based on these standards.[5,16–18] Some authors studied the literature about quality models in the Software Engineering field.[19,20] Moody[5] stressed that a conceptual model is simply a particular type of product and that quality characteristics should be consistent with ISO 9126, but he did not propose any quality characteristics. On the other hand, Qi *et al.*[18] proposed a quality model whose classification included syntactical, semantic, and pragmatic dimensions in the same level as reliability and maintainability; this does not appear to be very accurate. Other authors, Gucegliogl u and Demirors,[17] have associated quality characteristics with measures from the theoretical point of view. In our opinion, these problems indicate a lack of consensus as regards which quality characteristics are the most suitable, and there is neither a standardized classification of these quality characteristics nor a mapping between the quality characteristics proposed and appropriate measures with which to assess them. A thorough analysis of the literature about quality characteristics are detailed in Sec. 3.

2.2. *Measures for BP models*

A SLR concerning BP model measures was published in Ref. 21 and updated in Ref. 12. The most important aspect is that the measures are supported by some kind of empirical validation, which makes them reliable and facilitates the establishment of a more objective relationship between internal and external quality. If a particular quality characteristic was not supported by a measure with empirical validation, then representative measures were selected. The measures chosen are detailed in Table A.2. They can be used to measure or to predict some EQC, as is presented in Sec. 4.2.

3. Quality Characteristics for BP Models

This section tackles the relevant research concerning quality models for processes at the conceptual level. A systematic review was carried out for this purpose. The following subsections provide a description of the review process, along with the results obtained in the proposed approach, which homogenizes existing proposals and includes a selection of relevant quality characteristics that should be considered at the conceptual stage.

3.1. *Review process and the selected literature*

The identification and definition of quality characteristics for BP models is a no trivial matter, owing to the fact that quality is a fairly subjective term. Since there is a clear interest in approaching a definition of these characteristics, we have studied the published literature in a systematic manner. This was done by carrying out a SLR, following the recommendations in Ref. 22. The use of this methodology allows the most relevant studies with regard to a specific topic to be obtained in an unbiased manner. Table 1 summarizes the steps followed in the SLR to extract the relevant studies.

The execution of the steps described in Table 1 led to the extraction of relevant studies from literature (Table 2). The main threats to the validity of the review results are publication bias, undetected studies for keywords, and uncovered publication channels. We attempted to minimize the number of papers not included

Table 1. Systematic review steps.

Systematic Review Steps	Description
Question formularization	What initiatives concerning process model quality have been published to date?
Search string	("quality") AND ("process model" OR "conceptual model").
Source selection	(a) Science Direct, (b) Wiley InterScience, (c) ACM Digital Library, and (d) Scopus, on Computer Science. Proceedings of important conferences: (*International Conference on Software Engineering*, *International Conference on Conceptual Modeling*, etc.) and important journals such as *Journal of Systems and Software*, *Software Quality Journal*, etc.
Inclusion criteria	The inclusion criteria are based on the concept of taking into consideration only those studies whose principal subject includes quality in conceptual models and the specification of characteristics or dimensions of quality.
Exclusion criteria	The exclusion criteria mark out those studies which do not include quality characteristics for conceptual models (e.g. they are focused on data or information quality). Or it may be that the approach does not specify the characteristics or dimensions in detail.

Table 2. Distribution of the studies by source, selected on March 8, 2011.

Search Engine	Discovered	Relevant	Relevant not Repeated	Primaries
Science Direct	1,000	2	2	2
ACM	337	3	3	3
Wiley InterScience	122	0	0	0
Scopus	1,055	9	8	9
Results	**590**	**14**	**13**	**14**

Table 3. Questions and motivations of the SLR.

Question	Main Motivation
Q1. What are the quality characteristics of BP models?	To discover what are the most widely used quality characteristics specifically for BP models.
Q2. How are these quality characteristics classified?	To classify the selected quality characteristics appropriately.
Q3. What previous research pieces are the quality characteristics based on?	To discover what the reference documents are on which the selected quality characteristics are based.
Q4. Is any empirical validation included in the research?	To discover if the quality characteristics have been previously validated, to check if they are useful in practice.
Q5. Are the quality characteristics supported by measures?	To discover if the quality characteristics can be measured or predicted by measures.

by analyzing the works referenced in selected articles, but no work satisfied the inclusion criteria.

In order to extract the relevant information from the primary studies selected, the following questions were posed in Table 3.

The first question indicates the list of characteristics or attributes which authors have used to measure quality in process models. The second question tackles the quality dimensions used to classify the quality characteristics. For example, a list of characteristics can be classified as characteristics, sub-characteristics, and measures. The third question indicates whether the proposal is based on any international standards or previously published research works. The last question concerns the existence of supporting measures for quality characteristics. About the selected studies, the quality characteristics addressed in each work and other details are enumerated in Table A.1.

The results and discussion of the SLR are the following. We collected 135 quality characteristics from the papers chosen, 65 of which were not repeated. The characteristics which were tackled in more than one work are shown below.

The characteristics shown in Table 4 are, therefore, those most widely included in the quality models that were the scope of the SLR. In the case of comprehensibility/understandability, both terms are used to designate the same concept. The same is true of the terms modifiability and changeability or attractiveness and user interface aesthetics. This shows that it is necessary to homogenize quality characteristics ((1) *no homogenization of terms*).

However, the list of relevant quality characteristics is not the only important point; how to classify such characteristics is also a crucial aspect. Some of the selected proposals are based on the Lindland *et al.*[23] classification, which specifies syntactical, semantic, and pragmatic dimensions. On occasions, authors use the same classification but with adaptations. An example of this is the proposal of Qi *et al.*[18] which extends the three categories (syntax, semantic, and pragmatic) with others (maintainability, reliability, usage, and social). However, these seven categories seem not to be at the same level of syntax, semantics, and pragmatics, but perhaps as subcategories of them. Similarly, Rittgen's proposal[24] adds the social quality dimension.

Table 4. Number of repetitions of the quality characteristics.

Characteristic	Number of Proposals Which Include this Characteristic	Characteristic	Number of Proposals Which Include this Characteristic
Completeness	9	Suitability	3
Comprehensibility/ understandability	7	Security	3
		Reusability	3
Consistency	5	Accuracy	3
Reliability	4	Precision	2
Correctness	4	Safety	2
Changeability/ modifiability	3	Relevance	2
		Testability	2
Clarity	3	Stability	2
Generality	3	Learnability	2
Maturity	3	Effectiveness	2
Operability	3		

On the other hand, Mohagheghi *et al.*[25] indicate six quality goals, while Becker *et al.*[26] classify quality characteristics in six general guidelines. Satpathy *et al.*[16] organize the quality model into factors and subfactors, while Mehmood and Cherfi[27] classify it in dimensions, attributes, and metrics. Teeuw and van den Berg[28] do not indicate any classification, and Heravizadeh *et al.*[29] include quality characteristics in function, input/output, nonhuman resources and human resources. Gucegliogiu and Demirors[17] provide a classification that is composed of categories, characteristics, and sub-characteristics. Other authors do not specify any classification[30,31] ((2) *no homogenization of classification*).

On the other hand, in those studies selected which relate quality characteristics to measures, no references were found as to whether the measures chosen had undergone some kind of empirical validation. Moreover, the fact that a measure can be used to quantify several quality characteristics was not considered. Other proposals did not specify any measure for quality characteristics (4 of the 12 selected), which indicates that the relationship between quality characteristics and measurement is still immature ((3) *no empirical validation of measures related to quality characteristics*).

For all of the above reasons, this article proposes an attempt to minimize these disadvantages through a proposal of quality characteristics for BP models which considers a selection of some relevant quality characteristics extracted from this SLR and classifies them according to ISO/IEC 25010, System and Software Quality Models.[14] The classified quality characteristics are, therefore, related to measures with an empirical base.

3.2. *Quality characteristics for BP models*

This section addresses the issues set out above by describing a proposal which includes the relevant quality characteristics that were extracted from international standards and from relevant papers as a result of the SLR. The classification of quality characteristics is carried out by considering the ISO/IEC 25010 standard, which enhances the previous version ISO 9126.[13] ISO/IEC 25010 includes quality characteristics related to software system and data, along with the impact the system has on its stakeholders. This quality model for software systems has been applied and adapted to BPs owing to the similarities between business and software process models. For example, a "software product" logically matches a "business process", and a "function" of a software product matches the "activity" of BP models.

From our point of view, quality can be seen from two different dimensions: internal and external quality. Internal quality is about internal characteristics, which are based on inspecting the static properties of process models.[14] External quality considers quality characteristics from an external viewpoint, signifying the impact of models on users. Both dimensions are related through cognitive complexity, as was indicated in Ref. 32: "cognitive complexity is the mental burden of the persons

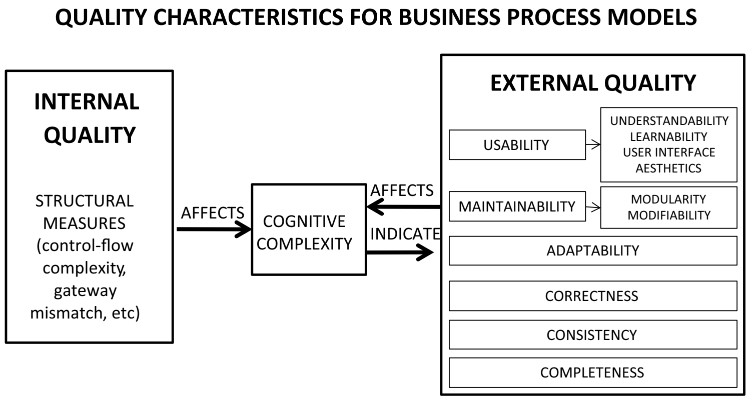

Fig. 1. Quality characteristics for BP models and their different parts, adapted from Ref. 32.

who have to deal with the model, so high cognitive complexity of a model causes it to display undesirable external qualities. The EQC are indicators of the cognitive complexity".

Figure 1 presents an overview of the proposed quality model, which is detailed in the following subsections, by indicating the quality characteristics for each dimension and the corresponding measures used to assess them.

3.2.1. *External quality*

External quality depends on the users' impact. External quality measures provide a black box view of the model and address properties related to the impact on humans. This is the most controversial part of the general quality model, because of the lack of consensus in the research community. In this case, some quality characteristics from international standards and other proposals have been selected. The selection criterion is based on the idea of adapting software quality characteristics to BP models. The quality characteristics are shown in Table 5.

The quality characteristics shown in Table 5 attempt to provide a general definition of quality for BP models that are organized by characteristics and sub-characteristics, which is the classification most frequently agreed on in the research community and international standards. The sub-characteristic *understandability* was adopted from ISO 9126, owing to the fact that the equivalent term in ISO 25010 is *appropriateness recognizability*, which we believe to be less intuitive. This characteristic is used in 54% of the proposals found through the SLR, so it is considered to be important in a quality model evaluation. Similarly, the quality characteristic of *user interface aesthetic* was used in ISO 25010, and it is called *attractiveness* in ISO 9126. In this case, we believe that the term *user interface aesthetic* reflects the concept we want to represent better. In the case of *modifiability*, it was used on both standards (ISO 9126 and ISO 25010) and in it 23% of the

Table 5. EQC for BP models.

Characteristic/ Sub-characteristic	Definition
Usability: Degree to which a model can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use	
Understandability	Attributes of models that have a bearing on the users' effort to recognize the logical concept and its applicability
Learnability	Degree to which a model can be used by specified users to achieve specified goals when learning to use it
User interface aesthetics	Degree to which the model provides the user with a pleasing and satisfying interaction
Maintainability: Degree of effectiveness and efficiency with which a model can be modified by the intended maintainers	
Modularity	Degree to which a BP design is composed of discrete models, such that a change to one model has a minimal impact on another
Modifiability	Degree to which a model can be effectively and efficiently modified without introducing defects or degrading existing product quality
Adaptability: Degree of effectiveness and efficiency with which a model can be adapted from one notation to another	
Correctness: Degree to which a model does not have workflow errors or faults, such as deadlocks	
Completeness: Degree to which a model has all the necessary, relevant information	
Consistency: Degree to which a process and the subprocesses in a model have no contradictions, together with the labels, the data across activities, and the requirements document of the model	

proposals studied were found. *Learnability* and *adaptability* were also used in ISO 9126 and ISO 25010, and in the case of *learnability*, 15% of the proposals found in the SLR also included it. *Modularity* was only included in ISO 25010.

We used the proposals found through the SLR for choosing other quality characteristics that authors considered important. That is the case of *completeness*, which was selected in 69% of the proposals, *consistency*, in 38% of the proposals, and *correctness* in 31% of the proposals. Since they are considered as relevant by most of the authors, we decided to include them.

3.2.2. *Internal quality*

The internal quality of BP models is tackled mainly by evaluating structural characteristics of the process models, along with other nonstructural aspects, such as, for example, those related to activity-labeling practices. This can be analyzed as a white box view of the model and addresses the static properties of models that are typically available for evaluation during the design. The design phase of a BP model generates a conceptual model that can be quantified by measures. For example, the control-flow complexity measure[33] is an indicator of the structural complexity of the model related to gateways, signifying that it provides information about the internal quality.

The principal idea involved in assessing the internal quality is the application of a group of measures (number of nodes, connector heterogeneity, etc.) which provide information about the model. These measures are used to support the BP models' quality. The measures chosen are presented in the following section.

4. Supporting Measures for BP Model Quality

In this section, measures are related to quality characteristics. The goal is to support the assessment of the model quality by analyzing the fulfillment of the specific quality characteristics. The published literature was studied in Sec. 2, and some measures for BP models were selected.[21] The relationships between these measures and quality characteristics will be specified in detail in this section.

4.1. *Influence of internal quality on external quality*

In order to establish the correspondence between measures and EQC, the software measurement ontology (SMO)[34] was considered. This ontology was defined to represent all the elements involved in the measurement process: "A quality model evaluates measurable concepts, which are related to attributes. Measurement is performed on attributes and entities. Measures are defined for attributes". This definition justifies the need to relate quality characteristics and measures in the effort to discover which measures are most capable of assessing the general quality of a conceptual model.

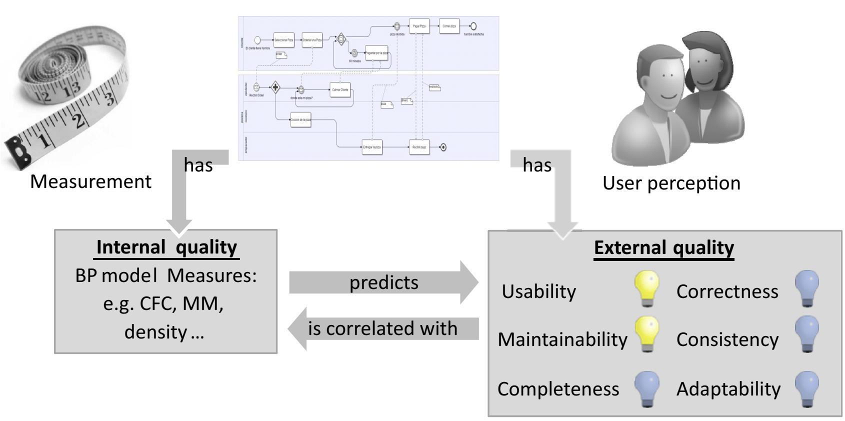

Fig. 2. Relationship between internal quality and external quality.

Figure 2 represents the relationship between internal quality and external quality in practice. First, measures are applied on BP models and measurement information is generated. These models are supposed to be syntactically correct, but their quality is still unknown. Since the relationship between EQC and measures was recognized through correlation analysis, measurement results are used to predict the level of each quality characteristic. Each quality characteristic can be evaluated through the application of a group of measures, and it is considered to be satisfied when the measures do not exceed a particular threshold. A bulb lights up when a group of measures indicates that a specific quality characteristic does not attain a desirable level. A light bulb is a metaphor for an activated trigger. The activation of a trigger indicates that some redesign initiatives should be applied to the specific model. After redesign, the model can again be a candidate for improvement.

How to deal with the analysis of EQC through the application of measures can be described as follows:

EQC1 — Understandability: Although this term is considered to be subjective, various experimental works have been developed in relation to it. A family of experiments to analyze the correlation between certain structural measures and understandability was carried out by Rolón *et al.*[35] This family was composed of three experiments in which a group of understandability tasks were carried out by subjects. Understandability was interpreted as the efficiency in realizing understandability tasks, which is calculated as the ratio between the number of correct answers and the time spent on tasks. The results of applying a correlation analysis with this experimental material and some BP measures were published.[11,35] These studies indicate that certain measures are significantly correlated with understandability efficiency (Table A.2). In this paper, we will complement the work of Rolón *et al.*[35] by using the experimental material of the aforementioned author, in addition to another group of measures.[12] This analysis is explained in Sec. 3.3.

However, not only structural measures should be taken into account when we analyze the understandability of a BP model. Another important aspect is related to the labels. As was described in Ref. 36, labels defined using the style *verb* +*object* are considered less ambiguous, and therefore more understandable. Other aspects related to labels are also studied in EQC9 – Consistency.

EQC2 — Modifiability: This quality characteristic was also analyzed in the research of Rolón *et al.*[10,35] with similar experiments to those used with the understandability characteristic. The specific measures correlated to modifiability are displayed in Table A.2. In this paper, we will complement that group of measures with others through a meta-analysis. This piece of research was started in Ref. 12 and will be explained in Sec. 3.3.

EQC3 — Learnability: The study of learnability for BP models requires certain assumptions. First, we must consider that the language used to model is already known by the user. That being the case, language learning is not taken into account in learnability evaluation. The second assumption is that the user has enough knowledge about the business topics. This is materialized in the BP model, which is already designed when the learnability evaluation is done. These assumptions led us to define learnability of BP models as the capability of learning a model, independently of the language with which it was designed (Table 5). A BP model is typically used as the communication vehicle among stakeholders, and using the model therefore signifies being able to understand the represented semantics, to then adapt it to the new business requirements.[37] In order to check the level of learnability, it is therefore necessary to demonstrate the understandability and modifiability capabilities, meaning that if the model is capable of being understood and modified, it satisfies the learnability requirements. Some factors should be taken into account if a model is to be considered as learnable, such as the modeling notation. As it was indicated in Refs. 38–40, Business Process Model and Notation (BPMN) has been conceived to facilitate users the understanding of BP models in comparison to other languages focused on more technical users. The prediction of this characteristic has not been talked about in related literature, so in this paper some measures for learnability are proposed, which may be seen in Table 9.

EQC4 — User Interface aesthetic: Carrying on with the similarities between a piece of software and a BP model, this characteristic refers to the interaction between user and the model. This is a complex concept and it can be faced from different perspectives, for example, the structural complexity of the model, the position of elements in the model (layout), or the correct definition of activity labels. Aesthetic considerations will affect design of the model, which is important when the business issues are presented to audiences, who have high expectations of aesthetic style.[41] In the study of Ref. 41 some guidelines for user interface aesthetic were collected: color, shape of elements, structured and consistent layout, etc. These guidelines need to be interpreted in the context of the application of user audience; in this case, layout is the most suitable for BP models. In the same way, Ref. 42 related design features, such as color, texture, and layout, to aesthetic and emotional responses by users. In this line, layout is one of the most important aspects for BP models from other fields and some measures for quantifying it have been published previously. Layout measures found in literature are shown in Tables 9 and A.2.

EQC5 — Adaptability: BPMN[43] is the de facto standard that is currently used to represent BPs at a conceptual stage. However, there are also other notations, such as Event-driven Process Chain (EPC),[44] Unified Modeling Language (UML)[45] etc. and it might eventually be interesting to adapt the model to another language. We believe that organizations tend to have *adaptable* BP models so that they can continuously update to new technologies or requirements. Certain structures cannot be represented in some languages, and the adaptability of the model is, therefore, limited. For example, a multi-merge pattern can be represented in BPMN but not in EPC, because the latter does not include an element with which to represent a synchronization of merging distinct branches in a single one. We consider this kind of structures as nonadaptable patterns. In the research of van der Aalst *et al.*[46] the different patterns were analyzed to discover whether they could be expressed in different notations. We must also take into consideration the different strengths of each modeling language, in order to value each language differently in each context. In this paper, we propose the notion of level of adaptability, to be specified through the analysis of each pattern in the model and the different notations which can represent them. For example, let us imagine a model with four patterns: explicit termination, synchronization, parallel split, and exclusive choice. Synchronization, parallel split, and exclusive choice are supported by seven languages or specific implementations (BPEL, Websphere Integration Developer, Oracle BPEL, BPMN, XPDL, UML, and EPC). However, explicit termination is not supported by three of them, EPC being a case in point. In this language, all the activated branches would continue to run until finished. Moreover, EPC is not the most suitable language for representing advanced behavior of models, so it should be evaluated differently than other behavior languages such as BPMN. In this case, the adaptability result of the example mentioned indicates that the model is not completely adaptable. A specific model is, therefore, more adaptable when it includes fewer nonadaptable patterns.

EQC6 — Modularity: BP models can be modularized. For example, models represented in BPMN[43] can be decomposed by using the subprocess element. The subprocess element helps to describe complex parts of a model in a separate one, thereby obtaining a better understandability of them. The modularization of a process model can have two effects: a benefit of information hiding and navigation costs.[47] Furthermore, modularization may encourage the understanding of a process model by its "information hiding" quality. The relationship between information hiding and navigation costs is subjective and depends on the personal opinion of the user. It is, therefore, necessary to find a suitable level of modularity, and this implies the modularization of a model when the number of nodes exceeds a specific number, which is called a threshold. Thresholds, such as the threshold value for modularization specified in Ref. 48, have been studied in various articles. In this paper, the authors proposed that a model with more than 50 elements should be modularized. Other authors in Ref. 12 proposed the number of 65 nodes previous to modularization tasks. These limit values help to detect when a process model needs to implement subprocesses. It could be specified by modularization when the number of nodes exceeds the threshold value.

EQC7 — Correctness: Error-probability measures are used to analyze whether or not the model has workflow errors. The prediction of the number of errors by using structural measures is possible, and was studied in Ref. 49. In this book, some structural measures were described and an empirical validation of predicting errors was included. These measures are shown in Tables 9 and A.2. The application of these measures to BP models makes it possible to detect error-models, resolving these before execution.

EQC8 — Completeness: Completeness measures can only indicate the users' satisfaction as regard the conversion of their business requirements into a conceptual model. This characteristic can be measured by precision and recall measures.[50] Precision indicates how exact the model is with regard to the user requirements, and recall describes how complete the model is. These measures provide a general indication of the completeness of the process model.

EQC9 — Consistency: The consistency quality attribute is widely used in software engineering with regard to, for example, the consistency between models in different abstraction levels.[51,52] The adaptation of the term consistency to software in BP models also generates the following classification:

- Consistency of activity labels: As was indicated in Ref. 53, activity labels should not be unnecessarily extensive or wordy and insufficiently informative. Names, roles, or definitions should be consistent with the referenced glossary.
- Data consistency: This can be viewed from a number of perspectives, such as data redundancy across activities. Codd's referential integrity constraint is an instantiation of this type of consistency.[54] In Ref. 55, the authors described a consistency measure based on the ratio of violations of a specific consistency type to the total number of consistency checks subtracted from one.
- Consistency with requirement specification: This consists of the appropriate alignment between requirements and their representation in a process model. It is difficult to measure automatically, and should therefore be assessed by checking requirements in the model. Many books on requirements engineering discuss requirement consistency checking informally, using reviewing method with conflict identification table and interaction matrix.[56,57] Other authors discussed consistency checking formally, using software cost reduction.[58] In this case, we use the concept of "requirements consistency" as an index of satisfaction by the domain experts. This can be measured through the use of checklists, as was indicated in Refs. 56 and 57.

4.2. *Empirical validation of model measures*

In the field of software and business measurement, a key step is to investigate whether the measure is actually effective in practice, that is, whether it is related to some quality characteristics worth studying and therefore helps in quality assessment. As stated in previous sections, there are diverse proposals on measures for BP models. In this section, we focus on selecting some relevant proposals from related literature which present empirical results about the possible effects of structural complexity on understandability and modifiability of BP models. Table 6 shows the classification of the measures chosen, which cover most of the characteristics of a model presented in this work. More details about those measures are on Table A.2.

Table 6. Classification of measures for applying meta-analysis.

Classification	Author	Measures
Size	Mendling[49]	Nodes, Diameter
	Rolón *et al.*[9]	NP, NEDDB, TNT, NEDEB, NID, NPG, NSFG, PDOPOUT, NL, NMF, NDOIN, NDOOUT, NCD, TNSE, TNIE, TNEE, NSFE, TNDO, TNCS, NSFA, TNG, TNA, TNSF, PDOPIN, PLT
Connection	Mendling[49]	Density, AGD, MGD
	Rolón *et al.*[9]	CLA, CNC
Modularity	Mendling[49]	Separability, Sequentiality, Structuredness, Depth
Connector	Mendling[49]	GM, GH
Interplay	Cardoso[33]	CFC
Complex behavior	Mendling[49]	Cyclicity, TS

All the measures presented in Table 6 have been validated with regard to their influence on understandability and modifiability by means of a family of experiments,[11,12,35] as a result of which some insights about their usefulness in the context of each experiment were obtained. However, it is important not only to extract specific conclusions but also to obtain general findings to reinforce their practical validity. This global analysis was achieved by means of applying meta-analysis, whose results are described in the next subsections, along with a summary of the research design and the results of the correlation analysis of the previous studies.

4.2.1. *Meta-analysis overview*

The main purpose of this section is to check the direct or indirect relationship between measures and quality characteristics, performing a meta-analysis to obtain conclusive correlation results. As we have said, the quality characteristics selected are understandability and modifiability.

"Meta-analysis refers to analysis of the analysis ... the statistical analysis of a collection of results from individual studies and whose purpose is to integrate those results. It is a rigorous alternative against purely narrative discussion of the results of a series of studies."[59] To carry out the meta-analysis, a family of experiments is needed. In this paper, we use a family of experiments designed by Ref. 35. A total of six experiments were conducted: three experiments to evaluate understandability and three to evaluate modifiability. In all, 127 students from four different universities took part in the experiments. The first experiment of understandability was done by 22 subjects (PhD students and students on 4th course of Computer Science), the second one by 40 subjects (students on 4th course of Computer Science), and the third one by 9 subjects (PhD students). In a similar way, the first experiment of modifiability was done by 29 subjects (students on 4th course of Computer Science), the second one was done by 15 subjects (students of Master in Information Systems), and the third one was done by 12 subjects (computer engineers). All the subjects had a high degree of knowledge about modeling, but very little experience in BPMN. To leverage the knowledge, all subjects received training in BPMN before the experiments.

The experimental material[a] for the first three experiments consisted of 15 BPMN models with different structural complexity. Each model included a questionnaire related to its understandability. The experiments on modifiability included 12 BPMN models (selected from the 15 models concerning understandability) and for each model some modification requests were proposed. For each model, the following data were collected: time of understandability or modifiability for each subject, number of correct answers in understandability or modifiability, and efficiency, defined as the number of correct answers divided by time.

Correlation results for measures and efficiency of understandability and modifiability are shown in Refs. 11, 12 and 35. This correlation analysis served to make the first filter, in order to get a group of measures considered to be good indicators of understandability or modifiability. To select the measures, the correlation results considered were those concerned with efficiency measure of dependent variables (understandability and modifiability). The correlation analysis lets us know the measures which can be useful to predict external quality of models. If the specific measure is significantly correlated with dependent variables measures (efficiency of understandability and modifiability) in the three experiments for each quality characteristic, the measure can be selected. If the measure is correlated in only one experiment, it will be eliminated directly. However, if the correlation is confirmed in only two of them, then a meta-analysis is necessary to obtain a general evaluation of the correlation analysis. All of these describe an algorithm for obtaining empirically validated measures in general, through the use of a family of experiments. These steps are described in Algorithm B.1 (see Appendix B).

4.2.2. *Meta-analysis results*

In this paper, the measures which were correlated in two of the three experiments need to be submitted to another test to confirm a significant correlation or not. There are several statistical methods that allow us to accumulate and interpret a set of results obtained through different experiments that are interrelated because they check similar hypotheses.[59–63] In this section, the meta-analysis is used to extract a general conclusion for the measures which were inconclusive results about correlation with understandability and modifiability. To carry out the meta-analysis presented, we used the Meta-Analysis v2 tool.[b] In this meta-analysis, we used the correlation results of each measure in each experiment, and from these values we obtained Hedges' g measure,[60] which we used as standardized measure. The Hedges'

[a]http://alarcos.inf-cr.uclm.es/bpmnexperiments.
[b]http://www.meta-analysis.com.

g measure is a weighted mean whose weights depend on the sample size of the experiments. The higher the value of Hedges' g is, the higher is the corresponding correlation coefficient. (Eq. (1), calculation of Hedges' g, $w_i = 1/(n_i - 3)$ and n_i is the sample size of the ith experiment).

$$\bar{Z} = \frac{\sum_i w_i z_i}{\sum_i w_i}. \tag{1}$$

A g between 1.01 and 3.4 depicts a high, between 0.38 and 1.00 is medium, and between 0.37 and 0 is low in terms of general correlation.[64] That means that, if the g indicates a high value for the TNE measure, then the TNE measure is strongly correlated. The results obtained are described in Table 7. In the first column, the measure is indicated. The column *correlation global effect size* indicates the global results of the correlation analysis. The next columns are inferior, superior limits, which represent the limits within which the global effect size can fluctuate, and the p-value determines if the conclusion of correlation is significant. The Hedges' g is then calculated, with the level of correlation indicated, which means that the labels high, medium, and low indicate the specific correlation level.

Table 7. Meta-analysis of the correlation analysis.

Measure	Correlation Global Effect Size	Inferior Limit	Superior Limit	p-Value	Hedges' g
		Efficiency of understandability			
CLP	−0.485	−0.652	−0.273	0.000	−1.063 (High)
NDOout	−0.332	−0.533	−0.096	0.007	−0.677 (Medium)
NDOin	−0.263	−0.477	−0.021	0.034	−0.523 (Medium)
NL	−0.013	−0.257	0.231	0.916	−0.026 (Low)
NSFE	−0.631	−0.758	−0.457	0.000	−1.558 (High)
PDOPin	−0.380	−0.584	0.394	0.704	−0.095 (Low)
PDOTout	−0.989	−0.739	0.243	0.323	−0.248 (Low)
PLT	0.538	−0.393	0.690	0.590	0.149 (Low)
TNCS	−1.451	−0.859	0.128	0.147	−0.365 (Low)
TNDO	−1.429	−0.850	0.133	0.153	−0.359 (Low)
TNIE	−0.665	−0.664	0.328	0.506	−0.168 (Low)
TNSE	−1.338	−0.832	−0.157	0.181	−0.337 (Low)
		Efficiency of modifiability			
CFC	−0.313	−0.544	−0.038	0.026	−0.608 (Medium)
CLA	0.209	0.073	0.461	0.145	0.406 (Medium)
Depth	−1.803	−1.090	−0.188	0.011	−0.818 (Medium)
NEDDB	−1.620	−1.028	0.098	0.105	−0.465 (Medium)
NEDEB	−1.628	−1.032	0.095	0.103	−0.469 (Medium)
NCD	−1.282	−0.923	0.193	0.200	−0.365 (Low)
NID	−1.618	−1.028	0.098	0.106	−0.465 (Medium)
NPF	−1.133	−0.874	0.234	0.257	−0.320 (Low)
NSFA	−0.075	−0.569	0.527	0.940	−0.021 (Low)
NSFG	−0.314	−0.545	−0.039	0.026	−0.611 (Medium)
PDOPout	0.272	−0.473	0.626	0.785	0.076 (Low)
TNA	−0.331	−0.643	0.458	0.741	−0.093 (Low)
TNG	−1.984	−1.155	−0.007	0.047	−0.581 (Medium)
TNSF	−1.555	−1.007	0.116	0.120	−0.445 (Medium)
TNT	0.263	−0.475	0.622	0.793	0.074 (Low)

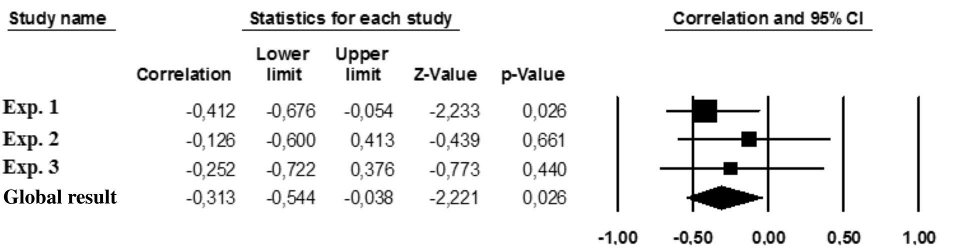

Fig. 3. Meta-analysis of CFC and efficiency of modifiability obtained by meta-analysis tool.

In Table 7, there are the meta-analysis results for the measures which obtained a significant correlation in two of the three experiments. The global results about understandability reveal that the 33% of the measures which obtained a significant correlation result with understandability in two of the three experiments give a significant global result and the correlation is satisfied because the g is medium or high. On the other hand, 27% of the measures and modifiability obtained a medium or high Hedges' g. We take into account that a number of them have a nonvalid p-value (p-value > 0.05) and these should be discarded.

For the readers' convenience, we show our meta-analysis results in diagram form, as provided by the Meta-Analysis v2 tool. Figure 3 shows the results of CFC and modifiability. Correlation results between the CFC measure and efficiency of modifiability in each experiment are represented, and finally, in the last row, the global results are provided. Not all the studies contribute equally to the overall conclusion, which is represented by the diamond in the last row of the figures. Each of them receives a specific weight in the meta-analysis, that is, the study's effect size, represented by the squares in the figures. The estimations for studies with a large sample size are more accurate, so they contribute more to the overall effect. However, sample size is not the only factor contributing to the weight of a study; for example, we should consider standard deviation.[65] In this case, we selected sample size for associating weights to each study. The weight of a study is proportional to the area of the corresponding square in the figures.

4.3. *Empirically validated measures for understandability and modifiability*

As a result of the application of the Algorithm B.1, the selected measures which are considered to be good predictors of understandability and modifiability are shown in Table 8.

More measures are significant for predicting understandability than modifiability, due to the fact that the structural complexity of models is directly related to the capacity of understanding them. Most of the measures such as n° nodes, diameter, TNE, TNA, etc. are counters of modeling elements and they seem to be relevant in

Table 8. Empirically validated measures for understandability and modifiability.

Quality attribute	Assumption Satisfied	Measures
Understandability	Correlated in the three experiments	$N°$ nodes, diameter, density, CNC, AGD, MGD, separability, sequentiality, depth, GM, TS, GH, NEEDB, NEDEB, NID, NCD, NPF, NSFG, TNG, NP, PDOPout, TNE, TNA, TNSF, CFC
	Correlation indicated by meta-analysis	CLP, NDOout, NDOin, NSFE
Modifiability	Correlated in the three experiments	AGD, MGD, separability, GM, GH
	Correlation indicated by meta-analysis	NSFG, CFC, TNG, depth

understandability terms. Few measures for modifiability are selected, in contrast to understandability. The capacity for modifying models is more complex to analyze than the other quality characteristic. In this case, most of the measures are related to *connections* or *connection interplay*. We, thus, conclude that the more sequential structures the model has, the easier to modify the model will be.

4.4. *Measures related to quality characteristics*

The application of meta-analysis helped us to choose a group of measures which are seen to be good indicators for understandability and modifiability. In addition, we resume all the measures (empirically validated or not) not only for those characteristics but also for the rest of the quality characteristics. For example, the measure $n°$ of nodes can predict the quality characteristics of understandability, learnability, modularity, and correctness, since the relationship among them was checked in past works. Table 9 can be useful in quality assessment of BP models. Validation of the relationship between measures and quality characteristics indicated in Table 9 requires to be addressed in future works.

5. Practical Example

This section provides an example to illustrate how the quality model and its corresponding measures can be applied in practice. The example BP has been adopted from a real case in a hospital and is named "incorporation of a new employee" (INE), which includes the training plan, information, and suitability of those people involved in the hospital, to facilitate their integration into the new job (Fig. 4). It is represented in BPMN. This model was created by a specific work group, consisting of specialists in modeling tasks (software engineers), in conjunction with health professionals at the hospital. This process was chosen as a low-complexity process, although the services provided are very important. It is a purely administrative process (i.e. it is not related to patient care) but has a throughput of a large number of users (in 2007, the hospital staff consisted of 2,600

Table 9. Measures related to each quality characteristic.

Empirically Validated Measure	Understandability	Learnability	User Interface Aesthetic	Modularity	Modifiability	Correctness	Completeness	Consistency
$N°$ Nodes	x	x		x		x		
Diameter	x					x		
Density	x					x		
CNC	x					x		
AGD	x				x	x		
MGD	x	x		x		x		
Separability	x	x		x		x		
Sequentiality	x					x		
Depth	x	x		x		x		
GM	x	x		x		x		
TS	x					x		
GH	x	x		x		x		
NEEDB	x							
NEDEB	x							
NID	x							
NCD	x							
NPF	x							
NSFG	x	x		x				
TNG	x	x		x				
NP	x							
PDOPout	x							
TNE	x							
TNA	x							
TNSF	x							
CFC	x	x		x		x		
CLP	x							
NDOout	x							
NDOin	x							
NSFE	x							
NSFG		x		x				
Structuredness						x		
Recall							x	
Precision							x	
Layout complexity			x					
Layout appropriateness			x					
Layout measure			x					
Consistency of activity labels								x
Data consistency								x
Consistency with requirement specification								x

workers, and 6,989 new contacts were made with regard to substitutions and new incorporations).

This process involves different professional categories: doctors, pharmacists, nurses, psychologists, administrative and technical staff, as well as others. Specific

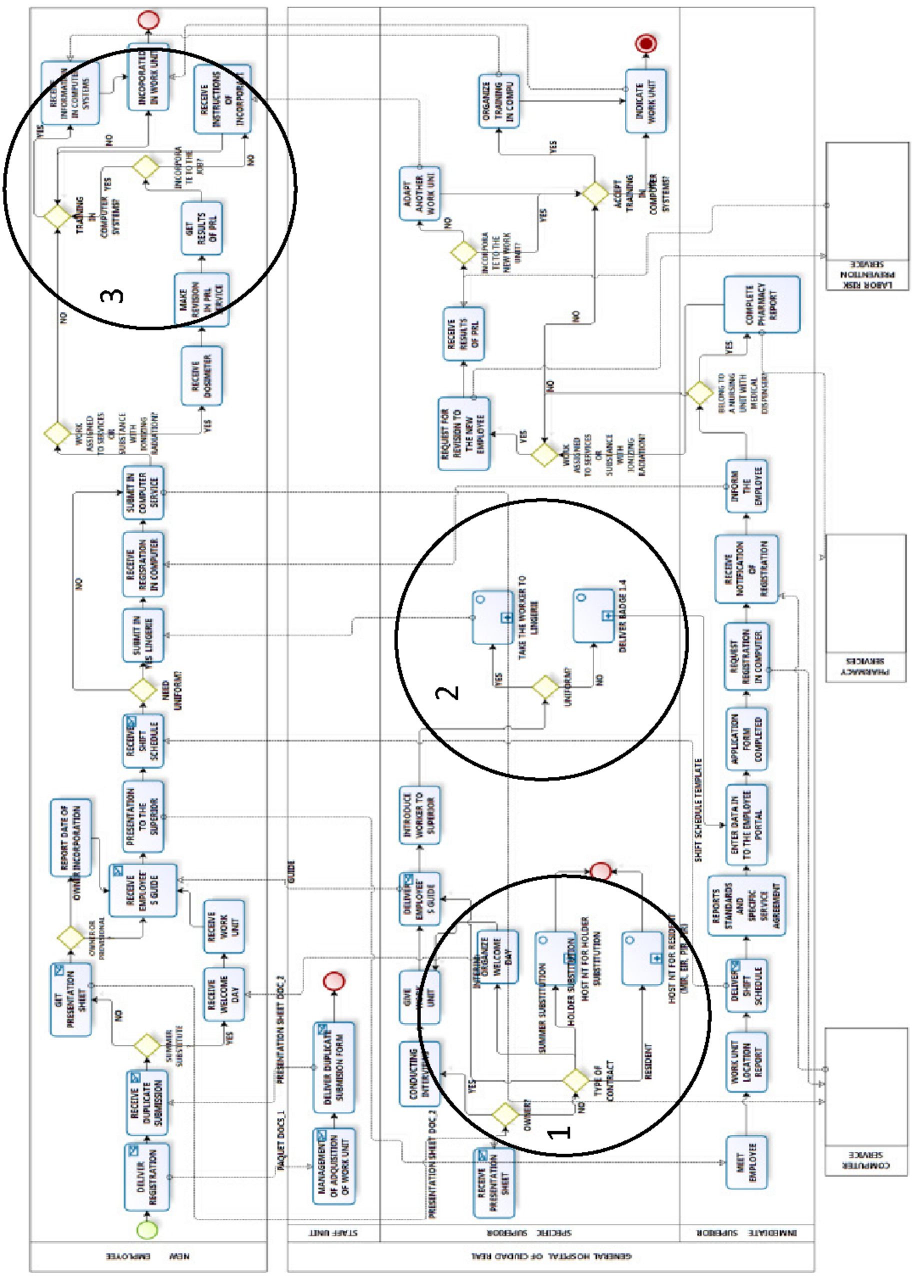

Fig. 4. BPMN model for the INE hospital process.

process characteristics were the following:

- Mission: to promote the organization of the INE process, which includes a plan for training and information, as well as the adaptation of the people involved to the hospital requirements, making their integration into the new job smoother.

- Limits: the INE process starts when the professional comes to the hospital and finishes when he/she is incorporated into the new job.
- Clients: new professionals.
- People responsible: those responsible for nursing, medical aspects, and management.
- Participants: new professionals, human resources, computer services, laundry and linen service, pharmacy, prevention services, nursing, and management service.
- Suppliers: human resources, provisions, maintenance, training, and information systems.

Figure 4 shows several signs of quality which were obtained by following the quality characteristics described in Sec. 2.2.1. The part of the diagram highlighted in Circle 1 indicates that some arcs cross with others, in addition to the overlapping labels; these are signs of a poor layout and may jeopardize the user interface aesthetics. The area of Circle 2 shows the use of subprocess elements, which indicates that the modularity quality characteristic has been taken into account. Circle 3 shows that the multiple-merge pattern has been used, which points to the fact that the adaptability of the model is not at its maximum, since this pattern is not adaptable to other languages such as EPC. However, the use of nonadaptable patterns could be more advantageous in other model aspects. The priority of one characteristic over another depends on the specific domain.

Other signs of quality for the model example in Fig. 4 are specified in Table 10, which shows the measurement result assessment according to the threshold obtained in Ref. 12 for *understandability* and *modifiability*. Thresholds divide the measures domains into levels: level 1, *very easy to understand/modify*, level 2, *easy to understand/modify*, level 3, *moderately modifiable/understandable*, level 4, *difficult to understand/modify*, and level 5, *very difficult to understand*. Example results show that *understandability* is poor in general, because 60% of the measures consider the model very difficult or difficult to understand. In contrast, *modifiability* measures indicate that the model is considered as moderately modifiable or easy to modify in 75% of the cases. Due to the poor results for *understandability*, *learnability* is also poor. Thresholds for *correctness* were published in Ref. 66, and we can observe the results in Table 10. Specifically, 60% of the measures indicate that the model is unlikely to have errors.

The number of nodes is 59, which indicates that, as was specified in Ref. 48, more subprocess elements are needed if a negative evaluation of *modularity* by the user is to be avoided. With regard to *completeness*, recall and precision measures should be quantified. In this case, it is not possible to specify the recall and precision measures and to give an assessment of them, because we lack the specific information. Completeness evaluation is outside the scope of this paper.

Finally, a checklist should be created by the user and the analyst in order to verify *consistency* with requirements and data. In this case, professionals in the

Table 10. Measurement results of the INE process.

Measure	Result	Understandability	Modifiability	Correctness
NP	5	Difficult to understand (L4)	No applicable	Not applicable
TNSF	75	Difficult to understand (L4)	No applicable	Not applicable
NMF	18	Difficult to understand (L4)	No applicable	Not applicable
Diam	25	Very difficult to understand (L5)	No applicable	Unlikely to have errors
AGD	3.5	Moderately understandable (L3)	Moderately modifiable (L3)	Likely to have errors
MGD	5	Moderately understandable (L3)	Moderately modifiable (L3)	Likely to have errors
Separability	0.66	Moderately understandable (L3)	Moderately modifiable (L3)	Unlikely to have errors
Sequentiality	0.39	Difficult to understand (L4)	No applicable	Unlikely to have errors
GM	19	Moderately understandable (L3)	Moderately modifiable (L3)	Likely to have errors
GH	0.23	Easy to understand (L2)	Easy to modify (L2)	Unlikely to have errors
CFC	20	Moderately understandable (L3)	Moderately modifiable (L3)	Unlikely to have errors
NEDDB	13	Very difficult to understand (L4)	Not applicable	Not applicable
NSFG	49	Very difficult to understand (L5)	Very difficult to modify (L5)	Not applicable
TNG	14	Difficult to understand (L4)	Difficult to modify (L4)	Not applicable
NSFE	6	Moderately understandable (L3)	Not applicable	Not applicable
TNE	5	Easy to understand (L2)	Not applicable	Not applicable
CLP	3	Moderately understandable (L3)	Not applicable	Not applicable
TNA	45	Difficult to understand (L4)	Not applicable	Not applicable
Nodes	60	Difficult to understand (L4)	Not applicable	Likely to have errors
Density	0.021	Very difficult to understand (L5)	Not applicable	Unlikely to have errors

Table 11. Assessment of quality for the model example.

(Sub)characteristic	Assessment
Understandability	Fairly inefficient
Learnability	Fairly inefficient
User interface aesthetics	Poor layout (arcs crossing with others and overlapping labels)
Modifiability	Efficient
Modularity	More subprocesses are needed
Adaptability	It is not maximum because of the use of a nonadaptable pattern
Correctness	Unlikely to have errors
Completeness	Unknown
Consistency	Professionals in health sector confirm this

health sector also confirmed label consistency because of the convention of using *infinitive verbs + complements.*

The global assessment of the quality (Table 11) of the model example indicates that several redesign initiatives should be applied, particularly to understandability, learnability, and modifiability, whose results were undesirable in each case.

6. Conclusion and Future Work

A growing number of works concerning process modeling has been published in the last few years in the BP research community. These works discuss the need to pay more attention to the quality of conceptual models. A conceptual model is not only a picture, but is also considered to be a key instrument in business development.

This paper has used an SLR to collect the various proposals related to quality characteristics published in literature. The analyses of these papers have revealed several problems: principally, the lack of consensus among researchers with regard to what the most important quality characteristics in process models are. While a difference of opinion on various issues is good, an agreement should at least be reached on the fundamentals. Since the uncontrolled proliferation of quality models is counterproductive to research progress, we have collected the most suitable quality characteristics for BP models, with the main purpose of unifying the evaluation of quality. In specific terms, the quality characteristic that is used most by authors as regards conceptual models in general has been *completeness*, but the proposals which focus more specifically on BP models agree on the study of the quality characteristic of *understandability*. Due to the immature nature of quality in BP models, it is very ambitious to validate a complete quality framework, but our set of quality characteristics for BP models is an attempt to contribute to the maturity of the field.

It is important to recall that quality characteristics are usually too subjective to be quantified with only one measure, and that a group of measures is needed. These measures should also be validated in order to ensure the usefulness of measurement results; however, there are still very few validated measures in this field. This is the main problem about a complete and useful definition of a quality framework. Although a BP model is similar to a typical conceptual model, some characteristics should be particularly emphasized because of the nature of their users (users are not always from the technical field), and *understandability* is recognized as being an important aspect that should be taken into account in quality assessment.

The proposed quality model is a first approach toward assessing the quality of conceptual models in a more objective manner. The characteristics presented cover an important part of the quality of models, but as quality is subjective in nature, the proposals should in future be enriched with further research and the application of the quality model in case studies.

Other issues intended for future work concern the empirical validation of more representative measures related to the proposed quality characteristics. After empirical validation of measures, reduction techniques may be needed, in order to avoid redundant information given by different measures; for example, by ABC analysis or PCA tests. The idea is to have a limited group of measures which can cover all the perspectives of the BP model, thereby enhancing the objectivity of the quality assessment.

Acknowledgments

This work was partially funded by the following projects: PEGASO/MAGO (Ministerio de Ciencia e Innovación MICINN and Fondo Europeo de Desarrollo Regional FEDER, TIN2009-13718-C02-01); ALTAMIRA (Junta de Comunidades de Castilla-La Mancha, Fondo Social Europeo, PII2I09-0106-2463); ESFINGE (Ministerio de Educación y Ciencia, Dirección General de Investigación/Fondos Europeos de Desarrollo Regional (FEDER), TIN2006-15175-C05-05); and INGENIOSO (Junta de Comunidades de Castilla La Mancha, PEII11-0025-9533).

Appendix A. Information Extracted in SLR

The information extracted from selected sources in the SLR and measures is shown in Tables A.1 and A.2.

Appendix B. Algorithm for Obtaining Empirically Validated Measures

This algorithm describes how to face the empirical validation of measures in a family of experiments. First, the correlation coefficient is calculated for each experiment between the dependent and the independent variable (e.g. the measure of CFC and the efficiency of understandability). If the measure is significantly correlated in all the experiments, this measure is concluded to have been correlated. However, if the measure is correlated in fewer experiments than ("number of experiments"/2) − 1, this measure is discarded because it cannot be considered to be correlated with the dependent variable. The last case is that the measure is correlated in more experiments than ("number of experiments"/2) − 1 but not in all the experiments. In this case, we do not give a conclusion about correlation, and another technique is needed, namely meta-analysis.

Table A.1. Information extracted from selected papers in SLR about quality characteristics.

Proposal	Based on	Validation	Supporting Measures?	Quality Characteristics
Mohagheghi *et al.*[25]	Several works found in a review	No	No	Correctness, completeness, consistency, comprehensibility, confinement, changeability
Qi *et al.*[18,67]	Lindland[23] and ISO 25010[14]	Survey	No	Syntactic (lucidity, soundness, laconicity), semantic (completeness, inherence, clarity, consistency, correctness, precision), pragmatic (simplicity, understandability, context conformity, context extendibility, accessibility), cognitive (user perceptibility, user satisfaction, credibility, attractiveness, technical accessibility), maintainability (modularity, reusability, modification stability, testability, traceability), reliability (maturity, availability), usage (ease of use, implementability), and social (agreeableness, time behavior, resource utilization)
Guceglioglu and Demirors[17]	ISO 9126[13]	No	Yes	Functionality (suitability, IT functionality, accuracy, interoperability, security), reliability (maturity, recoverability) usability (understandability, operability, attractiveness) maintainability (analyzability)
Lindland *et al.*[23]	Semiotic theory	Empirical	No	Collected characteristics: annotated, appropriate, complete, conceptually clean, consistent, constructable, correct, executable, expressive economy, formal, incompleteness tolerant, minimal, modifiable, precise Proposed quality dimensions: syntactic, semantic and pragmatic
Becker *et al.*[26]	Not specified	No	No	Correctness, relevance, economic efficiency, clarity, comparability and systematic design
Rittgen[24]	Lindland *et al.*[23]	Qualitative	Yes	Correctness, relevance, completeness, authenticity, understandability, performance, effectiveness, productivity

Table A.1. (*Continued*)

Proposal	Based on	Validation	Supporting Measures?	Quality Characteristics
Mehmood and Cherfi[68]	Unknown	Survey	Yes	Functionality (completeness, relevancy to requirements, practicability, expressiveness, reusability, reliability)
Satpathy *et al.*[16]	ISO 9001/9000-3, ISO 12207, CMM, ISO 9126 and FURPS model	No	No	Functionality (compliance, completeness, consistency, generality, suitability, inter-operability, security), usability (understandability, learnability, operability), efficiency and estimation (cost/effort estimation, cycle time, complexity estimation, resource usage, schedule/priority estimation, process maturity), visibility and control (automatic checks and feedback, progress monitoring, improvement measures), reliability (failure frequency, fault tolerance, recoverability), safety (risk avoidance), scalability, maintainability (analyzability, modifiability, stability, testability, defect trend, formal verifiability, informal verifiability, reusability, portability)
Shanks *et al.*[30]	Not specified	No	No	Accuracy, completeness, conflict-free, no redundancy
Matook and Indulska[31]	Several authors	Case study	yes	Generality, flexibility, completeness, usability, understandability
Teeuw and van der Berg[69]	Computer architecture in protocol specification[70]	No	No	Completeness, inherence (property), clarity, consistency, orthogonality, generality
Heravizadeh *et al.*[29]	Unknown	Case study	No	Suitability, accuracy, security, reliability, understandability, learnability, time efficiency, resource utilization, effectiveness, productivity, safety, user satisfaction, robustness

Table A.2. Measures for BP models.

Measure	Source	Empirical Validation?	Description
Average gateway degree (AGD)	49	Yes, correctness	Average of the number of both incoming and outgoing arcs of the gateway nodes
Coefficient of connectivity (CNC)	9	Yes, correctness	Ratio of the total number of arcs in a process model to its total number of nodes
Cohesion	71	No	To what degree the different operations within one activity are related
CLA	9	Yes, understandability and modifiability	Connectivity level between activities
CLP	9	Yes, understandability and modifiability	Connectivity level between participants
Consistency of activity labels	53	No	It indicates if the activity labels are unnecessarily extensive or wordy and insufficiently informative
Control-flow complexity	33	Yes, correctness	Complexity of the split gateways or, in other words, the number of mental states that have to be taken into account when a designer develops a process
Coupling	71	No	How strongly the activities in a workflow process are related/connected to each other
Cyclicity	49	Yes, correctness	Number of nodes in a cycle to the sum of all nodes
Data consistency	54, 55	No	It indicates data redundancy across activities, or the ratio of violations of a specific consistency type to the total number of consistency checks subtracted from one
Density	49	Yes, correctness	Ratio of the total number of arcs
Depth	49	Yes, correctness	Maximum nesting of structured blocs
Diameter	49	Yes, correctness	The length of the longest path from a start node to an end node
Gateway mismatch (GM)	49	Yes, correctness	Sum of gateway pairs that do not match with each other
Gateway heterogeneity (GH)	49	Yes, correctness	Different types of gateways that are used in a model
Layout appropriateness	72	Yes, user interface aesthetics	Efficiency of a screen in terms of cost involved in completing a collection of tasks
Layout complexity	73	Yes, user interface aesthetics	Evaluate the usability of different screen designs based on the Shannon formula
Layout measures	74	Yes, user interface aesthetics	A group of measures that quantify layout of models: number of edge crossing, number of non-rectilinear edges, overlapping area, etc.
Maximum gateway degree (MGD)	49	Yes, correctness	Maximum sum of incoming and outgoing arcs of these gateway nodes

Table A.2. (*Continued*)

Measure	Source	Empirical Validation?	Description
NCD	9	Yes, understandability and modifiability	Number of complex decision
NDOin/NDOout	9	Yes, understandability and modifiability	Number of data objects which are input/outputs of activities
NID	9	Yes, understandability and modifiability	Number of inclusive decision
NEDDB	9	Yes, understandability and modifiability	Number of exclusive data-based decision
NEDEB	9	Yes, understandability and modifiability	Number of exclusive event-based decision
NL	9	Yes, understandability and modifiability	Number of lanes
NMF	9	Yes, understandability and modifiability	Number of message flows
NP	9	Yes, understandability and modifiability	Number of pools
NPF	9	Yes, understandability and modifiability	Number of parallel forking
NSFA	9	Yes, understandability and modifiability	Number of sequence flows between activities
NSFE	9	Yes, understandability and modifiability	Number of sequence flows from events
NSFG	9	Yes, understandability and modifiability	Number of sequence flows from gateways
Number of nodes	49	Yes, correctness	Number of activities and routing elements
PDOPin/PDOPout	9	Yes, understandability and modifiability	Proportion of data objects as incoming/outgoing products and total data objects
PDOTout	9	Yes, understandability and modifiability	Proportion of data objects as outgoing product of activities of the model

Table A.2. (*Continued*)

Measure	Source	Empirical Validation?	Description
PLT	9	Yes, understandability and modifiability	Proportion of pools/lanes and activities
Recall and precision	50	No	Correctness and completeness between requirements and models
Sequentiality	49	Yes, correctness	Degree to which the model is constructed out of pure sequences of tasks
Separability	49	Yes, correctness	Number of cut-vertices, that is nodes that serve as bridges between otherwise disconnected components, to the total number of nodes in the process model on the other
Structuredness	75	Yes, correctness	Identification of unstructured patterns
TNA	9	Yes, understandability and modifiability	Total number of activities
TNCS	9	Yes, understandability and modifiability	Total number of collapsed subprocess
TNDO	9	Yes, understandability and modifiability	Total number of data objects in the model
TNE	9	Yes, understandability and modifiability	Total number of events
TNEE	9	Yes, understandability and modifiability	Total number of end events
TNG	9	Yes, understandability and modifiability	Total number of gateways
TNIE	9	Yes, understandability and modifiability	Total number of intermediate events
TNSE	9	Yes, understandability and modifiability	Total number of start events
TNSF	9	Yes, understandability and modifiability	Total number of sequence flows
TNT	9	Yes, understandability and modifiability	Total number of tasks
Token split (TS)	49	Yes, correctness	Maximum number of paths in a process model that may be concurrently activate

Input

`nme` is the number of measures
`ne` is the number of experiments
`ef[]` is a vector which contains the efficiency values for all the subjects, where `ef[j]` corresponds to the experiment j and `j ϵ [1,ne]`
`measure[]` is a vector of the measure values, where each element `measure[i]` corresponds to the measure values of `measure i`, where `i ϵ [1,nme]`

output

`validM` is a list which is initialized with all the measures

Algorithm

```
for i := 1 to nme
        for j := 1 to ne
              mcorrel = mcorrel + SpearmanCorrelation (ef[j],
measure[i])
              next j
        if mcorrel < ((ne/2)-1) then
              Delete measure[i] from validM
        else
              if ((ne/2)-1) ≤ mcorrel < ne then
                        meta = metaanalysis(ef, measure[i])
                        if meta = false then
                                Delete measure[i] from validM
                        end-if
              end-if
        next i
end-for
```

Algorithm B.1. Steps for obtaining empirically validated measures.

References

1. M. Weske, *Business Process Management*: *Concepts*, *Languages*, *Architectures*, 1st edn. (Springer-Verlag, Berlin, 2007).
2. I. Davies *et al.*, How do practitioners use conceptual modeling in practice? *Data Knowl. Eng.* **58** (2006) 358–380.
3. A. Enders and H. D. Rombach, *A Handbook of Software and Systems Engineering*: *Empirical Observations. Laws and Theories* (Addison-Wesley, Reading, MA, 2003).
4. B. W. Boehm, *Software Engineering Economics* (Prentice-Hall, Englewood Cliffs, 1981).
5. D. Moody, Theoretical and practical issues in evaluating the quality of conceptual models: Current state and future directions, *Data Knowl. Eng.* **55** (2005) 243–276.
6. J. V. Nickerson *et al.*, The spatial nature of thought: Understanding information systems design through diagrams, in *Proc. Int. Conf. on Information Systems* (2008).
7. B. Selic, The pragmatics of model-driven development, *Softw. IEEE* **20**(5) (2003) 19–25.

8. A. Fouad *et al.*, Embedding requirements within model-driven architecture, *Softw. Qual. Control* **19**(2) (2011) 411–430.
9. E. Rolón, F. García and F. Ruiz, Evaluation measures for business process models, *Symp. in Applied Computing SAC06*, Dijon, France (2006), pp. 1567–1568.
10. E. Rolón *et al.*, Prediction models for BPMN usability and maintainability, *BPMN 2009 — 1st Int. Workshop on BPMN*, Vienna, Austria (2009), pp. 383–390.
11. E. Rolón *et al.*, Analysis and validation of control-flow complexity measures with BPMN process models, *The 10th Workshop on Business Process Modeling Development and Support*, Amsterdam, The Netherlands (2009), pp. 58–70.
12. L. Sánchez-González *et al.*, Quality assessment of business process models based on thresholds, *CoopIS 2010 — 18th Int. Conf. Coop. Inf. Sys.*, Creta, Greece (2010), pp. 78–95.
13. ISO/IEC, *9126-1*, *Software Engineering — Product Quality — Part 1*: *Quality Model*, 2001.
14. ISO/IEC, *25010*, *Systems and Software Engineering — System and Software Product Quality Requirements and Evaluation* (*SQuaRE*) *— System and Software Quality Models*, 2011.
15. L. Osterweil, Software processes are software too, in *Proc. 9th Int. Conf. Software Engineering*, Monterey, CA, USA (1987), pp. 2–13.
16. M. Satpathy *et al.*, A generic model for assessing process quality, in *Proc. 10th Int. Workshop New Approaches in Software Measurement* (Springer-Verlag, 2000), pp. 94–110.
17. S. Guceglioglu and O. Demirors, Using software quality characteristics to measure business process quality, *Business Process Management* (2005) 374–379.
18. Y.-D. Qi, N. Qu and X.-F. Xie, Towards a preliminary ontology for conceptual model quality evaluating, *Web Information Systems and Mining Int. Conf.*, Sanya, China (2010), pp. 329–334.
19. M. Genero *et al.*, A systematic literature review on the quality of UML models, *J. Database Manage.* **22**(3) (2011) 46–70.
20. H. Nelson *et al.*, A conceptual modeling quality framework, *Softw. Qual. J.* **20**(1) (2012) 201–228.
21. L. Sánchez-González *et al.*, Measurement in business processes: A systematic review, *Bus. Process Manage. J.* **16**(1) (2010) 114–134.
22. B. Kitchenham and S. Charters, *Guidelines for Performing Systematic Literature Reviews in Software Engineering*, Technical Report (Keele University and University of Durham, 2007).
23. O. I. Lindland, G. Sindre and A. Solvberg, Understanding quality in conceptual modeling, *IEEE Softw.* **11**(2) (1994) 42–49.
24. P. Rittgen, Quality and perceived usefulneses of process models, in *SAC 10*, *Proc. ACM Symp. Applied Computing*, Lausanne, Switzerland (2010), pp. 65–72.
25. P. Mohagheghi, V. Dehlen and T. Neple, Definitions and approaches to model quality in model-based software development — A review of literature, *Inf. Softw. Technol.* **51**(12) (2009) 1646–1669.
26. J. Becker, M. Rosemann and C. von Uthmann, Guidelines of business process modeling, in *Business Process Management* (Springer, Berlin, Heidelberg, 2000), pp. 241–262.
27. K. Mehmood and S. Cherfi, Evaluating the functionality of conceptual models, in *Proc. ER 2009 Workshops on Advances in Conceptual Modeling — Challenging Perspectives* (Springer-Verlag, Gramado, Brazil, 2009), pp. 222–231.

28. B. Teeuw and H. van der Berg, On the quality of conceptual models, in *Proc. ER-97 Workshop on Behavioral Modeling and Design Transformation: Issues and Opportunities in Concept Model*, UCLA, Los Angeles, CA (1997).
29. M. Heravizadeh, J. Mendling and M. Rosemann, Dimensions of business processes quality (QOBP), *6th Int. Conf. on Business Process Management Workshops* (BPM Workshops, Milan, Italy, 2008), 2008.
30. G. Shanks, E. Tansley and R. Weber, Using ontology to validate conceptual models, *Commun. ACM* **46**(10) (2003) 85–89.
31. S. Matook and M. Indulska, Improving the quality of process reference models: A quality function deployment-based approach, *Decis. Support Syst.* **47**(1) (2009) 60–71.
32. L. Briand *et al.*, A comprehensive investigation of quality factors in object-oriented designs, An industrial case study, *Tech. Rep. ISERN-98-29*, Los Angeles, CA, USA (1998), pp. 345–354.
33. J. Cardoso, Process control-flow complexity metric: An empirical validation, in *SCC '06: Proc. IEEE Int. Conf. Services Computing*, Chicago, USA (2006), pp. 167–173.
34. F. García *et al.*, Towards a consistent terminology for software measurement, *Inf. Softw. Technol.* **48** (2005) 631–644.
35. E. Rolón *et al.*, Evaluation of BPMN models quality, A family of experiments, *ENASE — Int. Conf. Evaluation of Novel Approaches Software Engineering*, Frunchal, Madeira (2008), pp. 56–63.
36. J. Mendling, H. A. Reijers and J. Recker, Activity labeling in process modeling: Empirical insights and recommendations, *Inf. Sys.* **35**(4) (2010) 467–482.
37. A. Seffah *et al.*, Usability measurement and metrics: A consolidated model, *Softw. Qual. Control* **14**(2) (2006) 159–178.
38. D. Q. Birkmeier, S. Klöckner and S. Overhage, An empirical comparison of the usability of BPMN and UML activity diagrams for business users, *ECIS 2010*, Pretoria, South Africa (2010).
39. A. G. Nysetvold and J. Krogstie, Assessing business processing modeling languagues using a genering quality framework, *CAiSE 2005*, Porto, Portugal (2005), pp. 545–556.
40. S. A. White, *Process Modeling Notations and Workflow Patterns*, Workflow Handbook (Fischer, L., ED., Furutre Strategies Inc., Lighthouse Point, FL), 2004, pp. 265–294.
41. A. Sutcliffe, S. Kurniawan and J. E. Shin, A method and advisor tool for multimedia user interface design, *Int. J. Human-Comput. Stud.* **64** (2005) 375–392.
42. J. Kim, J. Lee and D. Choi, Designing emotionally evocative homepages: An empirical study of the quantitative relations between design factors and emotional dimensions, *Int. J. Human-Comput. Stud.* **59**(6) (2003) 899–940.
43. OMG, *Business Process Modeling Notation (BPMN), Final Adopted Specification*, 2006. http://www.omg.org/bpm (accessed 1 March 2012).
44. A. W. Scheer, *Business Processes Engineering: Reference Models for Industrial Enterprises* (Springer-Verlag, Berlin, 1994).
45. OMG, *UML Resource Page,* 2009. http://www.uml.org (accessed 1 March 2012).
46. W. M. P. van der Aalst *et al.*, Workflow patterns, *Distrib. Parallel Databases* **14**(1) (2003) 5–51.
47. H. A. Reijers, J. Mendling and R. Dijkman, On the usefulness of subprocesses in business process models, *BPM Center Rep. BPM-10-03*, BPMcenter.org, 2010.
48. J. Mendling, H. A. Reijers and W. M. P. van der Aalst, Seven process modeling guidelines (7PMG), *Inf. Softw. Technol.* **52**(2) (2010) 127–136.
49. J. Mendling, *Metrics for Process Models: Empirical Foundations of Verification, Error Prediction, and Guidelines for Correctness* (Springer Publishing Co., Incorp., 2008).

50. V. V. Raghavan, P. Bollmann and G. S. Jung, Retrieval system evaluation using recall and precision: Problems and answers, *SIGIR Forum* **23**(SI) (1989) 59–68.
51. C. F. J. Lange, Assessing and improving the quality of modeling, PhD thesis, Technische Universiteit Eindhoven (2007).
52. B. Berenbach, The evaluation of large, complex UML analysis and design models, in *Proc. 26th Int. Conf. Softw. Eng.*, Scotland, UK (2004), pp. 232–241.
53. B. W. Boehm, J. R. Brown and M. Lipow, Quantitative evaluation of software quality, in *Proc. 2nd Int. Conf. Softw. Eng.* (IEEE Computer Society Press, San Francisco, California, USA, 1976), pp. 592–605.
54. E. F. Codd, More commentary on missing information in relational databases (applicable and inapplicable information), *SIGMOD Rec.* **16**(1) (1987) 42–50.
55. L. L. Pipino, Y. W. Lee and R. Y. Wang, Data quality assessment, *Commun. ACM* **45**(4) (2002) 211–218.
56. G. Kotonya and I. Sommerville, *Requirements Engineering*: *Process and Techniques* (John Wiley, 1998).
57. I. Sommerville and P. Sawyer, *Requirements Engineering*: *A Good Practice Guide* (John Wiley, 1997).
58. C. Heitmeyer, R. Jeffords and B. Labaw, Automated consistency checking of requirements specification, *ACM Trans. Softw. Eng. Methodol.* **5** (1996) 231–261.
59. G. V. Glass, B. McGaw and M. L. Smith, *Meta-Analysis in Social Research* (Sage Publications, 1981).
60. L. Hedges and I. Olkin, *Statistical Methods for Meta-Analysis* (Academia Press, 1985).
61. R. Rosenthal, *Meta-Analytic Procedures for Social Research* (Sage publications, 1986).
62. J. Sutton *et al.*, *Methods for Meta Analysis in Medical Research Area* (John-Wiley, 2001).
63. F. M. Wolf, *Meta Analysis*: *Quantitiative Methods for Research Synthesis* (Sage Publications, 1986).
64. V. Kampenes *et al.*, A systematic review of effect size in software engineering experiments, *Inf. Softw. Technol.* **48**(8) (2007) 745–755.
65. M. W. Lipsey and D. B. Wilson, *Practical Meta-Analysis* (Sage Publications, 2001).
66. J. Mendling *et al.*, Thresholds for error probability measures of business process models, *Int. J. Sys. Softw.* **85**(5) (2012) 1188–1197.
67. Y.-D. Qi *et al.*, Analysis of contribution of conceptual model quality to software reliability, *Computer Application and System Modeling* (*ICCASM*), *2010 Int. Conf.*
68. K. Mehmood and S. S.-S. Cherfi, Evaluating the functionality of conceptual models, in *Proc. ER 2009 Workshops* (*CoMoL*, *ETheCoM*, *FP-UML*, *MOST-ONISW*, *QoIS*, *RIGiM*, *SeCoGIS*) *Adv. Concept. Model. — Challenging Perspect.* (Springer-Verlag: Gramado, Brazil, 2009), pp. 222–231.
69. B. Teeuw and H. van der Berg, On the quality of conceptual models, in *Proc. ER-97 Workshop on Behav. Model. Design Transform.*: *Issues and Opportunities in Concept. Model.*, UCLA, Los Angeles, CA, USA (1997).
70. G. A. Blaauw and F. P. Brooks Jr., *Computer Architecture.* Lecture notes University of Twente, Enschede, The Netherlands and University of Carolina, 1985.
71. I. Vanderfeesten, H. A. Reijers and W. M. P. van der Aalst, Evaluating workflow process designs using cohesion and coupling metrics, *Comput. Indus.* **50**(5) (2008) 420–437.
72. A. Sears, Layout appropriateness: A metric for evaluating user interface widget layout, *IEEE Trans. Softw. Eng.* **19** (1993) 707–719.

73. T. Comber and J. R. Maltby, Investigating layout complexity, in *Proc. Graphics Interface*, Toronto, Canada (1996), pp. 209–228.
74. H. Eichelberger and K. Schmid, Guidelines on the aesthetic quality of UML class diagrams, *Inf. Softw. Technol.* **51**(12) (2009) 1686–1698.
75. R. Laue and J. Mendling, Structuredness and its significance for correctness of process models, *Information Systems and E-Business Management* **8**(3) (2010) 287–307.