Requirements Engineering: The Quest for the Dependent Variable

Hannes Holm, Teodor Sommestad, Johan Bengtsson
Swedish Defence Research Agency (FOI)
Linköping, Sweden
hannes.holm@foi.se, teodor.sommestad@foi.se, johan.bengtsson@foi.se

Requirements engineering is a vibrant and broad research area. It covers a range of activities with different objectives. By reviewing experiments previously included in systematic literature reviews, this paper provides an overview of the dependent variables used in experimental requirements engineering research. This paper also identifies the theoretical motivation for the use of these variables in the experiments. The results show that a wide range of different variables has been applied in experiments and operationalized through both subjective assessments (e.g., subjects' perceived utility of a technique) and objective measurements (e.g., the number of defects found in a requirements specification). The theoretical basis for these variables and operationalizations are unclear in most cases. Directions for theoretical work to identify suitable dependent variables are provided.

***Index Terms*—Requirements engineering, experiments, dependent variables, frameworks, measurement, theory.**

I. Introduction

Software requirements engineering (RE) is seen as the process of determining the requirements that software systems should meet [1]. While variations exist, the process of doing requirements engineering (RE) has been more or less agreed upon. It starts with elicitation and continues with analysis, specification, and validation of requirements. Throughout these phases, statements and claims must be transformed to a set of quality-assured and well-formulated requirements [2]. Similar descriptions can be found throughout the literature [3]–[5], suggesting that RE is well-defined on a conceptual level. However, there is no widely agreed standard for how to measure the *success* of individual RE activities or RE as a whole. In the remainder of the paper, success refers to the ability of a RE activity to achieve its stated purpose.

Wieringa and Heerkens [6] found that RE research papers rarely validate proposed solutions or consider alternatives, which might be explained by the lack of standards for measuring success. Still, there is a need for an agreement within the RE community concerning measures of success to allow RE research to produce results that can be compared and evaluated effectively [7]. Similarly, a recent systematic literature review by Bano et al. [8] found that the practical effectiveness of RE research has not yet been properly addressed by software literature reviews.

The objective of this paper is to provide an overview of how successful RE is conceptualized and operationalized in existing research and to suggest directions for future research. By reviewing experimental RE research previously included in systematic reviews, this paper answers the following research question for each of the elicitation, analysis, specification, and validation phases:

RQ1: How are requirements engineering activities measured?

As a second objective, this study attempts to identify the theoretical and practical rationale behind the choices of dependent variables of different studies. In other words:

RQ2: Why are these measurements used?

The outline for the paper is as follows. Section II presents work related to the objectives of our study. Section III describes the review protocol. Section IV presents our results. Section V discusses the results and their limitations. Finally, Section VI concludes the paper and suggests future work.

II. An Overview of Dependent Variables in Requirements Engineering

As will be further described in section III.A, a systematic review was carried out to answer the two research questions. This review explicitly searched for previous literature reviews in RE and ought to have identified all existing overviews of measurement procedures used for RE activities.

Only one overview that identified dependent variables used in scholarly RE research was found – the review of techniques for the elicitation phase described in [9]. In this review, 50 different dependent variables were found and classified into seven classes such as quantity (20 variables) and time (six variables). In addition to [9], a number of other reviews touch upon the research questions discussed in this paper, and a few conceptual papers discuss similar questions. These are presented below.

A formative study of measurement of the quality of RE (mainly the requirements specification, RS) was presented in [10] and [11]. Thirty-four measures for RE success were identified based on a review of literature published at the time (almost 20 years ago) and interviews with experts. These 34 measures were clustered into three dimensions for RE success: (1) cost effectiveness of RE processes, (2) quality of RE products, and (3) quality of RE services. After further refinement, using a factor analysis of survey data from an

organization, 16 criteria were identified (of which none concerned the cost effectiveness of RE processes).

Another frame of reference for evaluation of research questions related to requirement specifications is provided in [12], which lists four qualities to assess: (1) quality of the requirements specification (e.g., understandability), (2) effectiveness/efficiency of specification activities (e.g., verification), (3) effort in subsequent processes (e.g., writing code), and (4) effectiveness/efficiency in software testing (e.g., acceptance tests).

[13] employed a process perspective and used an expert panel to validate a maturity model for RE. However, only 29 percent of the panel members considered the suggested RE capability maturity model complete. A process perspective was also used by Beecham et al. [14], who presented a framework for improving RE process management.

If the quest for dependent variables is broadened to information systems success in general, the consequence of RE success, the literature survey presented in [15] is often cited. It uses the categories of system quality, information quality, use, user satisfaction, individual impact, and organizational impact to classify works involving information systems success.

Finally, [16] provide a systematic literature review of how RE experiments employ effect sizes when presenting their results. The authors found that 29% of the studied experiments provide effect sizes. They do however not examine what dependent variables that are used by RE experiments.

In summary, while a number of ideas have been presented on how RE activities should be evaluated and assessed, there is only one overview of how evaluative research actually evaluates alternatives (i.e., [9], on requirements elicitation). The review conducted in the present study aims at providing such an overview by describing dependent variables used in experimental RE research and relating them to existing classification schemes. This also facilitates comparisons of the dependent variables involving different RE activities.

III. Review Protocol

The review method used in this study is described in sections III.A to III.E. These sections correspond to the steps of a systematic review listed in [17].

A. Identification of Research

RE is a broad concept related to many aspects of system science and software engineering research. Numerous papers concerning RE have been produced. As of August 5th 2014, the Scopus database held over ten thousand records containing "requirements engineering" in titles, keywords or abstracts. And these do not cover all records of relevance – for example, another 547 records were found that contain only "requirements elicitation". Thus, because of the sheer number of RE papers produced, simply searching scholarly databases for records that match certain keywords is not a viable option for a broad review such as this.

Instead of searching databases directly, this review includes records from previously performed systematic reviews. A systematic review is a structured approach for examining an existing body of literature to answer well-defined research questions; it is presumed to have clear criteria for including research and typically includes studies of a particular phenomenon [17]. Furthermore, systematic reviews are performed when a sizeable number of contributions related to a scientific issue have been made (i.e., when a synthesis is warranted). Thus, it can be expected that sub-domains addressed in systematic reviews are relatively mature and that papers included in systematic reviews are of relatively good quality. Therefore, previously published systematic reviews related to RE can be used to identify empirical studies that focus on relatively mature sub-domains of RE. This is the search method used in the present review. Some issues associated with this (unconventional) search method are discussed in Section VI.C.

To find suitable systematic reviews a search was made during April 2014 in Scopus and Emerald using the keywords "systematic review" and "requirements engineering". This search yielded 49 records describing systematic reviews. Another 16 records were found via the reference lists of these works and through less-structured searches using Google Scholar. These 65 records were reviewed by two reviewers to assess if they (1) reviewed RE research and (2) included empirical studies that measure RE success. The reviewers disagreed regarding 16 cases. Discussions between the reviewers were used to reach consensus for these cases. An example is [18], which address the RE activity of user involvement. There was disagreement regarding the inclusion of [18] since it focuses on identifying the users of medical devices and methods for collecting their feedback rather than RE success. [18] was removed after discussion. In total, 28 systematic reviews were found suitable to use as the basis for this review. These systematic reviews included a total of 915 papers, of which 893 were possible for us to retrieve.

B. Study Selection

To select papers suitable for inclusion in this review, the 893 papers were studied to identify experimental studies directly related to RE. The limitation to include only experimental studies was motivated by the natural role of dependent variables in them (e.g., as opposed to more qualitative case studies). The definition of experiment given by [19] was used. This definition states that an experiment is "a study in which an intervention is deliberately introduced to observe its effects".

After testing the agreement between reviewers' classifications and receiving a modest score (a Cohen's Kappa of 0.41), an inclusive strategy was used. If a reviewer was uncertain if the paper met the criteria, it was included for further analysis during the data-extraction phase. This inclusive strategy resulted in 93 papers. During the data extraction, where information required by the analysis (Section III.D) was also used as an inclusion criterion, 15 papers were removed. Thus, 78 studies were finally included.

C. Study Quality Assessment

The main quality criterion used in this review is the use of an experimental research method, i.e., observation of effects

from intervention. No other quality requirement was used after this during the screening process. However, multiple quality criteria were applied indirectly by using previously published reviews to identify relevant research. First, systematic reviews are typically performed on topics where multiple contributions have been made and a certain maturity can be expected with respect to research methods. Second, the systematic reviews used to identify primary research used quality criteria themselves. For instance, the review described in [20] required that the quality of studies exceeded a minimum threshold.

D. Data Extraction

The following data were originally extracted from each study and entered into a spreadsheet: the paper's title, the RE activity to which it relates (i.e., elicitation, analysis, specification, and/or validation), the domain (e.g., "creativity techniques"), the independent variable(s), the measurement method for independent variable(s), the dependent variable(s), the measurement method for dependent variable(s), how the dependent variables chosen were motivated, and a comment field (e.g., its relationship to other papers). Source material (citations from the articles) was given for data points when deemed necessary (which was usually the case). While independent variables were extracted, analysis of these is outside the scope of the present paper.

Papers that describe several experiments were treated as several independent experiments. For example, [21] describes two experiments and was therefore treated as two independent experiments. Papers that concern several phases (e.g., both elicitation and analysis) had response variables categorized according to the phase that the response variable belong to There were also relationships between studies. For instance, two experiments ([22] and [23]) replicated an experiment in [24]. All replications were included, but when two or more papers described the same experiment only the latest published paper of these was included.

E. Data Synthesis

Data synthesis was carried out as two separate and sequential activities. These activities are described in the following two subsections.

1) Synthesis of Dependent Variables and Measurements

To answer RQ1, the dependent variables and operationalization techniques for the 78 studies were classified. As a first step, the extracted information was employed with the aid of previously published categorizations [9]–[12] to identify a set of categories and states that could facilitate data synthesis. These are given in the bullet point list below. The purpose of this list is not to be holistic, but to reflect relevant properties related to the dependent variables studied.

- **Variable class** concerns the overall type of the dependent variable. For instance, the variable "True defects found out of all defects present" corresponds to the class "Defects found".
- **Measurement** is the technique that is employed to measure the state of the dependent variable. It contains six states:
  - *Answer key*: the researcher uses an answer key to correct given answers, e.g., a gold standard identified by the researchers performing the experiment.
  - *Count*: the number of occurrences of an item is counted by the subject of the study or the researcher. Unlike an answer key, counting does not require the researcher to identify correct answers.
  - *Time*: a mechanism that captures time.
  - *Judgment by expert*: subjective judgment by experts.
  - *Judgment by subject*: subjective judgment by subjects.
  - -: the study does not provide sufficient detail to identify the type of measurement employed.
- **Background scenario** concerns whether the scenario in the experiment is real, fictive, or a combination of both. It contains four states:
  - *Real*: the experiment is based on a real background (e.g., [25] based their work on a requirement specification for a Data Warehouse application produced by the Naval Oceanographic Office in Mississippi).
  - *Fictive*: the experiment is based on a fictive background (e.g., [26] employed 414 use cases developed by students).
  - *Combo*: the experiment is based on a combined real and fictive background (e.g., [1] is based on a real requirements specification (RS) that has had fictive defects injected by the researchers).
  - -: the study does not provide sufficient detail to identify whether its background is real.

The experiments studied were classified into these categories by one reviewer who did this classification iteratively, taking into account comments from the other authors obtained through workshops held between the iterations. To finally test the reliability of the scheme, 10% randomly chosen dependent variables corresponding to the validation phase were independently scored by three other reviewers. The result was more or less uniform, with only small differences between the results. For instance, one reviewer used "Defects" as a class when another used "Requirements defects found" and the third used "True defects found".

2) Synthesis of Motivations for Dependent Variables

The second research question (i.e., why is the success of requirements engineering activities measured as it is?) of this study involves determining why these particular dependent variables are employed. As for the categories related to RQ1, the purpose of this list is not to be holistic, but to reflect interesting properties related to the data studied. The motivations used in the studies were coded as follows:

- **Arguments:** the paper describes some logic or theory to support the choices of dependent variables.
- **References:** the paper cites references for the choices of dependent variables.
- **Alternative measures:** the paper discusses alternative types of dependent variables that could have been used, and why these were not used.
- **Framework:** the paper relates the chosen dependent variables to a framework of existing dependent variables.

One reviewer was responsible for the elicitation and validation phase, one for the analysis phase and one for the specification phase. This work was conducted iteratively in combination with discussions between these three reviewers.

IV. Dependent Variables and Measurements

A total of 78 articles were categorized, from which 298 dependent variables corresponding to 37 classes were elicited. An overview of these variables along with the RE phases they concern is described in Table I. The entire dataset, including all papers studied and their connected literature reviews, is available for download (www.foi.se/res-tqdv).

TABLE I. Overview of Dependent variables

Dependent variable	Elicitation	Analysis	Specification	Validation
Agreement	3	12	1	1
Ambiguity	0	8	0	0
Analysis base	0	1	0	0
Benefits	0	2	0	0
Breadth	5	0	0	0
Budget overrun	2	0	0	0
Completeness	0	3	1	0
Consistency	0	2	0	0
Correctness	13	9	10	0
Defects found	0	2	0	43
Dependencies	1	2	0	0
Depth	2	0	0	0
Ease of use	1	1	0	0
Effort	0	2	0	0
Feasibility	4	0	0	0
Group interaction	8	24	0	0
Managed issues	0	3	0	0
Memory recall	1	0	0	1
Novelty	5	0	0	0
Number of elicited items	31	2	1	0
Prioritization	0	3	0	0
Quality	7	2	2	0
Redundancy	0	1	0	0
Relevance	4	0	0	0
Reliability	1	0	0	2
Satisfaction	6	2	0	2
Structure	0	1	0	0
Suitability	1	0	0	1
Time	4	3	2	9
Traceability	0	1	1	0
Transfer	0	0	0	1
Uncertainty	0	7	0	0
Understandability	4	3	2	7
Usability	0	0	0	4
Usefulness	4	2	0	1
Verifiability	0	1	0	0
Operationalizations	*107*	*99*	*20*	*72*
Unique variables	*20*	*25*	*8*	*11*

A. Elicitation

The elicitation phase is addressed by 27 articles that include a total of 107 variables mapped to 20 different classes. Overall, the most common types of measurements are counting (31% of all cases), subjective judgment by experts (26%), and subjective judgment by subjects (25%). The experimental background is often fictive (63%), but there are experiments using real backgrounds (27%). As seen in Table I, the identified classes of variables are of different levels of abstraction. For instance, *understandability* can be seen as a component of *usefulness*. This is the case because some studies operationalize dependent variables as high-level constructs like "quality" or "usefulness". For example, in [27] it is stated that *"[t]he expert judges rated the quality of each group's solution on a scale from one (poor) to seven (excellent)"* while the variable *memory recall* used in [28] refers to "*the strict memorization of material being presented"*.

The most frequently used dependent variable is the *number of elicited items* (29%), where an "item" may refer to a requirement, goal, use case, event, system function, action, attribute, cultural issue, activity, or threat. This variable is measured by counting the number of items that are identified during an experiment (77%), often using a background scenario made up by the researcher (65%).

The second most common type of variable is *correctness* (12%). Correctness refers to how close results produced by some technique are to the ideal outcome. For instance, [29] operationalized correctness as the accuracy of Data Flow Diagrams, which was measured by judgment by 25 university students. Correctness is often measured by expert judgment (46%) and sometimes by an answer key (23%). The Background scenario of most of these studies is fictive (54%).
The third most common type of variable is *group interaction* (8%). This variable refers to how members of a group interact when performing some requirements elicitation activity. For example, in the experiment described in [30], the researchers used their judgment to rate whether any group members interacted in a dominant, destructive (for the group's performance) manner. The most common means of measuring correctness is by judgment from subjects, i.e., those targeted by the intervention. All experimental background scenarios for this variable are fictive.

There is a rather large distribution over the classes of dependent variables used within the elicitation phase, with at most 29% of the dependent variables corresponding to a single class (the number of elicited items found).

Five overall domains could be extracted for articles concerning RE elicitation. *Cooperation techniques* (11 variables) facilitate improved cooperation within groups of analysts. For example, [27] tested the difference in terms of cooperation during virtual meetings compared to face-to-face meetings. *Creativity techniques* (28 variables) concern methods that facilitate the generation of new ideas. For example, [31] tested whether specifically chosen mental operations help generate new, innovative ideas. *Document elicitation techniques* (29 variables) concern methods that facilitate information gathering from existing documents. For example,

[32] studied a method for automatic extraction of relevant information from legal documents. *Interview techniques* (16 variables) involve methods that help analysts to perform better interviews. For example, [33] tested whether a cognitive interview method incorporating five principles of memory retrieval aids an interviewee's recall ability. *Unknown domains* (4 variables) concern techniques that help analysts to elicit information about domains that are unknown to them. For example, [34] studied a method for training developers regarding domains that are previously unknown to them.

There is a great variance in terms of what to measure within these domains. Experiments involving interview techniques are the most similar with at most 44% of all variables corresponding to the same type (the number of elicited items); experiments involving creativity techniques or unknown domains are the least similar with at most 25% of all variables

B. Analysis

The analysis phase is addressed by 21 articles having a total of 99 variables that could be mapped to 25 classes (see Table I). Overall, the most common types of measurements are counting (45% of all cases) and subjective judgment by subjects (34%). This differs from the elicitation phase, where expert judgment is commonly used. The experimental background scenario is fictive for 67%, and real for 30%, of all variables.

The most common type of variable is *group interaction* (24%), measured by counting subject interactions (79%) or by asking the study's subjects (21%). For example, [35] employed seven dependent variables related to the effectiveness of asynchronous discussions, including the number of posted messages and the number of votes (i.e., counting), while [36] had participants rate each other regarding personal qualities such as trustworthiness and politeness (i.e., judgment by subject). All experimental background scenarios for this variable are fictive.

The second most common type of variable is *agreement* (12%). Agreement is similar to convergent validity and measure if the result from applying a technique agrees with results by another technique, or whether individuals agree on some topic. For instance, [37] operationalized agreement as "*the agreement between the different groups in terms of how they have prioritized the different features*". This category should not be confused with correctness (that concerns some notion of "truth") or group interaction (that concerns how individuals exchange information and behave in a group context). Agreement is typically measured by counting (67%) and always in the context of a fictive background scenario.

The third most common type of variable is *correctness* (9%), which is explained in Section IV.A. Correctness is typically measured by an answer key in the context of a real background scenario.

The synthesis of domains within the analysis phase identified four different domains. *Requirements negotiation techniques* (54 variables) concerns methods that facilitate the negotiation of requirements between different stakeholders during collaborative development of requirements specifications. For example, [38] investigated the use of a web-based meeting system for distributed requirements meetings. *Presentation techniques* (18 variables) concerns methods focused on facilitating the analysis process for the analyst. An example is [39], which describes an investigation of the effect on the performance of the analysts when using UML interaction diagrams. *Requirements prioritization* (22 variables) includes methods that support the process of prioritizing requirements, such as [37] that tested whether the initial order of requirements affects the final priorities. *Software release planning* (5 variables) concerns methods focused on facilitating better communication and knowledge sharing between stakeholders in software development projects. For example, [40] evaluated the use of a release planning method for web application development.

The greatest agreement concerns requirements negotiation techniques where at most 43% of all variables are of the same type (group interaction). The lowest agreement concerns requirements prioritization where at most 19% of all variables are of the same type (correctness).

C. Specification

The specification phase is addressed by ten articles and a total of 20 variables that could be mapped to eight classes. Overall, the most common types of measurements are answer key (50% of all cases) and subjective judgment by experts (30%). This differs from both the elicitation phase and the analysis phase where counting is heavily favored. The experimental background scenario is fictive for 70% of all variables and real for 25% of all variables.

The most common type of variable by far is *correctness* (50%). Here, correctness refers to the proper identification of relevant documents, classified requirements, traceability relations, changes or items. It is typically measured using an answer key (80%). For example, [41] measures it as the recall and precision related to the number of retrieved relevant documents. The background scenarios are evenly split between real and fictive cases.

The second most common types of variables are a three-way tie between *quality*, *understandability* and *time* (10% each). Quality refers to the overall quality of a produced RS and is measured using expert judgment involving a fictive background scenario. Understandability concerns how well users of a prescribed technique comprehend this technique and is measured using an answer key and fictive background scenario. Time is measured using a fictive or combined fictive and real background scenario.

The synthesis of domains within the specification phase identified three different domains. *Enhancing the usage of the RS* (3 variables) concerns how different notations and training affects the ability to comprehend specifications. An example is [42], which addressed the amount of training needed for understanding specifications based on formal notation as well as specifications based on an informal notation. *Decreasing the effort to produce the RS* (9 variables) concerns approaches that support the development of specifications. For example, [43] investigated the effectiveness of object-oriented analysis compared to structure analysis when compiling requirement

specifications. *RS Refinement* (7 variables) includes approaches that, based on existing specifications, support the process of making the specifications more comprehensive. This includes [44], which explored an approach for automatic detection of non-functional requirements.

The agreement is on overall greater for the specification phase than the elicitation or analysis phases, with at most 71% (RS Refinement and the variable correctness), and at worst 56% (Decreasing the effort to produce the RS and the variable correctness), of all variables corresponding to a single type.

D. Validation

The validation phase is addressed by 22 articles having a total of 72 variables within 11 classes. The most common types of measurements are answer keys (60% of all cases) and judgments by the studies' subjects (18%). Thus, this phase is similar to the specification phase with experts' judgments replaced by subjects' judgments. The experimental background scenario is fictive for 57% of all variables, real for 19% of all variables, and a combination of real and fictive for 24% of all variables. This differs from the remainder of the phases where a combination is rarely used.

The most common variable by far is *defects found* (60%). Defects found is conceptually similar to the variable *number of elicited items*, but focusing instead on the number of defects found in a requirements specification. Defects found is in the vast majority (88%) of the studies measured using an answer key over correct and incorrect elements and in approximately half of the studies was measured using a fictive requirements specification with artificial defects. When a combined fictive and real background scenario is employed (16%), fictive defects are injected into a real requirements specification. Whether to use real requirements specifications and defects is a debated topic that does not seem to have a well-defined answer. On the one hand, real specifications with ecologically valid defects (such as the one used in [23]) provide a valid view of the complexities involved in the real world, something that is difficult to manage artificially. On the other hand, a semi-artificial scenario (such as the one used in [1]) or completely artificial (such as the one used in [45]) offers greater control (e.g., in terms of unknown [to the researchers] defects) and could thus yield more reliable tests.

The second most common variable is *time* (13%). Here, time typically concerns the amount of time required to identify defects in a fictive RS. The third most common variable is *understandability* (10%). Understandability is measured using either an answer key or by counting, always in the context of a fictive background scenario.

A total of four different domains could be identified within the validation phase. *Presentation techniques* concern methods that alter the presentation of a RS in a way that is hypothesized to increase its readability and thus the possibility for analysts to spot errors within it. For instance, [21] tested a method that formally describes requirements specifications with the use of special diagrams. *Reading techniques* involve methods that help analyst to better understand a RS without altering it. This includes, for example, [24] which tested the difference between RS inspection based on checklists and those made ad-hoc.

These domains all share then same general viewpoint regarding what to measure – defects found (from 33% for Presentation techniques to 86% for Cooperation techniques). This differs from the other RE phases (especially elicitation and analysis) that have less agreement regarding this matter.

V. Motivation of Dependent Variables

Table II gives the frequencies with which studies have motivated the choice of dependent variable in different ways for the four phases and domains within these phases. More details of motivations used in the four phases are provided in sections V.A through V.D.

TABLE II. Motivation of dependent variables

Domain	***Variables***	***Argumentation***	***References***	***Alternatives***	***Framework***
Cooperation techniques	11	9	9	0	0
Creativity techniques	28	6	20	0	0
Document elicitation techniques	29	8	8	0	0
Interview techniques	16	4	6	0	0
Unknown domains	4	1	1	0	0
Other	19	0	0	0	0
Total Elicitation	*107*	*28*	*44*	*0*	*0*
Presentation techniques	18	9	6	0	3
Requirements negotiation models	54	33	6	0	0
Requirements prioritization	22	3	5	0	0
Software release planning	5	5	0	0	0
Total Analysis	*99*	*50*	*17*	*0*	*3*
Decreasing the effort to produce the RS	9	4	2	0	4
Enhancing the usage of the RS	3	2	2	0	0
Refining the RS	7	0	4	2	0
Unknown domains	1	0	0	0	0
Total Specification	*20*	*6*	*8*	*2*	*4*
Competence of analysts	1	0	0	0	0
Cooperation techniques	14	2	2	0	0
Presentation techniques	21	12	0	1	0
Reading techniques	36	29	29	16	16
Total Validation	*72*	*43*	*31*	*17*	*16*
Total	*298*	*127*	*100*	*19*	*23*

A. Elicitation

In [46] it is stated that "*[t]he true 'quality' of requirements elicited cannot, of course, be determined with any degree of certainty until much later in the system development process (e.g., during design or after implementation), if ever*". This dystopian notion might be a possible explanation for why elicitation studies omit motivating the choice of dependent variables. As seen from the Table II, the motivation is in general unclear for requirements elicitation papers. Not a single variable is motivated with discussions of alternative dependent variables or frameworks, less than half are motivated by citing others to support the choice, and less than a third of the time the selection of variable is motivated with logical arguments. For instance, the papers building on EPMCreate cite the original experiment involving the technique, but do not relate

to any overall framework that relates novelty and feasibility to other useful properties such as correctness and usefulness. In addition, note that the original experiment lacks a clear explanation behind its choice of dependent variable.

B. Analysis

Motivation of dependent variables within the analysis phase differs from the elicitation phase in the sense that variables are motivated to a greater extent by argument (51%) and to a lesser extent by references (17%). Because alternative dependent variables are not discussed, it is, as for the elicitation phase, generally unclear why one variable is chosen instead of another. One article ([47]) motivates its variables using the Bunge Wand Weber framework [48]. This framework is a modeling theory that describes a set of abstract ontological constructs that can be used to model information systems (e.g., thing, property, and system). The relation between this framework and concrete response variables is vague at best.

C. Specification

As seen in Table II, motivation is slightly better for the specification phase than for the elicitation and analysis phases Motivations are commonly done using references (40%) and/or arguments (30%). However, one study motivates its dependent variables by a comparison of alternatives, and one study motivates its dependent variables using a framework over quality attributes for requirements by [49].

D. Validation

As seen in Table II, motivation is in general higher for the validation phase than the elicitation, analysis or specification phases: 60% of the dependent variables are motivated by argument, 43% by references, 24% by discussions of alternative variables and 22% by referencing frameworks. This is especially clear for reading techniques, where 81% of all variables are motivated by argument and/or references, and 44% with alternatives and frameworks. The articles that mention frameworks (namely [50], [51]) refer to three articles:

- [52], which present an experiment that presents a list of fault classes. This work also ended up within this review.
- [53], which present a list of objectives for an RS and provide a test using these objectives for a flight software RS.
- The IEEE Guide to Software Requirements Specifications [54], which contains guidelines for how create a good (software) RS.
- The fault classes within these three articles are used to categorize faults identified during the experiments presented in [50] and [51].

VI. Discussion

The dependent variables and measurement procedures listed in this table represent only about a third (36%) of what is employed by the included studies. Thus, there is little consensus on how to measure the success of RE activities. This chapter first discusses the agreement in terms of dependent variables and measurement procedures. Thereafter, limitations of the present review are discussed.

A. Agreement on the Dependent Variables

While disagreement in terms of how to measure RE success was anticipated due to the complexity of the topic, the results were still rather surprising. In spite of a high abstraction level, 37 classes of dependent variables were identified, and most of the 298 identified measurements of dependent variables differ in one way or another. Thus, disagreement is considerable. This disagreement is to some extent contingent on whether one considers each RE phase as a whole or as the more specific domains they contain; however, it is always present.

Table III summarizes the agreement regarding the different RE phases. One observation is that the agreement increases along the RE process: a study addressing the validation phase introduces 0.5 new classes of dependent variables, while studies dealing with other phases introduce between 0.7 and 1.2. Furthermore, in the validation phase, 60% of all dependent variables correspond to a single class (defects found). In the other phases, the most frequently used variables are less dominant (29% - 50%). That is, when an existing RS should be analyzed in respect to requirement defects, there are some established methods for accomplishing it. For instance, several studies (e.g., [22], [24], [55]) employ the same experimental background scenario and methodology (using two RSs called CRUISE and WLMS), but with different tested independent variables.

TABLE III. Comparison of RE phases

Characteristic	Elicitation	Analysis	Specification	Validation
Novel classes introduced per paper	0.7	1.2	1.2	0.5
Frequency of the most common variable	29%	24%	50%	60%
Provides some motivation of variables	41%	59%	70%	64%

Research on the specification and validation phases is better at motivating their dependent variables than research regarding the elicitation and analysis phases. In particular, there are a number of frameworks describing RS fault types that can be used as a basis for RS inspections and the dependent variable defects found (the framework mentioned by the study involving RE analysis is abstract and not easily translated to dependent variables).

However, if one scratch the surface of dominant variables such as defects found, inconsistencies remain, and further theoretical work is needed. In particular, we found only one study that measured the number of falsely reported defects (using a single variable) [1]. This is rather odd considering that management of falsely reported defects can be a costly endeavor. When [15] studied dependent variables corresponding to the success of information systems, they argued that a great variety of response variables is necessary because each study is unique. Our analysis points to the opposite – it is generally unclear why certain dependent variables are employed before others that also should be possible to employ. For example, it is unclear why tests of the

creativity technique EPMCreate [56]–[59] typically employ novelty and feasibility instead of, for example, correctness or usefulness (cf. Section IV.A). A diversity of the magnitude identified by this review prohibits straightforward synthesis and analysis of experimental results, as exemplified by the complex aggregation rules employed in [9]. It is difficult to see any good reason for a diversity and disagreement of the magnitude present in the choice of dependent variables used in RE research; it is easy to see how standards and theoretical models would improve the utility of future research.

B. Agreement on the Measurement Procedures

There is a large variation in how different dependent variables are measured, especially for the elicitation and analysis phases. To employ an objective measurement (e.g., counting, answer key or time) is arguably a more reliable approach than the use of a subjective judgmental measurement. Subjective measures (i.e., judgment) are common for the elicitation and analysis phases (used in 51% and 37% of the variables). They are less common for the specification and validation phases (used in 30% and 18% of the variables). Studies of the specification and validation phases use an answer key to measure the dependent variable in the majority (50% and 60%) of the included variables. The reason behind this is likely that the use of an answer key requires the researchers to obtain the *ground truth* on the matter, something that the measurement frameworks within the specification and validation phases can be used as a basis to provide.

In regard to the degree of realism provided by the overall context of the experiment, the results are quite similar for all phases: the majority of variables are measured in the context of an experimental background scenario made up by the researchers. While a background scenario based on the real world naturally is more difficult to operationalize in an experiment, it ensures ecological validity. On the positive side of fictive scenarios, they do enable studying properties that otherwise might be difficult with real-world data. For instance, [1] found that trivial defects such as syntactical mistakes receive significant attention during RS reviews. If one wishes to study more complex defects, it is thus likely more effective to provide a fictive scenario without any syntactical mistakes.

A combined fictive and real background is not employed at all in the elicitation or analysis phases and seldom (5%) in the specification phase. The outlier is the validation phase, where it is used for a quarter of all variables. The reason behind this is that some experiments involving requirements validation concern real requirements specifications that are seeded with fictive defects. One reason why seeded defects are used is that it is simpler than finding actual defects (e.g., as in [50]). Another reason is that it (as for completely fictive designs) allows for a greater control over the experiment.

C. Limitations

Possibly the most apparent limitation of this study is that it makes no attempt to identify the best way of measuring the success of RE elicitation, analysis, specification, validation or any of the studied domains within these phases. It simply attempts to understand how these activities are measured by researchers, and why they are measured in these ways. As noted in the discussion above, the authors think that research should be directed towards defining theoretically sound and practically usable experimental dependent-variable protocols.

Apart from the fact that this study makes no attempt to identify the best way of measuring the success, the perhaps largest concern lies with the methodology employed to select studies: records included in previously performed systematic reviews were used instead of directly searching databases for these (as is typically done by systematic reviews). There are apparent issues with this search strategy. First, it purposely limits the review to a few selected topics in the RE domain, and these topics have been selected by others. Second, it is biased in the same way as the original reviews are with respect to search strategies, inclusion criterions, etc. Because these are likely to vary between the reviews, an undesirable variation may exist within the retrieved records. This is confirmed by [8], who identified large variations between reviews that covered the same or similar topics. Third, because the search relies on records established years ago, it does not guarantee that all present knowledge is included; more recent studies may exist. The 78 included articles were published between 1986 and 2011, with the majority (69%) published during the first decade of this millennium.

In spite of these issues, the strategy of using previously performed systematic reviews was judged as the best option compared to alternative search strategies. The use of previously published reviews was judged as a better alternative to drawing a random sample of papers from databases because it yields studies addressing well-defined and mature topics related to RE. Control over the research topics ensures that identified variation is not entirely attributed to different focus in the included studies. Another alternative would be to limit the search to journals, conferences and workshops with an explicit RE focus. This would introduce an undesirable bias because much of RE research is published in general system-science and software-engineering venues.

Another concern of the study lies with the coupling of dependent variables to classes, measurement types and experimental background scenarios (see Section IV). As no holistic framework or guidelines for this purpose was found, this mapping was conducted based primarily on the judgment of the authors. Thus, there is certainly a possibility that some dependent variables have been wrongly classified. It is also likely that there are better, e.g., more stringent and comprehensive, taxonomies of classifying dependent variables in RE research.

VII. Conclusions and Future Work

This study set out to answer two research questions: "*how are requirements engineering activities measured?"* (RQ1) and "*why are these measurements used?"* (RQ2).

Our results regarding RQ1 show that there is an extensive disagreement regarding what to measure, both within and between RE phases and domains of study. The most dominant variables for the four phases are the number of elicited items (the elicitation phase), group interaction (the analysis phase),

correctness (the specification phase) and defects found (the validation phase). The type of measurement and experimental context employed depend on the RE phase in question: works involving RE elicitation, analysis or specification favor subjective, judgmental methods, whereas works involving RE validation favor answer keys.

Our results regarding RQ2 show that it often is unclear why certain measurements are chosen. They also show that when motivation is provided, it often concerns referencing papers that in return do not motivate their own choices. Approximately one-third of the chosen variables are motivated with references to other researchers who have used them, almost half of them are motivated with arguments, and less than one-tenth are drawn from established frameworks. Furthermore, only 6% of the chosen dependent variables were chosen after contemplating and discussing alternative choices.

Our results suggest that future RE research needs to be directed towards development of measurement standards. For example, if the number of elicited items and their correctness is a sound means of measuring the effects of interventions on RE elicitation, then this method could be standardized. One means of accomplishing this could be to find well-conducted studies that provide background scenario information sufficient to enable replication. A good example of this is [24], which has been replicated by several other studies (e.g., [22]). Based on these, on literature on the problem areas and on best-practice guidelines on how to conduct experiments, optimal experimental configurations could be elicited. However, perhaps more important is to identify what the gold standard should be and identify its theoretical limitations. The results from this review could be used as a stepping stone to accomplish this objective.

VIII. Acknowledgments

The authors would like to thank Nina Lewau for her help to retrieve papers and Niklas Hallberg for his assistance in the review of papers.

References

[1] B. Cheng and R. Jeffery, "Comparing inspection strategies for software requirement specifications," in *Proceedings of 1996 Australian Software Engineering Conference*, 1996, pp. 203–211.

[2] N. Hallberg, S. Pilemalm, and T. Timpka, "Quality Driven Requirements Engineering for Development of Crisis Management Systems," *Int. J. Inf. Syst. Cris. Response Manag.*, vol. 4, no. 2, pp. 35–52, 2012.

[3] K. E. Wiegers, *Software Requirements*. Redmond, WA, USA: Microsoft Press, 2003.

[4] B. H. C. Cheng and J. M. Atlee, "Research Directions in Requirements Engineering," in *Future of Software Engineering (FOSE '07)*, 2007, pp. 285–303.

[5] B. Nuseibeh and S. Easterbrook, "Requirements engineering," in *Proceedings of the conference on The future of Software engineering - ICSE '00*, 2000, pp. 35–46.

[6] R. Wieringa and J. M. G. Heerkens, "The methodological soundness of requirements engineering papers: a conceptual framework and two case studies," *Requir. Eng.*, vol. 11, no. 4, pp. 295–307, Jun. 2006.

[7] D. Damian and J. Chisan, "An Empirical Study of the Complex Relationships between Requirements Engineering Processes and Other Processes that Lead to Payoffs in Productivity, Quality, and Risk Management," *IEEE Trans. Softw. Eng.*, vol. 32, no. 7, pp. 433–453, Jul. 2006.

[8] M. Bano, D. Zowghi, and I. Naveed, "Systematic Reviews in Requirements Engineering: A Tertiary Study," in *Fourth International Workshop on Empirical Requirements Engineering (EmpiRE)*, 2014, pp. 9–16.

[9] O. Dieste and N. Juristo, "Systematic review and aggregation of empirical studies on elicitation techniques," *IEEE Trans. Softw. Eng.*, vol. 37, no. 2, pp. 283–304, Mar. 2011.

[10] K. El Emam and N. H. Madhavji, "Measuring the success of requirements engineering processes," in *Proceedings of 1995 IEEE International Symposium on Requirements Engineering (RE'95)*, 1995, pp. 204–211.

[11] K. Emam and N. H. Madhavji, "An instrument for measuring the success of the requirements engineering process in information systems development," *Empir. Softw. Eng.*, vol. 1, no. 3, pp. 201–240, 1996.

[12] E. Kamsties and H. D. Rombach, "A Framework for Evaluating System and Software Requirements Specification Approaches," in *Workshop on Requirements Targeting Software and Systems Engineering*, 1999, pp. 203–222.

[13] D. Williams, T. Hall, and M. Kennedy, "A framework for improving the requirements engineering process management," *Softw. Qual. J.*, vol. 8, no. 2, pp. 133–147, 1999.

[14] S. Beecham, T. Hall, C. Britton, M. Cottee, and A. Rainer, "Using an expert panel to validate a requirements process improvement model," *J. Syst. Softw.*, vol. 76, no. 3, pp. 251–275, Jun. 2005.

[15] W. H. DeLone and E. R. McLean, "Information Systems Success: The Quest for the Dependent Variable," *Inf. Syst. Res.*, vol. 3, no. 1, pp. 60–95, Mar. 1992.

[16] V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. Sjøberg, "A systematic review of effect size in software engineering experiments," *Inf. Softw. Technol.*, vol. 49, no. 11, pp. 1073–1086, 2007.

[17] B. Kitchenham, "Procedures for performing systematic reviews," Keele, UK, 2004.

[18] S. G. S. Shah and I. Robinson, "User involvement in healthcare technology development and assessment: Structured literature review," *Int. J. Health Care Qual. Assur.*, vol. 19, no. 6, pp. 500–515, 2006.

[19] W. R. . Shadish, T. D. Cook, and D. T. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin, 2002.

[20] J. Li, H. Zhang, L. Zhu, R. Jeffery, Q. Wang, and M. Li, "Preliminary results of a systematic review on requirements evolution," *IET Semin. Dig.*, vol. 2012, pp. 12–21, 2012.

[21] K. Takahashi, A. Oka, S. Yamamoto, and S. Isoda, "A comparative study of structured and text-oriented analysis and design methodologies," *J. Syst. Softw.*, vol. 28, no. 1, pp. 69–75, 1995.

[22] K. Sandahl, O. Blomkvist, J. Karlsson, C. Krysander, M. Lindvall, and N. Ohlsson, "An Extended Replication of an Experiment for Assessing Methods for Software Requirements Inspections," *Empir. Softw. Eng.*, vol. 3, no. 4, pp. 1382–3256, 1998.

[23] F. Lanubile and G. Visaggio, "Assessing Defect Detection Methods for Software Requirement Inspection Through External Replication," Bari, 1996.

[24] A. A. Porter, Lawrence L. G. Jr., and V. R. Basili, "Comparing detection methods for software requirements inspections: A replicated experiment," *IEEE Trans. Softw. Eng.*, vol. 21, no. 6, pp. 563–575, Jun. 1995.

[25] G. S. Walia, J. C. Carver, and T. Philip, "Requirement Error Abstraction and Classification: A Control Group Replicated Study," in *The 18th IEEE International Symposium on Software Reliability (ISSRE '07)*, 2007, pp. 71–80.

[26] B. Bernardez, M. Genero, A. Duran, and M. Toro, “A Controlled Experiment for Evaluating a Metric-Based Reading Technique for Requirements Inspection,” in *Proceedings of the Software Metrics, 10th International Symposium*, 2004, pp. 257–268.

[27] R. Ocker, J. Fjermestad, S. R. Hiltz, and K. Johnson, “Effects of Four Modes of Group Communication on the Outcomes of Software Requirements Determination,” *J. Manag. Inf. Syst.*, vol. 15, no. 1, pp. 99–118, 1998.

[28] A. Gemino, “Empirical Comparisons of Animation and Narration in Requirements Validation,” *Requir. Eng.*, vol. 9, no. 3, pp. 153–168, Aug. 2004.

[29] L. A. Freeman, “The effects of concept maps on requirements elicitation and system models during information systems development,” in *Concept Maps: Theory, Methodology, Technology. Proceedings of theFirst International Conference on Concept Mapping*, 2004.

[30] E. W. Duggan, “Generating Systems Requirements with Facilitated Group Techniques,” *Hum.-Comput. Interact.*, vol. 18, no. 4, pp. 373–394, Dec. 2003.

[31] S. El-Sharkawy and K. Schmid, “A Heuristic Approach for Supporting Product Innovation in Requirements Engineering: A Controlled Experiment,” in *Requirements Engineering: Foundation for Software Quality*, 2011, vol. 6606, pp. 78–93.

[32] T. Breaux, “Exercising Due Diligence in Legal Requirements Acquisition: A Tool-supported, Frame-Based Approach,” in *2009 17th IEEE International Requirements Engineering Conference*, 2009, pp. 225–230.

[33] J. W. Moody, J. E. Blanton, and P. H. Cheney, “A Theoretically Grounded Approach to Assist Memory Recall During Information Requirements Determination,” *J. Manag. Inf. Syst.*, vol. 15, no. 1, pp. 99–118, 1998.

[34] K. M. de Oliveira, F. Zlot, A. R. Rocha, G. H. Travassos, C. Galotta, and C. S. de Menezes, “Domain-oriented software development environment,” *J. Syst. Softw.*, vol. 72, no. 2, pp. 145–161, Jul. 2004.

[35] D. Damian, F. Lanubile, and T. Mallardo, “An Empirical Study of the Impact of Asynchronous Discussions on Remote Synchronous Requirements Meetings,” in *Proceedings of 9th International Conference, FASE 2006, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS*, 2006.

[36] B. R. Gaines, M. L. G. Shaw, A. Eberlein, and D. E. H. Damian, “Using different communication media in requirements negotiation,” *IEEE Softw.*, vol. 17, no. 3, pp. 28–36, 2000.

[37] M. Svahnberg and A. Karasira, “A Study on the Importance of Order in Requirements Prioritisation,” in *2009 Third International Workshop on Software Product Management*, 2009, pp. 35–41.

[38] D. Damian, “An Empirical Study of a Multimedia Group Support System for Distributed Software Requirements Meetings,” *e-Service J.*, vol. 1, no. 3, pp. 43–60, 2002.

[39] J. Swan, T. Barker, C. Britton, and M. Kutar, “An empirical study of factors that affect user performance when using uml interaction diagrams,” in *Empirical Software Engineering, 2005. 2005 International Symposium on*, 2005, p. 10–pp.

[40] S. Ziemer and I. C. Calori, “An experiment with a release planning method for web application development,” in *Software Process Improvement*, Springer, 2007, pp. 106–117.

[41] A. Udomchaiporn, N. Prompoon, and P. Kanongchaiyos, “Software Requirements Retrieval Using Use Case Terms and Structure Similarity Computation,” in *2006 13th Asia Pacific Software Engineering Conference (APSEC’06)*, 2006, pp. 113–120.

[42] D. Carew, C. Exton, and J. Buckley, “An empirical investigation of the comprehensibility of requirements specifications,” in *Empirical Software Engineering, 2005. 2005 International Symposium on*, 2005, p. 10–pp.

[43] E. R. Sim, G. Forgionne, and B. Nag, “An experimental investigation into the effectiveness of OOA for specifying requirements,” *Requir. Eng.*, vol. 5, no. 4, pp. 199–207, 2000.

[44] J. Cleland-Huang, R. Settimi, X. Zou, and P. Solc, “Automated classification of non-functional requirements,” *Requir. Eng.*, vol. 12, no. 2, pp. 103–120, 2007.

[45] A. Bianchi, F. Lanubile, and G. Visaggio, “A controlled experiment to assess the effectiveness of inspection meetings,” in *Proceedings Seventh International Software Metrics Symposium*, 2001, pp. 42–50.

[46] M. G. Pitts and G. J. Browne, “Stopping Behavior of Systems Analysts During Information Requirements Elicitation,” *J. Manag. Inf. Syst.*, vol. 21, no. 1, pp. 203–226, 2004.

[47] A. Bajaj, “The effect of the number of concepts on the readability of schemas: an empirical study with data models,” *Requir. Eng.*, vol. 9, no. 4, pp. 261–270, 2004.

[48] Y. Wand and R. Weber, “On the deep structure of information systems,” *Inf. Syst. J.*, vol. 5, no. 3, pp. 203–223, 1995.

[49] A. M. Davis, *Software requirements: objects, functions, and states*. Prentice-Hall, Inc., 1993.

[50] F. Lanubile, F. Shull, and V. R. Basili, “Experimenting with error abstraction in requirements documents,” in *Proceedings Fifth International Software Metrics Symposium. Metrics (Cat. No.98TB100262)*, 1998, pp. 114–121.

[51] T. Berling and T. Thelin, “A case study of reading techniques in a software company,” in *Empirical Software Engineering, 2004. ISESE’04. Proceedings. 2004 International Symposium on*, 2004, pp. 229–238.

[52] G. M. Schneider, J. Martin, and W.-T. Tsai, “An experimental study of fault detection in user requirements documents,” *ACM Trans. Softw. Eng. Methodol.*, vol. 1, no. 2, pp. 188–204, 1992.

[53] V. R. Basili and D. M. Weiss, “Evaluation of a software requirements document by analysis of change data,” in *Proceedings of the 5th international conference on Software engineering*, 1981, pp. 314–323.

[54] ANSI/IEEE, “IEEE Guide to Software Requirements Specifications. Standard Std 830-1984,” 1984.

[55] F. Lanubile and G. Visaggio, “Assessing defect detection methods for software requirements inspections through external replication,” Bari, 1996.

[56] L. Mich, C. Anesi, and D. M. Berry, “Applying a pragmatics-based creativity-fostering technique to requirements elicitation,” *Requir. Eng.*, vol. 10, no. 4, pp. 262–275, 2005.

[57] L. Mich, M. Franch, and D. M. Berry, “Classifying web-application requirement ideas generated using creativity fostering techniques according to a quality model for web applications,” in *Proceedings of the 12th International Workshop on Requirements Engineering: Foundation for Software Quality, Luxembourg*, 2006.

[58] L. Mich, C. Anesi, and D. M. Berry, “Requirements engineering and creativity: An innovative approach based on a model of the pragmatics of communication,” in *in Proceedings of Requirements Engineering: Foundation of Software Quality REFSQ’04*, 2004.

[59] V. Sakhnini, D. M. Berry, and L. Mich, “Validation of the effectiveness of an optimized EPMcreate as an aid for creative requirements elicitation,” in *Requirements Engineering: Foundation for Software Quality*, Springer, 2010, pp. 91–105.