Software Quality Journal
https://doi.org/10.1007/s11219-020-09521-2

What lies behind requirements? A quality assessment of statement grounds in requirements elicitation

Corentin Burnay[1] · Sarah Bouraga[1] · Joseph Gillain[1] · Ivan J. Jureta[1,2]

Published online: 05 September 2020

Abstract
In requirements engineering (RE), an early yet critical activity consists in eliciting the requirements from various stakeholders, who usually have different assumptions, knowledge, and intentions. The goal during elicitation is to understand what stakeholders expect from a given software, expectations which then feed the analysis, prioritization, validation, and ultimately specification activities of the RE process. Elicitation is an interactive activity. It relies on verbal communication of statements of stakeholders about their requirements, their ideas, their assumptions, the constraints they know apply in the environment of the future software, and so forth. Statements, we claim, build either on a past experience of the stakeholder or are the result of reasoning from indirect experience, i.e., they have different grounds. In this paper, we introduce the concept of "Statement Ground" during RE, contrast it with the classical perspective on requirements elicitation, and position the concept in existing RE literature. We conduct an empirical assessment of the relative qualities of statements that have different grounds. Our work results in a better understanding of the statements produced by stakeholders during requirements elicitation, of their qualities, and of the interplay between those qualities and the concept of statement ground. It also results in the definition of a series of research questions which focus on the implications of our findings on the overall requirements engineering activity.

Keywords Requirement engineering · Elicitation · Stakeholders communication · Statement · Ground · Experience · Hypothetical statement

1 Introduction

1.1 Statements in requirements elicitation

Requirements elicitation, or simply *elicitation*, designates all activities conducted during requirements engineering (RE) in order to collect information from stakeholders about their

✉ Corentin Burnay
corentin.burnay@unamur.be

1 Department of Business Administration NADI Research Institute, PReCISE Research Center, University of Namur, Namur, Belgium

2 Fonds de la Recherche Scientifique - FNRS, University of Namur, Namur, Belgium

requirements toward a software-to-be and about the environment in which that software is supposed to operate (Nuseibeh and Easterbrook 2000). It is an interactive process, relying on verbal or written *statements* made by stakeholders to requirements engineers, who are in charge of summarizing the information and ultimately of producing the specification of a software that complies with those expectations. Techniques to elicit requirements efficiently are numerous (Zowghi and Coulin 2005). They include approaches as varied as interviews, surveys, questionnaires, brainstorming sessions, group meetings, ethnography, and so on. Elicitation techniques are tools used to collect statements from stakeholders of a software.

In this paper, the term *statement* is the representation of a piece of information, which an engineer judges to be sufficiently self-contained, so that it can be distinguished from other statements. A statement can be, for example, a paragraph describing an idea discussed during an interview, the written answer of a stakeholder to one particular survey question, drawings from a storyboard, a mind map, some notes, photos and other records that all seem to focus on one idea from a brainstorming session, etc. (Burnay 2016). It is important to distinguish a statement from a requirement, a goal, or any other RE entity in literature (Pohl 2010). Statements are what is communicated by stakeholders. They are informal by nature and have little "contractual" value, in the sense that they still require further discussion. The reader should bear in mind, for now, the informal aspect of a statement. RE entities on the other hand are structured; they are defined by an engineer who concluded that the RE entity was relevant for the rest of the RE process. As such, they therefore clearly differ from statements. Typically, an engineer will document an RE entity formally if several statements justify that RE entity, and enable to have a sufficiently clear view on it. In other words, statements form the baseline for the identification of various RE entities (goals, requirements, scenarios, domain assumptions, ...) which are used subsequently in the RE process. Table 1 summarizes this discussion with a quick comparison between statements and RE entities.

Note that we see no universal rules for mapping statements to instances of various RE entities; different RE ontologies and methods each use their own rules. Regardless of the specifics of these mappings, we consider that in all cases, elicitation statements are relevant to study when defining and formalizing RE entities. Basically, we get instances of RE entities by analyzing statements.

1.2 Quality of elicitation statements

Statements identified during elicitation are an input to downstream RE activities such as analysis, prioritization, or specification. It follows that the identification of quality statements—as a way to detect quality RE entities later on—is a key concern during elicitation. There is considerable research on how to collect more statements; various techniques and methods have been suggested to collect more systematically information about the domain and requirements (Zowghi and Coulin 2005), to avoid missing important questions

Table 1 Comparison of statements vs RE entities

	Statements	RE entities
Produced by	Stakeholders	Engineers
Degree of formalization	Low	Medium to high
RE step concerned with its identification	Elicitation	Analysis
Granularity	Atomic	Composite

during interviews (Burnay et al. 2014), to detect implicit or tacit information (Goguen and Linde 1993; Sutcliffe and Sawyer 2013), to collect novel/creative requirements (Maiden et al. 2004), or to select the most appropriate technique to elicit some particular content (Maiden and Rugg 1996; Hickey and Davis 2004). These approaches all focus primarily on the *quantity* of statements.

More, however, is not always better. Engineers may end up collecting a large set of statements, the quality of which is likely to be highly heterogeneous. Typically, problems of quality are managed by requirements engineers when RE entities are already identified and represented, during the next steps of the RE process, with mechanisms such as quality gateways (Robertson and Robertson 2012) or requirements quality indicators (Gėnova et al. 2013) to ensure the resulting specification is sufficiently good. In these approaches, statements are used regardless of their intrinsic quality to model RE entities, some of which are then withdrawn/revised during the validation step. We see here room for improvement; the many iterations, modeling and validation efforts, and the numerous interactions between stakeholders of the project represent many hours spent by the engineers. Our intuition in this paper is that it should be possible to reduce the time it takes to conduct RE by accounting *earlier* in the process for the *quality* of collected statements.

1.3 Grounds of elicitation statements

Of course, quality as such is difficult to observe or experience during elicitation, if observable at all. We therefore need a proxy to approach the notion of statement quality. This proxy should be simple, given the very early stage of the RE process this paper focuses on; it is indeed hardly feasible to compute a complex set of quality indicators for each statement collected during elicitation. An idea that appeared appealing to us when exploring the concept of statement was that it has a *ground*, which can be easily identified and which influences its quality, at least partially. By ground, we mean "the foundation," or more specifically "the underlying rules (either experience or speculation) that were used by stakeholders to share the Statement." Our premise is that two statements will likely have different grounds, and different grounds might provide information of various quality. Consider the following example to clarify our claim. A stakeholder shares a statement with an engineer during an elicitation interview, following the different recommendations from RE literature. There are two possible grounds for that statement. The stakeholder may describe something she actually experienced in the past; she knows and used a system that fulfills what she was asking for. In that case, we say the statement builds on experience, and we call it *experiential statement*. Alternatively, the statement may be the result of extrapolation from related experience, and/or from any other thinking about things and events which the stakeholders did not experience in fact. For example, the statement may reflect a functionality that the stakeholder never experienced, but heard of, or read about. She may however still share it, because she observed something similar in another context, and found it useful. In that case, we say that the statement builds on hypotheses, and we call it *hypothetical statement*. Our first contribution in this paper lies in the conceptualization of the concept of statement ground and the discussion of this concept in light of RE, software quality, and decision-making literature.

1.4 Qualities of statement grounds

In RE, a lot of attention has been paid to the qualities RE entities should have in order to minimize the risk of flaws during the process (Gėnova et al. 2013). In Christel and Kang

(1992) for instance, it is claimed that a major challenge during elicitation is to obtain information that is *understandable*, *stable* over time, and delimited by a clear *scope*. Similarly, engineers are often invited to stimulate stakeholders' during elicitation (Maiden et al. 2004), to provide *creative* information, in order to find novel and valuable solutions to long-term problems (Boden 2013). A last example is the need to rationalize the allocation of resources and therefore elicit *prioritized* requirements (Karlsson 1996). Just like RE entities, we consider statements have some properties that should matter to requirements engineers, values of which differ depending on the ground of that statement. This line of thinking about experiential and hypothetical statements lead us to the definition of a series of statement qualities that we study in the present paper:

- **Exhaustive**: how much relevant information a stakeholder will provide when sharing a statement during requirements elicitation?
- **Steady**: how likely is it that the stakeholder decides to change the statement in the future, after having communicated it?
- **Objective**: will another person understand the statement in the same way as the stakeholder who initially shared it?
- **Creative**: how novel, surprising and valuable is the statement compared with common knowledge?
- **Orderable**: how the statement can be ordered, prioritized or other-wisely ranked compared with other statements shared by the stakeholder?

A second contribution of this paper is to investigate empirically the difference in qualities between statements with different grounds along those five quality dimensions. It is important to distinguish those qualities from qualities of RE entities in general. A statement may be exhaustive, which does not necessarily suffice to ensure that the related RE entities will be exhaustive as well. Instead, we see a necessary condition; it is hardly feasible to have an exhaustive list of RE entities if all underlying statements are not exhaustive. We discuss this point with more details in Section 3.

We should point out that this line of thinking considers only one argument about RE entities, namely that RE entities should have multiple qualities in order to minimize the risk of flaws later in the process. However, vagueness and ambiguity have also been recognized as potential resources in RE. For instance, Ferrari et al. 2015 analyzed the role of ambiguity in the early stages of RE and, more specifically, discussed its role in disclosing tacit knowledge during interviews. In this paper, we purposefully focus on how well the quality of a statement can predict the quality of the related requirements, even if we are aware that this is probably a slightly too narrow approach. In future work, we should broaden the research to account for the other argument, more specifically, we should study the predictive power of "poor quality" statements to forecast the quality of requirements.

1.5 Organization

The rest of this paper is organized as follows. In Section 2, we discuss in greater details the distinction we make between experiential and hypothetical grounds. We review works in artificial intelligence and decision-making to justify that theoretical distinction. We also position this work in the field of requirements engineering. In Section 3, we present our research methodology; we define with more details the statement qualities introduced earlier, we define our research hypotheses and present the two experiments we designed in order to assess those qualities. We then discuss our results in Section 4, detail the possible practical implications and future works in Section 5, and finally put our work in perspective in Section 6.

2 Theoretical background

As a reminder, a statement is the recording of something expressed by a stakeholder, regardless of its potential to generate or not an RE entity; it is semantically close to the concept of speech act and utterances as defined in Searle (1969), or to the notion of utterance as discussed in Sutcliffe and Ryan (1998). In the core ontology (Jureta et al. 2008), Jureta et al.'s definition of utterance helps us clarify the concept; "utterances are actions intended to advance stakeholders' personal desires, intentions, beliefs, and attitudes, in the aim of ensuring that the engineer can produce a specification that then leads to the system responsive to the communicated concerns." We define the *ground* of a statement as "the experience or speculation which a stakeholder would use to justify the Statement." Any statement has a ground, regardless of its form: a sentence from an interview, an item checked in a survey form, a box suggested in a flow diagram, an idea in a brainstorming session, etc. We differentiate between two different types of grounds; *hypothetical grounds*, arising from assumptions and speculations of the stakeholder, and *experiential statements*, drawn from past experience of the stakeholder. This distinction is not a trivial choice and originates from theories in decision-making and artificial intelligence, where the very idea of statement ground and of justification is critical.

2.1 Experiential vs hypothetical statements

Experiential is understood in this paper as resulting from the "accumulation of knowledge or skill that results from direct participation in events or activities" (Princeton University: About WordNet 2010). *Experiential statements* (ES) then become the statements that are shared by a stakeholder which reflect her past experience, that is, the accumulation of knowledge or skill that results from her direct use of a software which verifies that statement. For instance, a stakeholder may ask for a search button in her future e-commerce web-page, because she already used Amazon and found it useful to search for some specific items. Experience may be either *direct* (ES+) when the stakeholder already used that functionality on that particular type of system (e.g., using Amazon, which is a e-commerce web-page) or *indirect* (ES-) when she experienced that functionality on another system (e.g., using the search bar in her office CRM, she experiences the functionality but in a totally different context). We clarify this difference later in Section 2.2.

Hypothetical on the other hand is understood as resulting from "opinions or ideas based on incomplete evidence" (Princeton University: About WordNet 2010). *Hypothetical statements* (HS) are the statements shared by a stakeholder based on hypotheses/assumptions instead of an experience with a similar system. For instance, the stakeholder may also request that her future e-commerce web-page identifies customers based on fingerprint and not a password. If she never experienced such a thing before, this will be considered as a hypothetical statement. Note that it does not mean the functionality does not exist; it simply means that the stakeholder never used it before sharing her statement. This brings us to the taxonomy of statement grounds as depicted in Fig. 1. One important remark is that the ground of a statement is subjective. A same statement could be hypothetical if shared by a stakeholder A, yet experiential if collected from a stakeholder B. The person at the origin of a statement is the one that defines its ground.

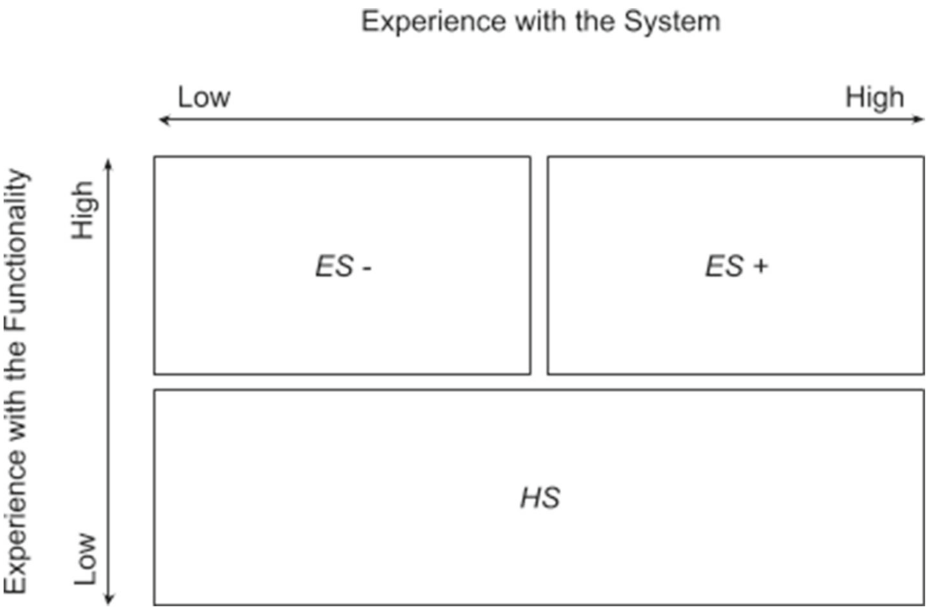

Fig. 1 Taxonomy of statements' grounds

2.2 Grounds in non-monotonic reasoning

The choice to work with ES-, ES+, and HS grounds is not trivial; it echoes a longstanding distinction of information between defaults and knowledge, as discussed in non-monotonic reasoning research, a subfield of artificial intelligence. Non-monotonic reasoning (NMR) research (Brewka et al. 2007) attempts to model reasoning where:

- Agents who do the reasoning (i.e., who makes the decision) face uncertainty;
- Agents make assumptions in order to make decisions in an uncertain environment;
- Agents can revise their initial conclusion anytime something perceived as uncertain actually happens, and new information arises.

It seems reasonable to assume that these conditions to NMR are verified in many human decisions settings, including the context of requirements elicitation where stakeholders have to decide which statements to share with requirements engineers. Although this remains an assumption, we observe that authors in psychology recognize how widespread non-monotonic logic really is, also observing its applicability to widely divergent fields (Pelletier and Elio 2005). This comforts us in the adoption of this NMR assumption in RE. Theories developed in artificial intelligence to formalize how NMR works are numerous; auto-epistemic logic, circumscription, or default logic are some examples (Reiter 1980; Moore 1984; McCarthy 1980; Brewka et al. 2007). In this paper, we use the default logic (Reiter 1980; Poole 1988) and therefore establish a distinction between defaults and facts in the context of RE and elicitation. We adopt default logic for the following reasons:

- Default logic has already found many applications in RE (see Greenspan et al. 1994, Zowghi et al. 1996, Zowghi and Offen 1997, Antoniou 1998, Billington et al. 2011).

– It entails many concepts that fit the problem we treat in this paper, i.e., a stakeholder distinguishes facts from defaults, or equivalently, facts from assumptions, when she decides which statement to share with the engineers during elicitation.

Default logic offers a relevant platform to discuss the concept of ground. A central concept in default logic is the *default theory*, according to which a person who has to make a decision under uncertainty will rely on some knowledge background she has combined with some default rules. The default rules are used by that person to fill in the gaps in her knowledge background. A default rule is very straightforward to define; for instance, a person may decide to believe something "by default" of any piece of knowledge in her knowledge background that would contradict that belief. Sometimes, it may happen that the default rule is invalidated by some new piece of information entering the knowledge background; this implies that default rules are defeasible (Reiter 1980).

While default logic is much more than this, this simple distinction between knowledge background and default rules is enough to justify the notion of ground in this paper. In our view, any time a stakeholder will share a statement in uncertain settings (that is, virtually all the time), we may use default logic to explain how she produced that statement; either (i) she resorted to her knowledge background, in which case the resulting statement is considered to be an experiential statement, or (ii) she resorted to a set of default rules, in which case the resulting statement is considered as a hypothetical statement.

2.3 Grounds in epistemology

Besides AI and default logic, *epistemology* offers another way to justify the use of experiential and hypothetical grounds in RE. This section builds on Steup's definition of experience (Steup 2014), and clarifies the notion of experience of a stakeholder with a system. In Steup's view, experience is a set of *justified true beliefs*, i.e., a statement can be considered to be experiential if and only if it is "Believed," "True," and "Justified" (Steup 2014). These three conditions "are individually necessary and jointly sufficient for knowledge."

– **Belief** refers to the "conviction of the truth of some statement [...] when based on examination of evidence" [?]. For a statement to be experiential, it is necessary but not sufficient that the stakeholder considers this statement to be possible, and is convinced of the relevance of that statement based on evidences at hand;
– **True belief** means that a belief has to be in accordance with the actual state of affairs. A statement which in fact is not true cannot be considered as experiential. For instance, a statement like "I must be able to search a customer based on his DNA," even if sincerely believed in by a stakeholder, could not be an experiential statement because it is not true (i.e., verified) in the state of affairs of the system;
– **Justified true belief** means that there must be a valid proof or justification, which supports the true belief. A stakeholder has a justified true belief, and hence an experiential statement, whenever there is no obligation to refrain from believing that thing is true. This is also the essence of the default theory discussed earlier; a stakeholder can believe that a system is secure, because there are no elements at hand that forces her to believe otherwise.

Now, notice how previous epistemological criteria offer additional ways to compare default rules and knowledge background in the default theory; elements within the knowledge background are justified true beliefs (knowledge), while elements within the default rules are simply beliefs (by definition of a default rule), which may not be true and which are not justified, i.e., defaults may turn out to be false as new information arise, and are therefore said to be defeasible. Steup also identifies several possible sources of justification (Steup 2014). We summarize these sources in our work, because they represent the different ways for an engineer to detect whether a stakeholder produces premises which are justified, i.e., if the stakeholder relies on her experience and hence shares experiential statements. In other words, in the absence of the following justifications, we consider a stakeholder will share statements that are hypothetical:

– Perception: refers to our five senses: sight, touch, hearing, smelling, and tasting. For example, a stakeholder knows the functionalities of a system because she saw the latter working, she touched the keyboard and the mouse when using it, and she heard noises produced by the system;
– Introspection: refers to the activity of inspecting the "inside of one's mind" (Steup 2014). For example, a stakeholder knows a system is not user-friendly, because it introspectively seemed that it was not easy to use and unintuitive;
– Memory: refers to the capacity of retaining knowledge acquired in the past. For example, a stakeholder knows a system has hourly back-ups because she remembers she shared that requirement some time ago;
– Reason: refers to knowledge derived from conceptual truths, mathematics, logic, etc. For example, a stakeholder knows a cloud system cannot work correctly without a connection to the Internet, because it is a conceptual truth;
– Testimony: refers to the acquisition of knowledge through other people. For example, a stakeholder knows a system is too slow, because she remembers many of her colleagues complaining about that.

At this point, it seems also interesting to relate our discussion to the question of argumentation, understood here as a form of reasoning that makes explicit the reasons for the conclusions that are drawn and how conflicts between reasons are resolved (Rahwan and Simari 2009). The use of argumentation has been discussed in multiple ways in RE (e.g., Ionita et al. (2014), Elrakaiby et al. (2017), Van Zee et al. (2015), or Yu et al. (2011)), and offers a complementary platform to discuss our proposition. Ionita et al. (2014) for instance propose an argumentation-based tool to support the elicitation of critical requirements, such as security requirements. They show that using an argumentation-based tool during elicitation can help uncover and refine more systematically and rigorously such critical requirements. For each argument, experts had to provide three elements: (i) a claim, (ii) one or more assumptions, (iii) one or more facts. Based on this rationalization, experts could accept the argument or not. By extension, the hypothetical statements discussed in this paper could also be handled using argumentation. During a RE effort, engineers could ask stakeholders to give a clear argumentation about their hypothetical statements in order to reduce the risk related to the hypothetical nature of the statement. This however remains outside the scope of this paper.

2.4 Statements and grounds in requirements engineering

It is very clear in RE literature that RE entities are not collected as such from stakeholders during an interview, i.e., simply asking to stakeholders what they expect is not a proper elicitation process. RE has come with a variety of techniques and approaches to make that process more reliable, and ensure collect as much information as possible. Frameworks exist to select the right technique in order to collect statements (Maiden and Rugg 1996; Hickey and Davis 2004). Engineers have the choice among a wide range of techniques (Zowghi and Coulin 2005; Davis et al. 2006; Goguen and Linde 1993), some focusing on the collection of statements from stakeholders (the ones that matter to us in this paper) like workshops, surveys, interviews, and others focusing on the collection of statements from the environment, like observation or ethnography. By applying those various techniques, engineers are capable of collecting information in order to define several RE entities (even in a preliminary form). Less attention has been paid in those techniques to the distinction between statements to RE entities; the closest contribution on the topic we found is in Rolland et al. (1999), where requirements chunks are gathered in order to produce more formal RE entities such as goals. In most cases however, the first representation of information is an instance of an RE entity, not the statement shared by the engineer.

Despite an agreement that stakeholders do not communicate requirements during elicitation, there is surprisingly little attention being paid to what stakeholders do actually share. Some references to the concept of statement exist, sometimes called requirements statements (Gotel and Finkelstein 1994), requirements utterance (Saeki et al. 1996), or speech act (Saeki et al. 1996; Jureta et al. 2008). We find however no research focusing solely on the study of the statements, of their characteristics, and of the qualities.

3 Empirical evaluation of statement grounds

We discussed thoroughly the theoretical difference between grounds of statements in Section 2. We now seek to provide evidence that the ground of a statement is something to account for in RE, from a practical point of view. Stated differently, we try to answer the following question: *For what reasons should requirements engineers distinguish between statements depending on their grounds?* To answer that question, we conduct two complementary empirical studies, under the form of two related experiments. Our goal is to clarify if yes or no, experiential (ES+ and ES-) and hypothetical statements do differ, and if that difference matters to RE in general. The experiments are still exploratory, given the early nature of this research. We are therefore cautious about the way our conclusions could be generalized. We consider however that this first series of empirical assessments offer enough indications to confirm the relevance of accounting for statement grounds.

3.1 Experiment 1—quality perception by stakeholders

The first experiment focuses on the stakeholder side. It intends to collect a set of statements, to identify the grounds for those statements, and to assess the perception of qualities from a stakeholder point of view. We proceed in four steps; firstly, we discuss the properties that we use to actually compare the three types of statements grounds, i.e., the variables we observe (independent variables) to quantify the difference between grounds (dependent variable). Secondly, we explain how these variables have been observed in this experiment. Thirdly, we provide details about the experiment itself, and how it was organized. Fourthly, we present our sample of subjects.

3.1.1 Definition of variables and hypotheses

We use the RE literature to identify a set of statement properties relevant to requirements engineers. First, we observe that the various activities in the RE process—elicitation, analysis, prioritization, validation, to cite only a few—all focus on one specific problem

(Nuseibeh and Easterbrook 2000; Pohl 2010). Elicitation for instance tries to gather as much information as possible, prioritization puts an order of priority on requirements, validation is an attempt to make sure documented requirements have been correctly understood, etc. Each step is therefore a source of candidate quality variables to include in our study. Second, we observe that specific attention has been paid to the qualities that RE entities should have in order to minimize the risk of flaws during the process. In Christel and Kang (1992) for instance, it is claimed that a major challenge during elicitation is to obtain information that is *understandable*, *stable* over time, and delimited by a clear *scope*. Similarly, engineers are often invited to stimulate stakeholders during elicitation (Maiden et al. 2004), to provide *creative* information, in order to find novel solutions to long-term problems. A last example is the need to rationalize the allocation of resources and therefore elicit *prioritized* requirements (Karlsson 1996). Just like RE entities, we consider elicitation statements have some properties that should matter to requirements engineers, values of which differ depending on the ground of that statement. This line of thinking about experiential and hypothetical statements lead us to the following list of elicitation statements qualities/properties, and to a set of hypotheses that we want to test. Table 2 lists those null hypotheses that we expect to reject; we hope to demonstrate that there is a statistically significant difference between grounds, for those different statements properties:

- **Creativity**: the extent to which a statement will be seen as novel, valuable, and disruptive. A statement can range from *creative*—it is perceived as novel compared with the usual way things are done—to *conservative*—it is not perceived as something disruptive compared with the standard way of doing things;
- **Objectivity**: the extent to which a statement reflects what is actually intended by stakeholders, without any bias. A statement can range from *transparent*—the statement will be understood in the same way by different people—to *oblivious*—it is unclear if different people will understand the statement in the same way;
- **Exhaustivity**: the extent to which a statement contains relevant information. A statement can range from *exhaustive*—all relevant information is included in the statement—to *superficial*—there may be information missing or remaining implicit;

Table 2 Hypotheses on the perception by stakeholders of HS, ES-, and ES+ statements

How do stakeholders perceive the quality of statements with different grounds?
$H_0^1 : CreativityStake_{ES} = CreativityStake_{HS}$ (Alt: $\neq$)
$H_0^2 : ObjectivityStake_{ES} = ObjectivityStake_{HS}$ (Alt: $\neq$)
$H_0^3 : ExhaustivityStake_{ES} = ExhaustivityStake_{HS}$ (Alt: $\neq$)
$H_0^4 : SteadinessStake_{ES} = SteadinessStake_{HS}$ (Alt: $\neq$)
$H_0^5 : OrderabilityStake_{ES} = OrderabilityStake_{HS}$ (Alt: $\neq$)
$H_0^6 : CreativityStake_{ES+} = CreativityStake_{ES-}$ (Alt: $\neq$)
$H_0^7 : ObjectivityStake_{ES+} = ObjectivityStake_{ES-}$ (Alt: $\neq$)
$H_0^8 : ExhaustivityStake_{ES+} = ExhaustivityStake_{ES-}$ (Alt: $\neq$)
$H_0^9 : SteadinessStake_{ES+} = Steadiness_{ES-}$ (Alt: $\neq$)
$H_0^{10} : OrderabilityStake_{ES+} = OrderabilityStake_{ES-}$ (Alt: $\neq$)

- **Steadiness**: the extent to which a statement will remain unchanged over time. A statement can range from *tentative*—there is no certainty about whether the statement will remain true or not in the future—to *steady*—there will not be changes in the future;
- **Orderability**: the extent to which the statement can be ranked against other statements, in terms of importance and urgency. A statement can range from *Orderable*—there is a clear indication of whether it is an important/significant statement compared with other statements—to *fuzzy*—there are no clues about its importance compared with other statements.

3.1.2 Measurement of variables

Our variables are latent variables. Quantifying it accurately would therefore require the use of several observable variables to be estimated by several different people involved in the interaction. In this first experiment, we approximate the measure of each variable by asking directly to authors of statements how they would evaluate each variable, for each statement. What we measure, therefore, is not the objectivity but rather the perception of objectivity by the author of a statement. We consider that measuring stakeholders' ratings of their respective statements provides valuable information about the overall quality of those statements. Given the fact that stakeholders are also the owners of their statements, forcing introspection of what they uttered is likely to reveal important indications for the analysts that we also want to capture as part of this study. We acknowledge the inherent limitations and impact this has on the generalizability of our results. This work is a first step aiming to confirm the relevance of the topic. Positive answers to the questions above would confirm the need to conduct more accurate studies using, for instance, a structural equation modeling approach. In any case, this remains for future works.

3.1.3 Experimental design

To test our hypotheses, we need to collect a set of statements and, for each of these statements, determine the value of the *independent variables* as well as the nature of their *ground*. We need to collect those data in a context that is as representative as possible of how requirements elicitation happens. We therefore opt for an experiment, in which actual requirements are shared by stakeholders about an actual software. Controlled experiment indeed "is an investigation of a testable hypothesis where one or more independent variables are manipulated to measure their effect on one or more dependent variables" (Easterbrook et al. 2008), which fits the objectives of this paper. The experiment took place in three different steps, as described below.

Step 1—Collecting a sample of statements: this first part of the experiment acts as an overall introduction: subjects are explained during a workshop—usually involving 10 subjects at a time—that a new web platform has to be designed for the University of Namur, in order to share courses material, to discuss group assignments, to ask questions to professors and teaching assistants, etc. An old system exists, but needs to be replaced. Subjects are told a project is initiated by the University to ensure the new platform fits well with the needs of students and staff members, and that they are involved as stakeholders of that future system. In other words, we asked them to share requirements for a system of which they are actual stakeholders.

Subjects were first invited to discuss together about the platform and their requirements, afterwards their were asked to actually share their statements with the authors. Interviews are not optimal to collect a large number of statements, so that we designed a web interface on which subjects could encode their statements about the future system. Subjects used this interface during the workshop itself. The interface consists of different tabs, in which stakeholders can write down their statements. There is a limit of 500 characters for each tab. An important constraint is to encode at least three different statements. Subjects are able to give more than three statements. There is no time constraint to complete the form. Examples of statements are also provided to help stakeholders understand what is expected from them. In this step, no evaluation of the statements occurs, and the quality variables we use in step 2 are not visible to subjects. This is important to avoid anticipation or order effects.

Step 2—Collecting the statements variables: the second step of the experiment focuses on the evaluation of independent variables. Below each statement encoded in step 1 by the stakeholder, the five quality variables are displayed and stakeholders are invited to evaluate them using a 6-level scale (see Fig. 2). From this moment, the statements cannot be edited anymore (they appear in read-only mode). Each variable was discussed separately with participants during the workshop, to make sure everything was clear to them. Following discussion from Subsection 3.1.2, we measure the perception stakeholders have of their statements, in terms of the five variables we focus on in this experiment. A translation of the questions submitted to stakeholders is reported in Table 3. The translation of the scale available to stakeholders to answer those questions is reported in Table 4. Note again that the variables are not collected at the same moment as the statements themselves, to avoid order effects.

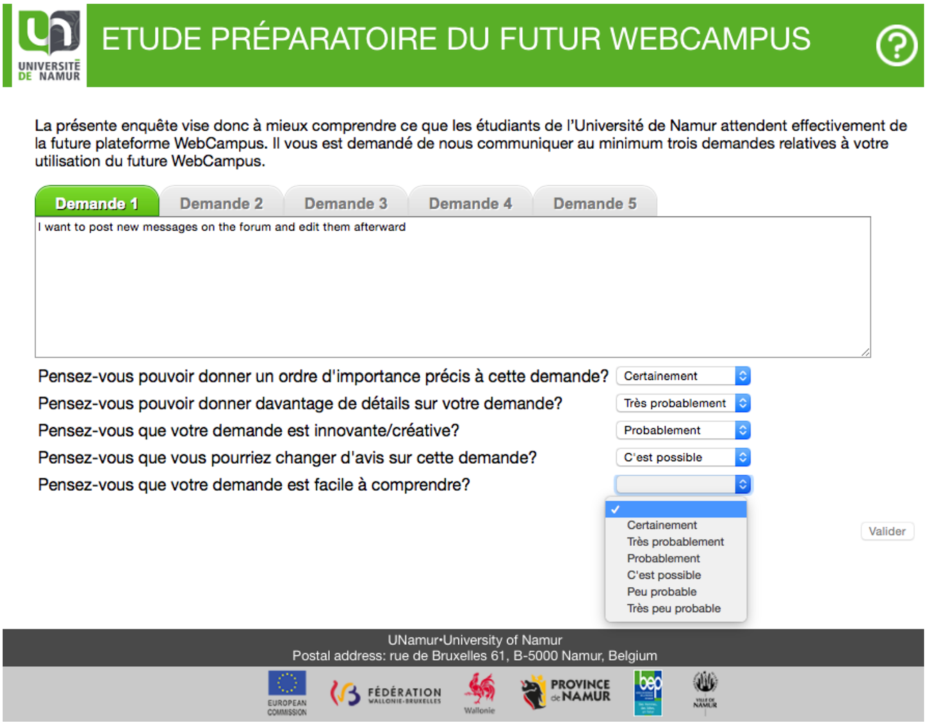

Fig. 2 Experiment—collecting statements

Table 3 Study variables to approximate statement quality

Variable	Questions
Orderability	Do you think you can give a specific order of importance to this statement?
Exhaustivity	Do you think you could provide more details about this statement?
Creativity	Do you think that your statement is innovative / creative?
Steadiness	Do you think that your statement will still apply and matter to you in a year?
Objectivity	Do you think that your statement is easy to understand for others?

Step 3—Collecting the statement grounds : the third part focuses on the identification of the ground of each requirement. Each statement that stakeholders encoded in the first part of the survey is shown to its author. The stakeholder is then asked to tell whether she already used a same system satisfying that demand (e.g., the old Web platform to be replaced) (ES+), another similar system satisfying that demand (ES-), or whether she never experienced that demand before (HS). It is impossible for the stakeholder to get back to step 2, so that the identification of the grounds systematically happens after—and independently from—the writing of statements themselves. Such design gives the chance for a given stakeholder to share both experiential and hypothetical statements, i.e., a stakeholder is not experiential, some of her statements are. Once the step is concluded, stakeholders are told that the workshop was actually part of an experiment.

3.1.4 Subjects

We solicited people from the University of Namur (Belgium) having a stake in the development of the platform. Stakeholders of the platform include students, teaching assistants, and teachers. A total of 98 subjects took part to our experiment, including computer science, law, economics, and management students, as well as a dozen members from the teaching staff. Students were heterogeneous, with undergraduate and graduate students. This increases the probability of collecting both hypothetical and experiential statements, as some students had no experience with the old system. We used a random sampling method to select groups of students who participated in the workshops; all students were listed, and we randomly picked some of them and invited them to take part in our experiment. Subjects were not compensated for participating in the study; the only motivation was to contribute to the design of a new platform on a voluntary basis. Subjects did not know beforehand that the case was a fictive one. Subjects were informed beforehand that this was part of a research project. Subjects were free to refuse participation, or to remove their answers after the workshop. The entire process took place in French, i.e., the online questionnaire, the assignment, the statements, and explanations of the experimenters were all in French.

Table 4 List of scales submitted to stakeholders

Scale +	Scale item -
1. Definitely	6. Definitely not
2. Very probably	5. Very probably not
3. Probably	4. Probably not

3.1.5 Preliminary study

Our survey is the result of some prior work which has failed studying reliably the grounds of statements. Our initial approach was to perform a 2-step interview with a dozen stakeholders, before and after the use of a system they initially had never used. Our goal in doing so was to identify some statements, and to quantify the value of properties associated with the latter before and after use. Stakeholders told us they were not comfortable in evaluating the properties as asked, so that we could not use the data collected through interviews. We learnt the following from this first attempt:

- Subjects were afraid to give a "wrong" answer, due to the physical presence of the experimenter in front of them;
- Subjects were not committed enough, due to poor context during the interviews (subjects were not questioned in the context of a real RE project of a system);
- The delay between the two interviews was too short, so that stakeholders could not be considered to be experienced at the moment of the second interview;
- Data was qualitative, and it was difficult to collect a sufficiently large number of statements via interviews, so that statistical analysis was hardly feasible;
- Our design based on repeated measures (in this study, before and after use) introduced several biases in the results which made results insignificant.

These observations lead to our second radically different design. We opted for an online questionnaire, to avoid the influence of the interviewer. This enabled us to define a more concrete context for subjects, to ensure higher commitment. The questionnaire evaluates grounds at the level of the statement, not at the level of the system, i.e., there is not one system that only leads to experiential or hypothetical statements, there are systems, which may lead to some hypothetical and some experiential statements. The online questionnaire made it more straightforward to collect such thinner granularity information, in larger quantity. In our second design, we do not perform repeated measure analysis, we simply compare groups of hypothetical and experiential statements to see how they differ.

Another advantage of this first study is that we clarified with stakeholders which question was best to ask in our online questionnaire as a way to actually measure our five qualities of variables. Our initial questions to stakeholders, for instance, were asking directly about the quality: "do you think this statement is objective?". This was not easy for stakeholders to interpret, so that we improved the questions throughout the preliminary story to obtain questions listed in Table 3. While this does not guarantee perfect internal validity of our study, we consider it provides some clues that our questions to stakeholders are valid.

3.1.6 Results

We collected 632 statements during our experiments. After reporting and evaluating their statements, each subject was asked to confirm or not each statement. Subjects were told that confirming amounts to communicate the statement to analysts (i.e., the stakeholder confirms this is something he/she actually wants) while rejecting implies to withdraw the statement from the process (i.e., the stakeholder is not sure about his/her statement and prefer to not communicate it for further analysis). We received a confirmation for 353 statements. We removed 35 observations due to quality issues (subjects did not evaluate the five variables for the statement, text was not readable, etc.), to end up with 318 usable statements. We use resulting data to test hypotheses listed in Table 2. As a reminder, our main objective in this first experiment is to show that experiential statements differ significantly from hypothetical ones, in terms of the properties we defined in Section 3, i.e., we test whether some variables are significantly different between two groups. We collected ordinal data, under the form of a scale described in Table 4. A Kolmogorov-Smirnov test on the variables confirms that our distributions are not normal, so that we need to resort to a non-parametric test to compare the groups; we use the Mann-Whitney U test. The test works under the main assumptions that all observations from the considered groups are independent from each other, and that data are ordinal, i.e., it must be possible to tell which of two observations is the greater. These two assumptions are verified in our case. The Mann-Whitney U tests that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample (H_0). Note that we also report adjusted P values using the Benjamini-Hochberg procedure (with $N = 5$) to account for the multiple comparison we perform on a same set of data. Finally, we report an effect size for each inference, using the r of Cohen ($r = Z/\sqrt{N}$) (Fritz et al. 2012). A value of 0.1 suggests a small effect, 0.3 a medium effect, and 0.5 a large effect.

Experiential vs hypothetical statements (HS vs ES) are first compared, regardless of the experience being direct or not. From our 318 initial statements, we identify 186 observations in the HS group, and 132 in the ES group. We report the main descriptive statistics for each group in Table 5. For each variable of the study, we run a Mann-Whitney U test to test the null hypothesis that the two groups are similar. Table 5 summarizes our results. We observe a significant difference in perception between ES and HS in terms of creativity, objectivity, steadiness, and orderability. This means that stakeholders perceive differently a statement depending on its ground. We cannot conclude anything about the exhaustivity of statements. We report a graphical representation of our results under the form of bar plot in Fig. 3a to d, to facilitate interpretation. Each bar is divided in two to reflect the distribution of ES vs HS for each level of the scale. As a reminder, 1 reflects a strong agreement of the stakeholder with the variable, while 6 represents a strong disagreement. On that scale, smaller values indicate higher perceived quality.

We observe that a large part of ES are considered as steady by the stakeholders. This proportion however reduces in a constant way, as the steadiness decreases (getting closer to 6 on the scale). The opposite trend is observed for HS; a larger proportion of stakeholders consider their statements to be more changing over time. Mean values of the HS vs ES group for the perception of steadiness (available in Table 5) confirm that hypothetical statements tend to be perceived as less steady by stakeholders than experiential ones. The exact same pattern is observed for objectivity and orderability variables. The case of creativity is different; we observe that the proportion of experiential statements is larger for conservative (less creative) levels of our scale, and that this ratio reduces constantly as statements become more creative. In other words, we observe a U-shape distribution of statements along the creativity dimension, with experiential statements being perceived as either very creative or very conservative. We will discuss this result a bit later, in light of the ES-/ES+ analysis conducted below.

Direct vs indirect experiential statements are now analyzed. Here, we compare two groups in the subset of our data composed of experiential statements: direct (ES+) and indirect (ES-) experiential statements. As a reminder, a stakeholder may share a statement about a system using past experience with that system (ES+) or with another similar system (ES-). From the 132 experiential statements, we identify 70 observations in the ES+ group, and 62 in the ES- group. We adopt the same approach as for hypotheses H_1 to H_5; we run a Mann-Whitney U test to test the null hypothesis that the two groups are similar. Table 6

Table 5 Mann-Whitney *U* tests—stakeholders perception of HS vs ES statements

	Creativity H_0^1	Objectivity H_0^2	Exhaustivity H_0^3	Steadiness H_0^4	Orderability H_0^5
Mean ES	3.36	2.09	2.97	4.71	2.43
Mean HS	3.73	1.89	2.84	4.93	2.10
Std Dev.	1.562	1.207	1.490	1.226	1.290
Mann-Whitney *U*	10507.500	10935.500	11467.000	10499.000	10284.000
Z-value	2.230	1.775	1.021	2.296	2.568
P value	0.026**	0.076*	0.307	0.022**	0.010***
Adj. *P* value	0.043**	0.043**	0.307	0.022**	0.050**
Rank ES	149.99	166.71	163.85	169.05	170.21
Rank HS	172.90	149.34	153.37	146.04	144.41
Size effect	0.125	0.09	0.06	0.129	0.144

Sign: *** = 1%, ** = 5%, * = 10%

summarizes our results for this second phase. We observe a significant difference between ES+ and ES- in terms of creativity and steadiness. This means that stakeholders perceive differently a statement depending on the robustness of the experience they leveraged to produce a statement. We cannot conclude anything about the exhaustivity, orderability, and objectivity of statements. A representation of our results is depicted in Fig. 3a to d.

We observe that a much higher proportion of ES+ is observed for low levels of the stability scale, suggesting direct experiential statements are seen as more stable over time than indirect experiential statements. The distribution of ES- against ES+ is bigger for lower levels of the scale. This is confirmed in Table 6 where the mean level for ES+ is lower than for ES-. Regarding creativity, we observe a clearer pattern than in the ES/HS analysis: Indirect experiential statements are seen more frequently by stakeholders as creative (low levels

Table 6 Mann-Whitney *U* tests—stakeholders perception of ES- vs ES+ statements

	Creativity H_0^6	Objectivity H_0^7	Exhaustivity H_0^8	Steadiness H_0^9	Orderability H_0^{10}
Mean ES+	4.39	1.85	3.10	1.83	2.26
Mean ES-	2.98	1.98	2.55	2.24	1.92
Std Dev.	1.704	1.219	1.591	1.285	1.277
Mann-Whitney *U*	1102.500	1974.500	1769.000	1634.500	1908.000
Z-value	4.953	0.977	1.867	2.586	1.263
P value	0.000***	0.328	0.062*	0.010***	0.207
Adj. *P* value	0.000***	0.328	0.103	0.025**	0.259
Rank ES+	81.75	63.71	72.23	57.86	70.24
Rank ES-	49.28	69.65	60.03	74.15	62.27
Size effect	0.278	0.054	0.104	0.145	0.071

Sign: *** = 1%, ** = 5%, * = 10%

Fig. 3 Graphical representation of stakeholders' perception of HS, ES-, and ES+. **a** Creativity according to stakeholders. **b** Objectivity according to stakeholders. **c** Steadiness according to stakeholders. **d** Orderability according to stakeholders

of the scale) than direct experiential statements. The effect is very significant, according to Table 6. Following Table 6, mean level for this variable is 2984 in the ES- group against 4386 in the ES+, suggesting direct experience is perceived by stakeholders as something harmful to creativity.

3.2 Experiment 2—Quality perception by requirements engineers

The second experiment focuses on the requirements engineer side. It intends to assess the perception of qualities of statements collected during our first experiment from an engineer point of view. We proceed in the same way as in experiment 1; we first provide details about the hypotheses we want to test, describe the procedure used in this second experiment, discuss our sample, and finally detail the results of the study.

3.2.1 Definition of variables and hypotheses

We use exactly the same variables as in the first experiment, but change the point of view. Our objective here is to measure the perception of quality of statements by requirements engineers who are supposed to use those statements to produce actual requirements, and show that this perception changes significantly with the ground of the statement. Our variables are therefore the perception by engineers of creativity, exhaustivity, steadiness, objectivity, and orderability. The dependent variable remains the ground of the statement, as defined by stakeholders in experiment 1, i.e., HS, ES+, or ES-. Table 7 lists those new null hypotheses that we expect to reject; we hope to demonstrate that there is a statistically significant difference in perception of quality by engineers between grounds, for those different statements properties. Another important question we try to deal with in this second experiment is the one of alignment; do stakeholders and engineers have an aligned view

Table 7 Hypotheses on the perception by engineers of HS, ES-, and ES+ statements

How do engineers perceive the quality of statements with different grounds?
H_0^{11} : $CreativityEng_{ES} = CreativityEng_{HS}$ (Alt: $\neq$)
H_0^{12} : $ObjectivityEng_{ES} = ObjectivityEng_{HS}$ (Alt: $\neq$)
H_0^{13} : $ExhaustivityEng_{ES} = ExhaustivityEng_{HS}$ (Alt: $\neq$)
H_0^{14} : $SteadinessEng_{ES} = SteadinessEng_{HS}$ (Alt: $\neq$)
H_0^{15} : $OrderabilityEng_{ES} = OrderabilityEng_{HS}$ (Alt: $\neq$)
H_0^{16} : $CreativityEng_{ES+} = CreativityEng_{ES-}$ (Alt: $\neq$)
H_0^{17} : $ObjectivityEng_{ES+} = ObjectivityEng_{ES-}$ (Alt: $\neq$)
H_0^{18} : $ExhaustivityEng_{ES+} = ExhaustivityEng_{ES-}$ (Alt: $\neq$)
H_0^{19} : $SteadinessEng_{ES+} = SteadinessEng_{ES-}$ (Alt: $\neq$)
H_0^{20} : $OrderabilityEng_{ES+} = OrderabilityEng_{ES-}$ (Alt: $\neq$)
Does the perception of engineers differ from the perception of stakeholders?
H_0^{21} : $CreativityEng = CreativityStake$ (Alt: $\neq$)
H_0^{22} : $ObjectivityEng = ObjectivityStake$ (Alt: $\neq$)
H_0^{23} : $ExhaustivityEng = ExhaustivityStake$ (Alt: $\neq$)
H_0^{24} : $SteadinessEng = SteadinessStake$ (Alt: $\neq$)
H_0^{25} : $OrderabilityEng = OrderabilityStake$ (Alt: $\neq$)

on the quality of statements depending on the ground? Hypotheses related to this second question are also included as part of Table 7.

3.2.2 Experimental design

The procedure in this second experiment is rather simple. To test our hypotheses, we reuse the statements collected from stakeholders during experiment 1. We submit each of the 318 selected statements to requirements engineers, and ask them to evaluate for each statement the different quality variables. The procedure is similar; for each statement, engineers have to answer five different questions, as reported in Table 8. Requirements engineers could answer those questions using the same scale as used in experiment 1 (see Table 4). It is important to note that no transformations were made on statements produced by stakeholders; analysts were exposed to the statements "as produced," with their typos, ambiguities, etc. in order to avoid any bias.

Table 8 Study variables to approximate statement quality

Variable	Questions
Orderability	Do you think this statement is easy to prioritize over other requirements?
Exhaustivity	Do you think this statement is complete and detailed enough to specify a requirement for a future system?
Creativity	Do you think this statement is innovative / creative?
Steadiness	Do you think this statement is likely to change in the future?
Objectivity	Do you think this statement is easy to understand to produce requirements?

3.2.3 Subjects

We invited 12 business analysts from Technofutur TIC, a large Belgian competence center offering training to professionals and unemployed people on information technologies. Technofutur TIC is using a Moodle platform (the same technology used by Webcampus) to share material with their trainees, so that the business analysts from Technofutur TIC have a good understanding of what Webcampus is and of how it works. As a reminder, Webcampus was used in experiment 1 as a pretext to collect statements from stakeholders. Business analysts from Technofutur are therefore good candidates to gather data about the perceived quality of statements produced by our stakeholders. Each analyst received from 25 to 27 statements to evaluate across the five qualities of our study. They completed the evaluation using a Qualtrics questionnaire. Statements and qualities were displayed randomly to analysts, to avoid any bias in the sequence of reply.

3.2.4 Results

We collected quality evaluations for each of our 318 statements during our second experiment. We combined these evaluations with the grounds identified in the first experiment (HS, ES+, and ES-) to constitute our second dataset. We report in Table 9 the descriptive statistics for each analyst. Similarly to experiment 1, the data we collected are ordinal, under the form of a 6-level scale. A Kolmogorov-Smirnov test on the variables confirmed that our distributions are again not normal. We therefore resorted to the same test as the one used above, the Mann-Whitney U test. To test hypotheses H_0^{21} to H_0^{25}, we used the Wilcoxon signed-ranked test, a variant of the Mann-Whitney U which allows us to compare paired observation (the evaluation from a stakeholder and from a business analyst of the same statement).

Experiential vs hypothetical statements (HS vs ES) are first compared, regardless of the experience being direct or not. We report the main descriptive statistics for each group in Table 10. For each variable of the study, we run a Mann-Whitney U test to test the null

Table 9 Descriptive statistics—engineers perception of statements

	A_1	A_2	A_3	A_4	A_5	A_6
Number of HS	15	14	16	10	15	14
Number of ES-	6	6	7	8	5	7
Number of ES+	6	7	4	9	7	6
Mean score	3.34	3.92	2.49	2.95	5.15	3.60
Std Dev.	1.41	1.21	1.18	1.45	0.96	1.28
Skewness	0.06	-0.48	0.05	0.22	-1.34	0.11
	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}
Number of HS	14	18	18	18	19	15
Number of ES-	5	4	4	2	2	6
Number of ES+	7	4	4	6	5	5
Mean score	3.77	2.95	4.65	4.33	3.51	3.88
Std Dev.	1.81	1.32	1.17	1.84	1.78	1.46
Skewness	0.24	0.02	-0.84	-0.96	-0.17	-0.29

Table 10 Mann-Whitney U tests—engineers perception of HS vs ES statements

	Creativity H_0^{11}	Objectivity H_0^{12}	Exhaustivity H_0^{13}	Steadiness H_0^{14}	Orderability H_0^{15}
Mean ES	3.92	2.68	3.24	3.13	3.01
Mean HS	3.39	3.09	3.58	3.36	3.37
Std Dev.	1.635	1.584	1.655	1.647	1.595
Mann-Whitney U	10,034.500	10,586.500	10,918.500	11,314.000	10,621.500
Z-value	2.823	2.133	1.705	1.211	2.087
P value	0.005***	0.033**	0.088*	0.226	0.037**
Adj. P value	0.025**	0.062*	0.110	0.226	0.062*
Rank ES	147.45	146.70	149.22	152.21	146.97
Rank HS	176.48	168.58	166.80	164.67	168.40
Size effect	0.158	0.120	0.096	0.068	0.117

Sign: *** = 1%, ** = 5%, * = 10%

hypothesis that the two groups are similar. Table 10 summarizes our results. We observe a significant difference in perception of engineers between ES and HS in terms of creativity, objectivity, exhaustivity, and orderability. This means that engineers perceive differently a statement depending on its ground. We cannot conclude anything about the stability of statements. Again, we report a graphical representation of our results under the form of bar plots in Fig. 4a to d, to facilitate interpretation. We observe that a large part of HS are considered as creative. This is in line with the perception of stakeholders; our two groups tend to agree that less experience generates more creative statements (this point is further discussed later in the paper). Objectivity on the contrary displays a higher proportion of disagreement from engineers for the HS statements, suggesting an HS statement is seen as harder to interpret and understand than an experiential one by engineers, i.e., the more experience, the easier the understanding. The same pattern is also observed for exhaustivity, where ES tend to have more "1" answers. Orderability is more nuanced; although we observe a statistically significant effect, the interpretation of the result turns out to be harder; overall, we observe fewer HS for high orderability perception (1), and more for low orderability (5 and 6). Peaks of HS are however observed for 3 and 5 answers, so that no clear pattern can be extracted. Overall, we remain cautious about this last quality. One possible explanation of this is that prioritization, unlike the three other significant qualities of the experiment, is something depending strongly on the business and its stakeholders; engineers therefore may have troubles assessing the orderability of a statement.

Direct vs indirect experiential statements generate no statistically significant result in terms of perception by the engineers. It seems that business analysts from our experiment do not perceive a difference in quality between statements produced with ES- and those produced with ES+, i.e., the type of experience used by the stakeholder does not matter. This is also very clear from Fig. 4a to d, where the number of ES+ vs ES- is basically the same for each level of the scale (Table 11).

Stakeholders vs engineers perception is the last set of questions that remains to be answered. As discussed earlier, we used a Wilcoxon signed-rank test to compare the paired observations; for each statement, we have a perception from a stakeholder and one from an

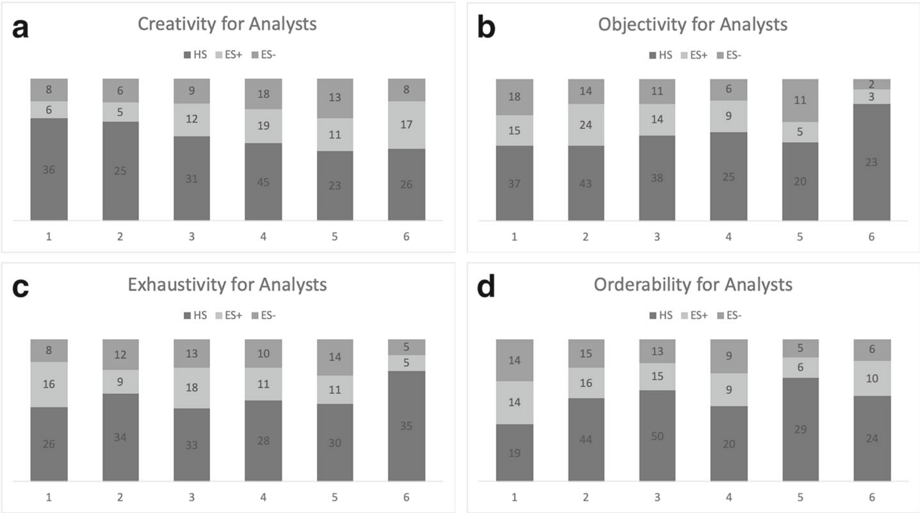

Fig. 4 Graphical representation of analysts' perception of HS, ES-, and ES+. **a** Creativity according to analysts. **b** Objectivity according to analysts. **c** Exhaustivity according to analysts. **d** Orderability according to analysts

engineer. We want to see if, in general, there is a difference between the two types of actor. Results are reported in Table 12. We observe that creativity is the only quality where both engineers and stakeholders are aligned (we cannot reject the null hypothesis that the median difference between the paired observation is equal to zero); this also reflects clearly when comparing Figs. 3a and 4a; the patterns are very similar. We do not observe any other matching. This means that, for exhaustivity, completeness, steadiness, and objectivity, engineers and stakeholders have diverging perceptions on the quality of the statements (regardless of its ground).

Table 11 Mann-Whitney *U* tests—engineers perception of ES- vs ES+ statements

	Creativity H_0^{16}	Objectivity H_0^{17}	Exhaustivity H_0^{18}	Steadiness H_0^{19}	Orderability H_0^{20}
Mean ES+	4.07	2.63	3.10	3.20	3.10
Mean ES-	3.74	2.74	3.40	3.05	2.90
Std Dev.	1.548	1.464	1.564	1.603	1.632
Mann-Whitney *U*	1922.500	2129.500	1932.500	2052.500	2036.000
Z-value	1.152	0.189	1.102	0.546	0.623
P value	0.249	0.850	0.271	0.585	0.534
Adj. *P* value	0.678	0.850	0.678	0.731	0.731
Rank ES+	70.04	65.92	63.11	68.18	68.41
Rank ES-	62.51	67.15	70.33	64.60	64.34
Size effect	0.065	0.010	0.062	0.031	0.035

Sign: *** = 1%, ** = 5%, * = 10%

Table 12 Wilcoxon signed-rank tests—engineers vs stakeholders perception

	Creativity H_0^{21}	Objectivity H_0^{22}	Exhaustivity H_0^{23}	Steadiness H_0^{24}	Orderability H_0^{25}
Negative ranks	126.04	98.63	126.99	147.32	103.84
Negative *N*	127	64	100	218	79
Positive ranks	134.76	127.18	135.07	92.81	142.14
Positive *N*	133	174	163	54	181
Ex aequo	58	80	55	46	58
Z-value	0.799	7.519	3.813	10.507	7.304
P value	0.249	0.000***	0.000***	0.000***	0.000***
Adj. *P* value	0.249	0.000***	0.000***	0.000***	0.000***
Size effect	0.045	0.422	0.214	0.590	0.409

Sign: *** = 1%, ** = 5%, * = 10%. Negative case: engineer's answer $<$ stakeholder's answer, positive case: engineer's answer $>$ stakeholder's answer, ex aequo case: engineer's answer = stakeholder's answer

4 Discussion

Our results suggest that stakeholders and engineers associate different quality properties to the statements, depending on their grounds. In most cases, we also observed that engineers and stakeholders perceive differently a same statement. We summarize the main conclusions of our experiment in Table 13 and discuss them below.

H_1, H_6, and H_{11} confirmed—Experience blocks creativity: Statements collected during our experiment are perceived as more creative by stakeholders who have no experience with the system under consideration and therefore resort to hypothetical statements. Feel of creativity however is higher when experience is indirect. On the contrary, stakeholders with direct experience perceive their requirements as being more conservative. The same pattern is observed for the group of engineers, although we could not detect a significant effect on the type of experience (direct of indirect) for them.

Table 13 Summary of effects of grounds on statements perception

	Creativity	Objectivity	Exhaustivity	Steadiness	Orderability
Stakeholders	Effect	Effect	No effect	Effect	Effect
HS vs ES	H_1^1	H_1^2	H_0^3	H_1^4	H_1^5
Stakeholders	Effect	No effect	No effect	Effect	No effect
ES- vs ES+	H_1^6	H_0^7	H_1^8	H_1^9	H_0^{10}
Engineers HS	Effect	Effect	No effect	No effect	Effect
HS vs ES	H_1^{11}	H_1^{12}	H_1^{13}	H_0^{14}	H_1^{15}
Engineers HS	No effect	No effect	No effect	No effect	Effect
ES- vs ES+	H_0^{16}	H_0^{17}	H_0^{18}	H_0^{19}	H_0^{20}
Stakeholders	No effect	Effect	Effect	Effect	Effect
Engineers	H_0^{21}	H_1^{22}	H_1^{23}	H_1^{24}	H_1^{25}

H_2 and H_{12} confirmed—Experience helps objectivity: we find that stakeholders with experience are more confident in the objectivity of the statements they communicate, i.e., they believe the statements they share could be understood as intended by other people, without additional information. This is also confirmed for engineers, who tend to find hypothetical statements harder to interpret than experiential ones. Despite this, we observe that engineers and stakeholders are not perfectly aligned in their perception; this can be explained by a much higher frequency of "Probably" answers from stakeholders, which shapes the distribution of answers differently than for engineers. Finally, it was impossible to conclude a difference in objectivity between direct and indirect experience, from the stakeholders and engineers perspective.

H_8, H_{13}, and H_{23}—Exhaustivity comes with experience for engineers only: it seems that stakeholders with direct experience tend to consider their statements as more complete than when they only have indirect experience. However, we cannot conclude that hypothetical statements are seen as less complete by their authors, so that this quality overall has no effect on the perception of stakeholders. We could explain this by the fact that exhaustivity is hard to assess for non-experts; it requires a good understanding of the RE process to know if a statement contains all the necessary information, which is hardly the case of stakeholders in general. On the contrary, we observe that engineers perceive experiential statements as more complete than hypothetical ones, suggesting stakeholders with experience produce more detailed statements that are seen as more exhaustive by engineers. Caution is required however, since we find not significant adjusted P values for those hypotheses (only the non-adjusted P values are below the significance level of 10%). In any case, we observed no significant effect of the type of experience. We conclude that engineers and stakeholders have different perceptions on this quality.

H_4, H_9, and H_{24} confirmed—Experience stabilizes statements for stakeholders only: hypothetical statements are seen as more likely to evolve or to be withdrawn as time goes by. The effect of experience is clear in our sample; stakeholders with experience claim more frequently that their statements are steady. We also find that direct experience produces statements which are perceived as more steady than those produced through indirect experience. In other words, the bigger the experience, the more stable the statements. Those effects do not exist for engineers; the ground of a statement did not influence the perception of analysts in our study. This could be explained by the fact that steadiness implies a deep understanding of the environment in which the statement is made. Engineers in our experiment did not have that understanding, and therefore massively replied with the neutral answers (3 and 4 on our scale), for any type of ground. This results naturally in a significant difference between the perception of engineers and stakeholders.

H_5, H_{15}, and H_{25} confirmed—Experience to sort things out: we find that experience, indirect or direct, helps stakeholders to put an order of priority on the importance of their statements, i.e., stakeholders expect to have a clearer idea about the priority of a statement when it builds on experience. The same pattern was also observed, although in a less extreme way, for engineers. We cannot conclude anything about the impact of experience type (direct vs indirect), for both engineers and stakeholders.

Ultimately, it is interesting to note that some of our qualities appear to be universal, in the sense that the ground affects in the same way the perception of both engineers and stakeholders (this is the case of creativity). This can be explained by the fact that creativity is not something proper to RE in general. On the contrary, some of our qualities are proper to one actor. For instance, we discussed the fact that exhaustivity is a concern only in RE, while it remains a relatively vague notion to stakeholders who are not familiar with the RE process and can hardly judge if their statements are complete enough to generate an RE entity. Steadiness, on the other hand, is something that matters to engineers but which is mostly related to the business side and the changes occurring in that business. As such, it is relatively hard for engineers to capture it via a standalone statement. Finally, objectivity and orderability relate to the communication about the RE process between the two actors. Actors may not have aligned views on these, but both of them have significantly different perceptions on them depending on the ground of the statement.

5 Practical implications and future work

Our experiment made it clear that the ground of a statement influences the way stakeholders and engineers perceive the quality of that statement, at least in terms of the variables we explored in this paper. We did not measure in any way the *actual quality* of related RE entities due to the well-known practical difficulties of measuring quality (see Section 3.1.2). Our claim however is that the perception of quality by both stakeholders and engineers of a same statement will likely influence the qualities of future RE entities documented through it. To leverage previous empirical results and produce practical conclusions, a first important question is therefore to explore empirically what has remained a working assumption in this paper:

> A direct and positive relation exists between the quality of a statement perceived by a stakeholder and an engineer, and the actual quality of any related RE entity.

For instance, a stakeholder who shares a statement that is perceived as exhaustive by that stakeholder and a requirement engineer has more chance to produce an RE entity that is actually more exhaustive, all else being equal. Stated differently, we make the assumption that a statement perceived as more exhaustive presents a higher probability to lead to an exhaustive RE entity, *although it is not a sufficient condition*. Considering that stakeholders are the source of information, experts in their fields and engineers are owner of RE entities comforts us in the use of this working assumption. Additional empirical evaluations are however needed to confirm its relevance.

Another important question following the previous assumption is: “How can we account for the Ground of Statements – and therefore for related qualities – when modeling RE entities?”. The intuition here is that establishing a RE model based on a bad mix of statement grounds may accentuate some risks for the RE process, which could be mitigated if modeled and dealt with properly. For instance, a model building only on hypothetical statements has a higher chance to be creative, but also presents a higher probability of flaws during prioritization or validation of requirements. Modeling RE entities together with their statements grounds is a way to identify, early in the RE process, potential flaws in the RE activities and take actions in order to mitigate them. In order to attain such results, evolutions to existing requirements modeling language will have to be made. A first potential contribution could be to introduce a notation for statement grounds, and to formalize the different ways in which that notation could be used to analyze the inherent quality of a RE model. Many related questions can also be identified, like the one of computing a risk score for a RE model based on the statements that have been used, or the question of how we can further automate risk detection by using the present and future empirical contributions on the concept of ground, and its impact on RE entities quality. More generally, practical implications can also be identified in other sub-fields of RE like elicitation or prioritization, to name only two. For instance, investigating the potential impact of statement grounds on the selection of a particular elicitation technique, on the method used to conduct the entire elicitation process, and on the procedure applied to negotiate or prioritize requirements are some examples of questions that we consider interesting to investigate in the future.

6 Limitations

Our work relies on the use of an important working assumption: "a direct and positive relation exists between the quality of a statement perceived by an engineer and the actual quality of any related RE entity." It could be argued that the way engineers perceive the qualities of statements they collect may be very different from the actual quality of that statement. We acknowledge that limitation, but stick to the primary objective of the paper: provide a first empirical evaluation of the impact of statement grounds on the early stages of the RE process. Even though our approach is probably not perfectly representative of the actual quality of a resulting RE entity, it comes as a support to anticipate potential weaknesses in the elicitation process. Besides, the identification of the grounds is only one quality indicator that should be used in combination with many others. Finally, we consider that the combination of stakeholders' and engineers perceptions still provides a valuable indication of the overall quality of a resulting RE entity. This is because stakeholders are experts in their field, owner of the information, and at the origin of most RE entities, while engineers have expertise in the analysis and formalization of those RE entities.

Another potential limitation of our work is the external validity of our experiments, conducted with a hundred people from the University of Namur and a dozen business analysts from Technofutur TIC. The generalizability of our conclusions is therefore limited. It is important however to remember that the very concept of ground is a human-centered one, not an RE-specific one. What we mean is that we study something related to human psychology, namely how people produce and justify some statements. As a consequence, the specifics of the RE project in which the experiment took place have relatively low impact on the nature of our conclusions. Replicating our results in different RE project should therefore lead to relatively similar results. Additional experiments could help reduce the eventual bias due to the subjects involved in our experiment (gender, cultural factors, etc.).

A last limitation is related to the internal validity of our experiment, conducted on a relatively small number of quality variables. The notion of quality entails much more than the five variables we used, and many more factors worth being investigated. We justify this choice by the early nature of the present contribution; the concept of statement ground is new, and not much is known about it. In any case, it is important to remember that the five quality properties we selected are those which seem the most prominent among RE community, i.e., we captured quality factors which matter the most the RE practitioners in general. Future research could also focus on the inclusion of additional quality variables, to make our risk assessment approach more robust.

7 Conclusion

Stakeholders, when sharing information about a software-to-be, will likely brainstorm and produce a lot of statements. Some will just be simple ideas, some will be more complex demands based on things they experienced several times in the past, others will simply be approximations of something they experienced in a different context, etc. Those statements, we claim, are not proper requirements, and more generally not proper RE entities. Among the various collected statements, some will likely be withdrawn later in the process, as stakeholders realize this might not be the thing they actually needed. Others may be robust, and become actual requirements. This paper investigates the connections between the so-called statements and RE entities. It suggests that different statements shared by stakeholders build on different grounds, and that those grounds likely influence the intrinsic quality of the statement, and hence of related RE entities. To explore these hypotheses, we conducted an experiment where subjects were asked to share various statements, and then to evaluate the perceived quality of those different statements. Doing so, we found out that statements built on experience increase the objectivity of a statement and reduce the risk of volatility. On the other hand, experience is not always desirable as it leads stakeholders to produce less creative RE entities. To explore these hypotheses, we conducted two experiments. For the first one, stakeholders were asked to share various statements, and then to evaluate the perceived quality of those different statements. For the second experiment, engineers were asked to evaluate the perceived quality of the previously collected statements. Doing so, we found out that, for both the stakeholders and the engineers, statements built on experience increase the objectivity of a statement, but that experience is not always desirable as it leads stakeholders to produce less creative RE entities. We also discovered that stakeholders perceive that experience reduces the risk of volatility while engineers perceive experiential statements as more complete than hypothetical statements. The work presented in this paper is still at an early stage, and opens a new avenue for research on software quality. We plan to keep working on the distinction between statements and RE entity, their modeling, and on the empirical evaluation.

References

Antoniou, G. (1998). The role of nonmonotonic representations in requirements engineering. *International Journal of Software Engineering and Knowledge Engineering*, *8*, 385–399.

Billington, D., Estivill-Castro, V., Hexel, R., Rock, A. (2011). Requirements engineering via non-monotonic logics and state diagrams. In *Maciaszek, LeszekA. and Loucopoulos, P., ed.: Evaluation of novel approaches to software engineering. Springer Berlin Heidelberg* (pp. 121–135).

Boden, M.A. (2013). Creativity In Berys Gaut and Dominic McIver Lopes, ed. The Routledge companion to aesthetics. Routledge Handbooks Online.

Brewka, G., Niemelä, I., Truszczynski, M. (2007). Nonmonotonic reasoning. In *Handbook of Knowledge Representation. Volume 3* (pp. 239–284).

Brewka, G., Roelofsen, F., Serafini, L. (2007). Contextual default reasoning. In *IJCAI International Joint Conference on Artificial Intelligence* (pp. 268–273).

Burnay, C. (2016). Are stakeholders the only source of information for requirements engineers? Toward a taxonomy of elicitation information sources.

Burnay, C., Jureta, I.J., Faulkner, S. (2014). An exploratory study of topic importance in requirements elicitation interviews. In *Proc. 26th International Conference on Advanced Information Systems Engineering (CAiSE'14)* (pp. 180–195).

Christel, M.G., & Kang, K.C. (1992). Issues in requirements elicitation. Technical Report CMU/SEI-92-TR-12 ESC-TR-92-012.

Davis, A.M., Dieste, O., Hickey, A.M., Juristo, N., Moreno, A.M. (2006). Effectiveness of requirements elicitation techniques: empirical results derived from a systematic review. In *Proc. 14th IEEE International Conference on Requirements Engineering* (pp. 179–188).

Easterbrook, S., Singer, J., Storey, M.A., Damian, D. (2008). Selecting empirical methods for software engineering research. In *Guide to Advanced Empirical Software Engineering* (pp. 285–311).

Elrakaiby, Y., Ferrari, A., Spoletini, P., Gnesi, S., Nuseibeh, B. (2017). Using argumentation to explain ambiguity in requirements elicitation interviews. In *Proceedings-2017 IEEE 25th International Requirements Engineering Conference RE 2017* (pp. 51–60).

Ferrari, A., Spoletini, P., Gnesi, S. (2015). Ambiguity as a resource to disclose tacit knowledge. In *2015 IEEE 23rd International Requirements Engineering Conference (RE)*, IEE (pp. pp. 26–35).

Fritz, C.O., Morris, P.E., Richler, J.J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology:, General*, *141*, 2–18.

Goguen, J.A., & Linde, C. (1993). Techniques for requirements elicitation. In *Proc. IEEE International Symposium on Requirements Engineering* (pp. 152–164).

Gotel, O.C.Z., & Finkelstein, A.C.W. (1994). An analysis of the requirements traceability problem. In *Proceedings of the First International Conference on Requirements Engineering, IEEE* (pp. 94–101).

Gėnova, G., Fuentes, J.M., Llorens, J., Hurtado, O., Moreno, V. (2013). A framework to measure and improve the quality of textual requirements. *Requirements Engineering*, *18*, 25–41.

Greenspan, S., Mylopoulos, J., Borgida, A. (1994). On formal requirements modeling languages: RML revisited. In *Proc. 16th International Conference on Software engineering* (pp. 135–147).

Hickey, A.M., & Davis, A.M. (2004). A unified model of requirements elicitation. *Journal of Management Information Systems*, *20*, 65–84.

Ionita, D., Bullee, J.W., Wieringa, R.J. (2014). Argumentation-based security requirements elicitation: the next round. In *2014 IEEE 1st International Workshop on Evolving Security and Privacy Requirements Engineering ESPRE 2014 - Proceedings* (pp. 7–12).

Jureta, I.J., Mylopoulos, J., Faulkner, S. (2008). Revisiting the core ontology and problem in requirements engineering. In *Proc. 16th IEEE International Conference on Requirements Engineering* (pp. 71–80).

Karlsson, J. (1996). Software requirements prioritizing. In *Proc. 2nd IEEE International Conference on Requirements Engineering* (pp. 110–116).

Maiden, N., Gizikis, A., Robertson, S. (2004). Provoking creativity: imagine what your requirements could be like. *IEEE Software*, *21*, 68–75.

Maiden, N., & Rugg, G. (1996). ACRE: selecting methods for requirements acquisition. *Software Engineering Journal*, *11*, 183–192.

McCarthy, J. (1980). Circumscription - a form of non-monotonic reasoning. *Artificial intelligence*, *13*, 27–39.

Merriam-Webster online (2019). Belief. https://www.merriamwebster.com/dictionary/Belief?utm_campaign=sd&utm_medium=serp&utm_source=jsonld.

Moore, R. (1984). Possible-world semantics for autoepistemic logic.

Nuseibeh, B., & Easterbrook, S. (2000). Requirements engineering: a roadmap. In *Proc. Workshop on the Future of Software Engineering. Volume 1. 35–46*.

Pelletier, F.J., & Elio, R. (2005). The case for psychologism in default and inheritance reasoning. *Synthese*, *146*, 7–35.

Pohl, K. (2010). Requirements engineering: fundamentals, principles, and techniques. Springer Publishing Company Incorporated.

Poole, D. (1988). A logical framework for default reasoning. *Artificial Intelligence*, *36*, 27–47.

Princeton University: About WordNet (2010). WordNetPrinceton University. <http://wordnet.princeton.edu>.

Rahwan, I., & Simari, G. (2009). Argumentation in artificial intelligence. Springer Berlin Heidelberg. Heidelberg.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, *13*, 81–132.

Robertson, S., & Robertson, J. (2012). Mastering the requirements process: getting requirements right Addison-wesley.

Rolland, C., Grosz, G., Kla, R. (1999). Experience with goal-scenario coupling in requirements engineering. In *Proc IEEE International Symposium on Requirements Engineering IEEE Comput. Soc* (pp. 74-81).

Saeki, M., Matsumura, K., ichi Shimoda Kaiya, H. (1996). Structuring utterance records of requirements elicitation meetings based on speech act theory. In *ICRE. 21*.

Searle, J.R. (1969). Speech acts: an essay in the philosophy of language. Cambridge university press.

Steup, M. (2014). Epistemology. In *Edward N. Zalta, ed.: The Stanford Encyclopedia of Philosophy. Spring 201 edn.* (pp. 1–34).

Sutcliffe, A., & Ryan, M. (1998). Experience with SCRAM, a scenario requirements analysis method. In *Proc. 3rd IEEE International Conference on Requirements Engineering* (pp. 164–171).

Sutcliffe, A., & Sawyer, P. (2013). Requirements elicitation: towards the unknown unknowns. In *Proc. 21st IEEE International Requirements Engineering Conference RE IEEE* (pp. 92–104).

Van Zee, M., Bex, F., Ghanavati, S. (2015). Rationalization of goal models in GRL using formal argumentation. In *2015 IEEE 23rd International Requirements Engineering Conference RE 2015 - Proceedings* (pp. 220–225).

Yu, Y., Tun, T.T., Tedeschi, A., Franqueira, V.N., Nuseibeh, B. (2011). Openargue: supporting argumentation to evolve secure software systems. In *Proceedings of the 2011 IEEE 19th International Requirements Engineering Conference RE 2011* (pp. 351–352).

Zowghi, D., & Coulin, C. (2005). Requirements elicitation: a survey of techniques, approaches, and tools. In *Aurum, Aybüke and Wohlin, C., ed.: Engineering and managing software requirements. Springer Berlin Heidelberg 19–46*.

Zowghi, D., Ghose, A., Peppas, P. (1996). A framework for reasoning about requirements evolution.

Zowghi, D., & Offen, R. (1997). A logical framework for modeling and reasoning about the evolution of requirements. In *Proc. 3rd IEEE International Symposium on Requirements Engineering* (pp. 247–257).

Corentin Burnay is assistant professor at the University of Namur. After a Ph.D. and a postdoc on business analysis and incompleteness in information elicitation, he developed a significant interest in the engineering of Business Intelligence solutions, with a focus on the practice of corporate dashboard designs. Corentin was the recipient of a Research Fellow (Aspirant) position with the Belgian national research fund (Fonds de la Recherche Scientifique FNRS, Brussels) from 2012 to 2016. He has published over fifteen research papers within the fields of requirements engineering and information systems. He is the recipient of the distinguished paper award from CAiSE 2014, and the organizer of the international MORE-BI workshop since 2015.

Sarah Bouraga is assistant professor at the University of Namur. She completed a Ph.D on requirements engineering for online social networks. Sarahs other research interests include business and data analysis, business intelligence, and decision-making. She has also been coorganizing the international MORE-BI workshop since 2015.

Joseph Gillain is Researcher at the University of Namur. He is currently completing a Ph.D. thesis on the formalization of optimization in requirement engineering of information systems. His work deals with Operational Research and Variability in Requirement Problems. Joseph designed several optimization models dealing amogn others with software product lines, self-configurable systems.

Ivan J. Jureta holds the Senior Researcher (Chercheur qualifi) position with the Belgian national research fund (Fonds de la Recherche Scientifique FNRS, Brussels) and is Associate Professor with the Department of Business Administration, University of Namur.

Ivans principal research interest is in method engineering and method automation, focusing on the elicitation, modelling and analysis of knowledge that human experts apply in problem solving and decision making, the engineering of ontologies and processes capturing that knowledge, and the automation of the said processes. This interest falls within the various research fields concerned with the transfer, preservation and automation of knowledge.

Ivan is the author of Analysis and Design of Advice (Springer, 2011), has published over 60 research papers on these topics within the fields of requirements engineering, business analysis, and conceptual modeling of information systems, and serves on scientific committees of the International Conference on Advanced Information Systems Engineering (CAiSE), the International Conference on Conceptual Modeling (ER), and the International Requirements Engineering Conference (RE). He is the recipient of the 2008 IEEE RE conference best paper award, and the 2014 CAiSE conference distinguished paper award. He has organized and chaired the series of International Workshops on Modeling and Reasoning for Business Intelligence (MORE-BI), held in Brussels in 2011 and Florence in 2012. For the most current list of research publications, visit DBLP.