CrossMark

ORIGINAL ARTICLE

A questionnaire-based survey methodology for systematically validating goal-oriented models

Jameleddine Hassine · Daniel Amyot

Abstract Goal models represent interests, intentions, and strategies of different stakeholders. Reasoning about the goals of a system unavoidably involves the transformation of unclear stakeholder requirements into goal-oriented models. The ability to validate goal models would support the early detection of unclear requirements, ambiguities, and conflicts. In this paper, we propose a novel validation approach based on the Goal-oriented Requirement Language (GRL) to check the correctness of GRL goal models through statistical analyses of data collected from generated questionnaires. System stakeholders (e.g., customers, shareholders, and managers) may have different objectives, interests, and priorities. Stakeholder conflicts arise when the needs of some group of stakeholder compromise the expectations of some other group(s) of stakeholders. Our proposed approach allows for early detection of potential conflicts amongst intervening stakeholders of the system. In order to illustrate and demonstrate the feasibility of the approach, we apply it to a case study of a GRL model describing the fostering of the relationship between the university and its alumni. The approach brings unique benefits over the state of the art and is complementary to existing validation approaches.

J. Hassine (✉)
College of Computer Science and Engineering,
King Fahd University of Petroleum and Minerals,
P.O. Box 1621, Dhahran 31261, Kingdom of Saudi Arabia
e-mail: jhassine@kfupm.edu.sa

D. Amyot
School of Electrical Engineering and Computer Science,
University of Ottawa, 800 King Edward St., Ottawa,
ON K1N 6N5, Canada
e-mail: damyot@eecs.uottawa.ca

1 Introduction

There is a general consensus on the importance of good requirements engineering (RE) approaches for achieving high-quality software. Requirements elicitation, modeling, analysis, validation, and management are amongst the main challenges faced during the development of complex systems.

A common starting point in requirements engineering approaches is the elicitation of goals that the targeted system will need to achieve once developed and deployed. Goal modeling can be defined as the activity of representing and reasoning about stakeholder goals using models, in which goals are related through links with other goals and/or other model elements such as tasks that the system is expected to execute, resources that can be used, or roles that can be played [29].

Over the past two decades, several goal modeling languages have been developed. Amongst the most popular ones, we find i* [54], the NFR Framework [9], Keep All Objects Satisfied (KAOS) [50], Tropos [16], the Business Intelligence Model (BIM) [20], and the Goal-oriented Requirement Language (GRL) [26] part of the ITU-T standard User Requirements Notation (URN). Ayala et al. [6] have presented a comparative study of i* [54], Tropos [16], and GRL [26]. The authors have identified (1) eight structural criteria that consider the characteristics of the language constructs and are related to models, actors, intentional elements, decomposition elements, additional reasoning elements, and external model elements, and (2)

Springer

six non-structural criteria that analyze the definition of the languages, their use, and also the elements that complement them such as formalizations, methodologies, and software tools. These criteria are mainly syntactical. Several tools supporting modeling and analysis features for these languages are available, and Almeida et al. recently published a systematic comparison [2].

As goal models gain in complexity (e.g., large systems involving many stakeholders who are often distributed), they become difficult to analyze and to validate. Indeed, tentative requirements provided by the stakeholders of complex systems may be, among others, ambiguous, contradictory, and vague, which may cause many issues when the requirements engineer transforms such requirements (expressed usually in natural language) into a formal syntax in a specific goal description language. As incorrect system requirements generated from goals can lead to cost, delay, and quality issues during system development, it is essential to ensure the validity of the source goal models. Most of the work targeting the validation of goal models focuses on the dynamic evaluation of satisfaction through propagation algorithms [3, 5, 8, 9, 16, 21–23, 29]. To the best of our knowledge, no empirical approach has been proposed to establish the validity of goal models with stakeholders.

Conflicts emanate when system stakeholders (i.e., individuals) or categories/groups of stakeholders (e.g., employees, customers, shareholders, and managers) have different perceptions of a particular aspect of the system. It is important to detect such conflicts early in the system development life cycle. The main motivation of this research is to validate goal models. In particular, we focus on the use of empirical data collected from surveys administered to the system stakeholders to help validate GRL model artifacts and to identify and detect conflicts (that may arise amongst intervening stakeholders) in GRL goal models. In the rest of paper, we use the term *group* to denote a category of stakeholders.

This paper aims to:

- Systematically create questionnaires from goal models (i.e., map GRL elements and relationships to natural language survey questions readable by non-experts) that are administered to system stakeholders, according to a proposed selection policy. Furthermore, our survey-based approach allows for the validation of goal models by a large number of stakeholders who can be distributed geographically. Indeed, involving a large number of stakeholders in the validation process can increase our confidence in the goal model and promote its acceptance.
- Extend our previous work [18] by providing an enhanced survey vocabulary and by extending the

questions to include: (1) quantitative contributions, (2) qualitative and quantitative importance of goals, tasks, and resources, (3) many dependency configurations, and (4) indicators and actors.

- Tackle the issue of validating complex goal models using empirical data that can be analyzed using proven statistical methods such as *t* test [17] and *ANOVA (Analysis of Variance)* [13]. Applying statistical analysis to the collected data provides a way to quantify the confidence we have in the validity of goal models. In addition, statistical methods, e.g., ANOVA and *t* test, allow for the measurement and the assessment of the differences between respondent groups, which represents an essential tool to detect significant conflicts between intervening stakeholders.
- Address scalability issues when validating large models (with hundreds of links). We provide an optimization strategy to minimize the number of generated questions.
- Identify and analyze the possible interactions between system stakeholders, in order to avoid the administration of all questions to all stakeholders.

We have chosen GRL [26] as target language, given its status as an international standard, but our proposed methodology can likely be applied to other goal-oriented languages that visually support actors, intentional elements, and their relationships (including *i** and Tropos), thus maintaining the discussion generic.

The remainder of this paper is organized as follows. The features of the GRL language are briefly overviewed in Sect. 2. In Sect. 3, we present and discuss the proposed validation approach for GRL models. Section 4 provides guidelines for designing the validation survey from a goal model. The analysis of stakeholder dependencies and interactions is discussed in Sect. 5. Next, empirical data analysis is presented in Sect. 6 and applied to a case study of a GRL model describing the fostering of the relationship between a university and its alumni (Sect. 7). A discussion of benefits and threats to validity is provided in Sect. 8. A comparison with related work is presented in Sect. 9. Finally, conclusions and future work are presented in Sect. 10.

2 GRL in a nutshell

The Goal-oriented Requirement Language (GRL) [26] is a visual notation used to model stakeholders' intentions, business goals, and non-functional requirements (NFR). GRL integrates important concepts from:

1. The NFR Framework [9], which focuses on the modeling of NFRs and the various types of

relationships between them (e.g., AND/OR decomposition, as well as positive and negative contributions). The NFR Framework comes with goal decomposition strategies along with propagation algorithms to estimate the satisfaction of higher-level goals given the attainment or non-attainment of lower-level ones.
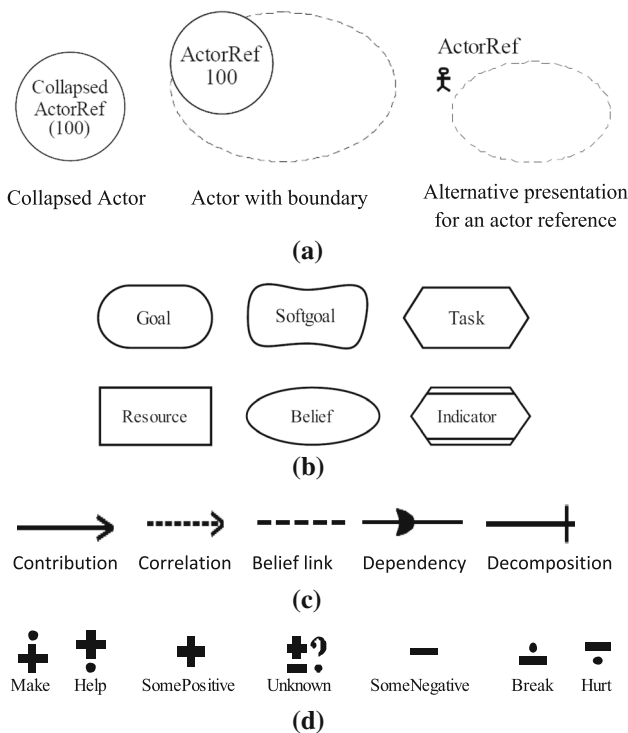
2. The $i*$ goal modeling language [54], which has as primary concern the modeling of intentions and strategic dependencies between actors. Dependencies between actors concern goals, softgoals, resources and tasks.

The basic notational elements of GRL are summarized in Fig. 1. Actors (see Fig. 1a) are holders of intentions; they are the active entities in the system or its environment who want goals to be achieved, tasks to be performed, resources to be available, and softgoals to be satisfied [26]. Actor definitions are often used to represent stakeholders as well as systems. A GRL actor may contain intentional elements and indicators describing its intentions, capabilities, and related measures. Figure 1b illustrates the GRL intentional elements (i.e., goal, task, softgoal, resource, and belief) that optionally reside within an actor. A GRL indicator is a GRL-containable element used for models that base GRL model analysis on real-world measurements. An indicator may convert real-world values into GRL satisfaction values by comparing the former against target, threshold, and worst-case values.

Figure 1c illustrates the various kinds of links in a goal model. Decomposition links allow an intentional element to be decomposed into sub-elements (using AND, OR, or XOR). Beliefs, used to represent rationales from model creators, are connected to GRL intentional elements using belief links. Contribution links indicate desired impacts of one element (intentional or indicator) on another element. A contribution link has a qualitative contribution type (see Fig. 1d) and/or a quantitative contribution (an integer value between -100 and 100 in standard GRL). Correlation links are similar to contribution links but describe side effects rather than desired impacts. Finally, dependency links model relationships between actors. For a detailed description of the GRL language, the reader is invited to consult the URN standard [26].

In such goal models, determining the type of intentional element is usually simple. The type of decomposition links (AND, OR, or XOR) is also usually simple to assess, although deciding we have enough (or too many) sub-elements or alternatives is a harder question. The existence of dependencies is also subject to discussions among stakeholders. In addition, whether contributions and correlations from indicators and intentional elements to other intentional elements should be positive or negative, and to what extent (including nonexistence), often leads to divergent opinions. The need for contributions and correlations, as well as that level they should have, are often the subject of re-assessment when a goal model evolves, for example, through additional refinement or decomposition [10, 39, 40]. These are amongst the many reasons why goal-oriented model validation is required.
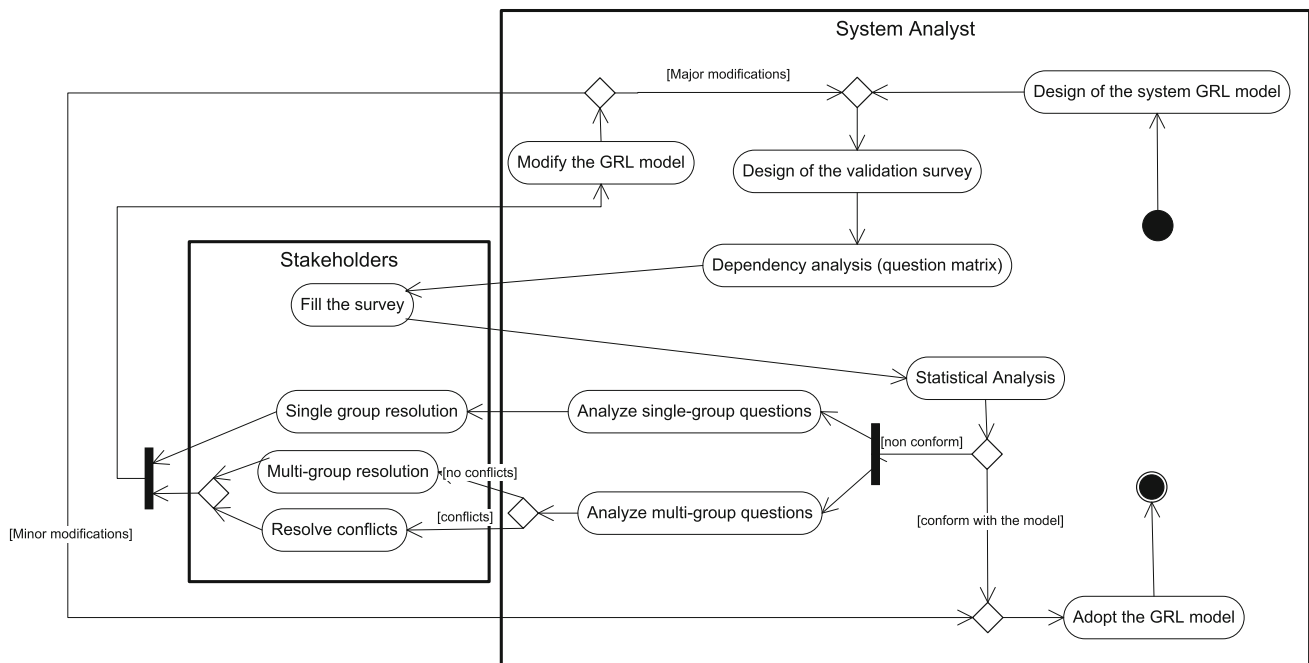
3 GRL-based validation approach

The activity diagram in Fig. 2 illustrates the steps of our GRL-based goal model validation approach. The process starts with the design of the system GRL model. In this step, the requirements engineer or system analyst plays a central role in shaping the problem and solution knowledge, provided by the system stakeholders (through interviews, surveys, workshops, or other such requirements elicitation techniques), into a GRL model. The resulting GRL model is used to design the *validation survey*, i.e., a questionnaire to be given to the system stakeholders. Questions often involve simple answers on a scale that goes from 1 (strongly agree) to 7 (strongly disagree). The generated survey maps closely the structure and the semantics of the GRL model under analysis.

Next, a dependency analysis and a classification of the survey questions are performed (see Sect. 5). The output of

Fig. 1 GRL notational elements, **a** GRL actors, **b** GRL intentional elements and indicators, **c** GRL links, **d** GRL qualitative contribution types

Fig. 2 GRL validation approach

the later step is a questionnaire matrix that specifies which question is associated with which stakeholder group, so only relevant questions are asked to a given stakeholder. The resulting stakeholder-specific surveys are then filled by their respective stakeholders. In case the GRL model describes a socio-technical system (STS) that involves social actors (human and organizations) and technical subsystems, developers may also be considered as stakeholders, amongst others, of the technical systems. Hence, the survey questions associated with the technical system GRL sub-model may be administered to the system developers. Finally, the collected survey data are statistically analyzed to determine whether the GRL model requires modifications (which should be re-assessed if they are major) or whether it can be adopted.

Difficulties arise when many stakeholders with different backgrounds and goal modeling expertise levels (if any) participate in the engineering of requirements over a long period of time, which hinders the quality of the goal model.

We define a conflict, with respect to a specific sub-model, as a situation in which at least two groups of stakeholders have different views of the sub-model under analysis. These views are captured quantitatively using one attitudinal question generated from the sub-model. The agreement of a group of stakeholders with the sub-model is expressed by a positive answer (i.e., that corresponds to a value within a [1, 3] interval), while a disagreement is expressed by a negative answer (i.e., that corresponds to a value within a [5, 7] interval). The disagreement between intervening groups is exhibited when we have major

differences (i.e., statistically significant differences) in the answers of the stakeholders to the survey questions associated with the GRL sub-model. Note that we do not distinguish here the sources of conflicts, but the latter may arise for various reasons such as different needs, understandings, priorities, or social values amongst stakeholders. The outcome of the statistical analysis is discussed in the following sections.

3.1 Analysis data conforming with the GRL model

Participants' answers conform with the expected values derived from the GRL model (i.e., the stakeholders agree on all survey questions). In this case, the GRL model is considered valid and it is adopted as is.

3.2 Analysis data not conforming with the GRL model

Answers to some questions do not conform with the GRL model (i.e., responses within the [5, 7] range, indicating disagreement). In this case, the questions with such negative answers are analyzed in order to understand the nature of the disagreement (different needs, understanding, priorities, social values, etc.). Two major situations can be distinguished:

1. Questions administered to a single group of stakeholders: The analysis of statistical descriptives (e.g., mean, standard deviation, skewness, and kurtosis.) along with the histograms may lead to:

- A statistically significant portion of the participants disagree with the question statement. In this case, the stakeholders within the group should be consulted and the problematic GRL artifact (corresponding to this specific question) should be modified to reflect the stakeholders' view.
- Participants' answers are substantially different from one another (e.g., flat histogram). These differences should be discussed within the stakeholders' group. The GRL artifact might be modified as a result of a resolution within the group.

2. Questions administered to more than one group of stakeholders: The statistical data analysis (e.g., t test [17] and ANOVA [13]) may lead to:

- Different groups of participants agree between themselves (i.e., no conflicts) but their agreement is not conform with the model. For example, all involved stakeholders disagree with a question statement (corresponding to one specific model artifact). In this case, the stakeholders should be consulted and the problematic GRL artifact (corresponding to those questions) should be revised and modified in order to reflect the view of all stakeholders.
- There is a conflict between the intervening groups. For example, one group agrees with a question statement while another group disagrees with it. In such a situation, the stakeholders are consulted and made part of the conflict resolution process. Resolution can be achieved through a mediation process or through less formal discussions between the different parties. Please note that conflict resolution could be done through many existing approaches, and as such is out of the scope of this paper.

Next, the requested modifications are incorporated into the GRL model. For major modifications, such as the deletion of many GRL elements/links or the modification of link decompositions types, an additional iteration may be required. The process stops when satisfactory results are obtained, leading to a valid GRL model.

4 Design of the validation survey

The design of the validation survey, to be administered to the system stakeholders, represents the corner stone of our proposed validation approach. Stakeholders include anyone who has a vested interest in the system (e.g., customers, different types of end users, system developers, and regulators). The survey questions should be produced based on the GRL model's intentional elements (goals, tasks, etc.) and links (dependencies, decompositions, contributions, etc.) for the target stakeholders (GRL actors).

4.1 Types of survey questions

Depending on the type of collected information, different types of questions can be designed [45]:

1. *Classification questions:* They are used to collect and classify information about participants. Classification questions are required, for example, to check that an acceptable quota of people or companies has been interviewed. Note that the number of groups depends on the number of categories of stakeholders involved being modeled. Typical classification questions provide a profile of the participants—by finding out their age, their sex, their social class, their marital status, their position (role) within a company, etc. Furthermore, they are used to compare and contrast the different answers of one group of participants with those of other groups.

2. *Attitudinal questions:* Attitudes are opinions or beliefs that people have about subjects or topics—they can be deemed to be favorable or unfavorable. Matters of opinion are collected by attitudinal questions. As the term suggests, these questions seek to uncover people's beliefs and thoughts on a subject. Attitudinal questions address, among others, the following: "What do you think of ...?", "Do you agree or disagree with ...?", "How do you rate ...?", etc. Typically, an attitudinal question consists of a statement for which stakeholders are asked to express their agreement or disagreement. A 5-point or 7-point Likert scale [36] may be used to measure the level of agreement or disagreement. The format of a typical 7-point Likert-item scale is as follows:

1. Strongly agree
2. Agree
3. Agree somewhat
4. Neither agree nor disagree (Undecided)
5. Disagree somewhat
6. Disagree
7. Strongly disagree

As our survey targets validation, it is composed of attitudinal questions only.

A further variation on classification or attitudinal questions is that they can be open-ended or closed. Open-ended questions, as the name suggests, leave the participant free to give any answer. A closed question is a question that

limits participants with a list of answer choices that have been anticipated (e.g., a scale). We focus here on closed questions, for which analysis can be automated. Additional open-ended contingency questions could be used to collect suggestions in case of disagreement, as explored in our previous work [18]. However, they require much effort from participants (which might prevent them from participating fully; this effort might also be wasted if everyone disagrees), they require manual interpretation by analysts, and they are actually not needed for conflict resolution if stakeholders are involved through other means.

4.2 Survey questions vocabulary

Creating well-structured, simply-written survey questions helps in collecting valid responses. While there are no predefined rules on the wording of survey questions, there are some basic principles such as *relevance* and *accuracy* [24] that do work to improve the overall survey design. Although generic, the intent of Table 1 is to provide some tips on how to derive question vocabulary from goal model constructs. The presented examples of question vocabulary are actually derived from the inherent definition of goal model constructs. However, to produce relevant, accurate, and well-understood surveys, the designed questions may include technical words from the targeted domain. Therefore, this exercise is currently done manually, which also enables one to accommodate the many ways in which people write goals (e.g., "maximize profits" versus "profits be maximized"). The automation of survey question generation for models in specific goal-oriented languages is left for future work.

Table 1 Examples of GRL constructs and their corresponding questions vocabulary

GRL intentional elements	
Construct	Question vocabulary
Goal <goal-id>	The full realization/fulfillment of <goal-id>
Softgoal <softgoal-id>	The sufficient achievement of <softgoal-id>
Task <task-id>	The completion/execution of <task-id>
Resource <res-id>	Uses <res-id>
Indicator <ind-id>	Reaching the target of <ind-id>
Belief <text>	We believe that <text>
Actor <actor-id>	Actor <actor-id> participates in
Actor with boundary <actor-id>	Actor <actor-id> encloses

GRL intentional relations	
GRL contribution links	Question vocabulary
Make	Contributes positively and sufficiently to
SomePositive	Has some positive contribution (with unknown extent) to
Help	Helps (has a positive but insufficient contribution)
Unknown	Has some contribution to (but the extent and the degree (positive or negative) are unknown)
Hurt	Hurts (has a negative but insufficient contribution)
SomeNegative	Has some negative contribution (with unknown extent) to
Break	Contributes negatively and sufficiently to

GRL correlation links	Question vocabulary
SomePositive	Has some positive correlation (side effect) on
SomeNegative	Has some negative correlation (side effect) on

GRL decomposition links	Question vocabulary
Decomposition AND	The satisfaction of each of the sub-intentional elements is necessary to achieve the target
Decomposition OR	The satisfaction of one of the sub-intentional elements is sufficient to achieve the target, but many sub-intentional elements can be satisfied
Decomposition XOR	The satisfaction of one and only one of the sub-intentional elements is necessary to achieve the target

Attitudinal Question:
"The <u>**execution**</u> of **Task1** <u>**helps**</u> the <u>**realization**</u> of **Goal1**".
Please tell us to what extent you agree with this statement ?
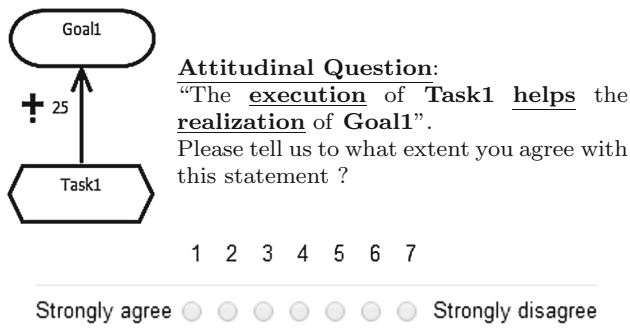
Fig. 3 Contribution of type **Help** and its associated question

Figure 3 illustrates an example of a contribution relationship of type **HELP** between task *Task1* and goal *Goal1*, and its associated attitudinal survey question. Specific words are used to describe the relationship type (i.e., verb **helps**), and the involved participants with appropriate achievement description (i.e., **completion of task Task1**, **realization of goal Goal1**).

A GRL dependency describes how a source actor definition (the depender) depends on a destination actor definition (the dependee) for an intentional element (the dependum). Often, the modeler will use two consecutive dependency links (from depender to dependum, and from dependum to dependee) to express detailed dependencies, but dependencies can be used in more generic situations as well (see Table 2). The dependum specifies what the dependency is about, i.e., the intentional element around which a dependency relationship centers. With an intentional element as a source of the dependency, the depender may specify why it depends on the depender for the dependum. With an intentional element as a target of the dependency, the dependee may specify how it is required to provide or satisfy the dependum.

Dependencies enable reasoning about how actor definitions depend on each other to achieve their goals. Dependency links can be used in a number of configurations including but not limited to the ones described in Table 2. According to the required level of detail, intentional elements inside actor definitions can be used as source and/or destination of a dependency link.

4.3 Validation of quantitative contributions

In addition to the qualitative validation, presented in the previous section, we may want to validate the quantitative weight of contributions. To do this, a preliminary interval decomposition is needed. One possible mapping (which is implemented in jUCMNav [28], an eclipse-based tool that supports GRL modeling) is depicted in Table 3. It is worth noting that crossing the boundaries between "qualitative" and "quantitative" realms is recognized as a challenge [19].

Although the proposed partitioning of the range of numbers is a reasonable one, it is not necessarily absolute. Our proposed analysis approach still works even if the interval boundaries are changed or relaxed, as needed.

For example, in addition to the qualitative question presented in Fig. 3 and assuming a positive answer, we may validate the weight of the help contribution using the following question:

"The <u>**execution**</u> of **Task1** <u>**helps**</u> the <u>**realization**</u> of **Goal1**. The weight of the contribution is evaluated to 25 on a [1, 49] scale (1 being the lowest value, 49 being the highest value). Should the contribution weight be:

1. Much more
2. Somewhat more
3. Fine as is
4. Somewhat less
5. Much less"

Similarly, we can validate quantitative contributions expressed as interval ranges.

Another alternative would be a quantitative validation question, where the participant provides his input in terms of a contribution weight (i.e., without using a Likert scale). Such a question may have the following form:

"The <u>**execution**</u> of **Task1** <u>**helps**</u> the <u>**realization**</u> of **Goal1**. The weight of the contribution is evaluated to 25 on a [1, 49] scale (1 being the lowest value, 49 being the highest value). In your opinion, what would be the appropriate weight of the contribution? Please rate the weight on a [1, 49] scale: ..."

However, such questions would slowly bring participants into a conflict *resolution* space, for which better techniques (some of which being discussed in Sect. 8) are available.
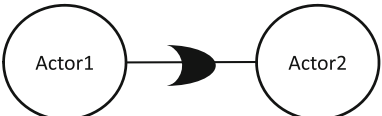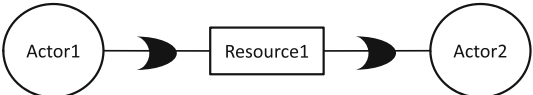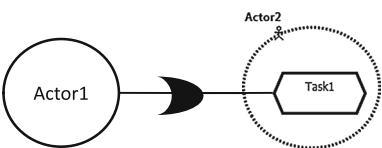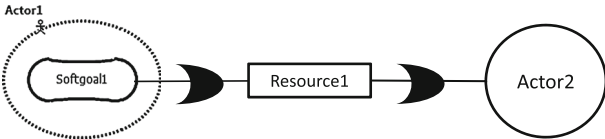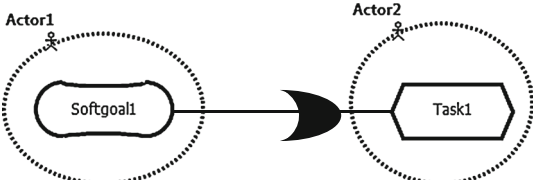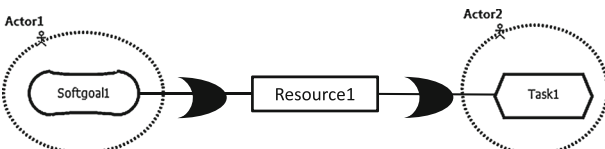
4.4 Validating the importance of intentional elements

A GRL intentional element can be given an importance level when included in an actor. GRL-based analysis approaches [3] use this importance to compute the overall satisfaction level of the containing actor. Such importance level can be qualitative (e.g., *High*, *Medium*, *Low*, or *None*, which is the default) or quantitative (between 0 and 100, 0 being the default and lowest value and 100 being the highest). The GRL diagram in Fig. 4 shows a goal and a task contained by an actor. The goal has an importance value of 100 (i.e., *High*) whereas the task has an importance value of 25 (i.e., *Low*).

In our approach, to validate the importance of an intentional element, we may use:

1. A validation question similar to the one used in Fig. 3, where the participant expresses his agreement or disagreement with the existing importance values.

Table 2 GRL dependencies configurations and their corresponding questions vocabulary

GRL dependencies

GRL dependency configuration	Question vocabulary
	Actor1 depends on Actor2
This is a configuration that is typical of preliminary goal models that require further refinement. A dependency is identified, but what, why, and how are still unknown	
	Actor1 depends on Actor2 to provide Resource1
This is a configuration that focuses solely on strategic dependencies between actors. Why and how the dependum is provided are unknown	
	Actor1 depends on Actor2 to execute Task1.
The dependum and why it is required are unknown	
	To achieve Softgoal1, Actor1 depends on Actor2 to provide Resource1
How the dependum is provided is unknown	
	To achieve Softgoal1, Actor1 depends on Actor2 to execute Task1
The dependum is unknown	
	To achieve Softgoal1, Actor1 depends on Actor2 to provide Resource1 by executing Task1
This is a configuration that details the dependum (Resource1) together with why it is required and how it is provided	

2. A qualitative validation question, where the participant provides his input in terms of a qualitative evaluation of the element importance. Such a question may have the following form (corresponding to the *ActorGoal1* in Fig. 4):

   "In your opinion, how important to you is *ActorGoal1*:

   1. Not at all important
   2. Low importance
   3. Slightly important
   4. Neutral
   5. Moderately important
   6. Very important
   7. Extremely important"

3. A quantitative validation question, where the participant provides his input in terms of a quantitative evaluation of the element importance. Such a question may have the following form:

   "In your opinion, how important to you is ActorGoal1. Please rate the importance from 0 and 100 [0 being not important at all, 100 being extremely important]": ...

Using options 2 and 3 would help determine the potential qualitative/quantitative deviations between the original

Table 3 Quantitative contribution intervals

Contribution type	Quantitative interval
Make	100
SomePositive	[50, 99]
Help	[1, 49]
Unknown	0
Hurt	[−49, −1]
SomeNegative	[−99, −50]
Break	−100
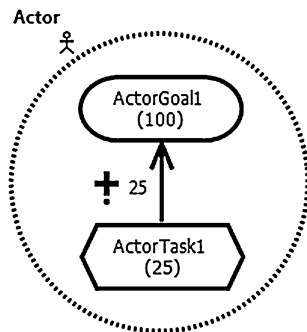
Fig. 4 GRL actor that contains intentional elements with importance values

value (taken from the GRL model) and the participant's response.

4.5 Validation survey optimization

The validation survey should reflect all model artifacts and cover both qualitative and quantitative aspects of a GRL model. Although many useful GRL models we have observed in the past had a handful of actors and a few dozen intentional elements, some models are much larger in size [46]. Deriving survey questions from GRL models with hundreds of elements and links may lead to scalability issues. One mitigation approach is to minimize the number of generated questions. This can be achieved by clustering questions for GRL *sub-models*, as follows:

- In decomposition links, all children are included in the sub-model.
- Beliefs, associated with intentional elements, are included in sub-models and are not assessed separately.
- The importance of intentional elements can be included in a sub-model question and is not assessed separately.

Figure 5 illustrates an example of an attitudinal question describing a *help* relationship and assessing the *belief* as part of the sub-model.

Although such optimization helps reduce the number of questions and improve the manageability of the survey, the applicability of such recommendations is left to the full discretion of the system analyst.

Table 4 characterizes the number of expected survey questions corresponding to each GRL construct.
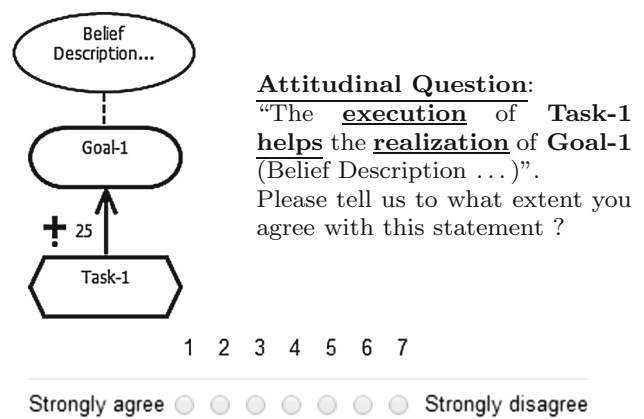
5 Analysis of stakeholder interactions

Due to the number of stakeholders and potential complexity of GRL models, survey optimization is helpful but may not be sufficient in reducing the cost of the overall validation process. Indeed, deriving and administrating all questions to all stakeholders quickly become very costly. Asking too many questions may lead to decreased data quality and/or abandoned surveys.

In order to reduce costs and the risk of losing participants, a general guideline is to have one survey per type of stakeholders (i.e., per GRL actor), which only includes questions that pertain to the intentional elements and relationships contained by that actor. However, stakeholders often have dependencies and other types of interactions that can lead to additional shared questions. We propose an approach to identify and analyze the possible interactions between system stakeholders. In this research, stakeholders are captured using GRL actors (and their corresponding boundaries) and the interactions between stakeholders are expressed using explicit and/or implicit dependencies (Sect. 5.1), and URN links (Sect. 5.2).

In this approach, we first identify the participating stakeholders and then determine their potential interactions.

5.1 GRL actor dependencies

Dependencies between stakeholders can be classified as explicit or implicit. *Explicit dependencies* are modeled as dependency links, while *implicit dependencies* are modeled

Attitudinal Question:
"The **execution** of **Task-1 helps** the **realization** of **Goal-1** (Belief Description …)".
Please tell us to what extent you agree with this statement ?

**Fig. 5** GRL sub-model including a belief element and its corresponding question

Table 4 Characterization of the validation survey size

GRL construct	Expected number of generated questions per GRL construct
Contribution link	1 (either qualitative or quantitative)
AND decomposition	2 (one for checking necessity and one for checking sufficiency)
OR decomposition	2 (one for checking necessity and one for checking sufficiency)
XOR decomposition	2 (one for checking necessity and one for checking sufficiency)
Intentional element importance	1 (either qualitative or quantitative)
Actor importance	1 (either qualitative or quantitative)
Group of qualitative contributions (connected to the same element)	3 (one for checking necessity, one for checking sufficiency, and one for checking the qualitative weight distribution)
Group of quantitative contributions (connected to the same element)	3 (one for checking necessity, one for checking sufficiency, and one for checking the quantitative weight distribution)
Dependency	1
Belief	1
Indicator	3 (for checking its quantitative parameters: target, threshold, and worst-case values)

using contributions crossing actor boundaries. Figure 6 illustrates a GRL model having one explicit dependency between *Actor1Goal2* and *Actor2Goal1* goals, and one implicit dependency between task *Actor1Task1* and goal *Actor2Goal1*. Furthermore, since the execution of task *Actor2Task1* contributes to the realization of the goal *Actor2Goal1*, such a contribution has an indirect impact on the achievement of goal *Actor1Goal2*. Not only do we include the intentional elements involved in the implicit dependencies, but also their other direct incoming/outgoing contributions, if any (only one level, not complete subtrees).

Hence, the survey questions derived from the model in Fig. 6 is administered to actors *Actor1* and *Actor2* as follows:

- The explicit dependency question involving goals *Actor1Goal2* and *Actor2Goal1* is administered to both actors.
- The implicit dependency's questions relative to the help contributions involving (1) *Actor1Task1* and *Actor2Goal1*, (2) *Actor2Task1* and *Actor2Goal1*, and (3) *Actor1Task1* and *Actor1Goal1* are administered to both actors.
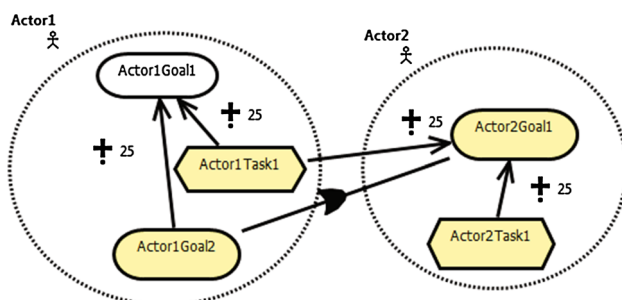
Fig. 6 Actors' dependencies example

- The question relative to the help contribution from *Actor1Goal2* to *Actor1Goal1* is administered to Actor1 only.

5.2 GRL actors with URN links

The use of URN links is another form of dependency description. Figure 7 illustrates the use of a URN link between two actors *Agent1* and *Agent2*. The symbols ▶ and ◀ indicate the presence of a source and a target URN link, respectively.

The association between *Agent1* and *Agent2* is of type *IsAssociatedWith* here. As URN links are typed user-defined extensions, they do not have a visual representation in GRL. However, such relationship can still be exploited during analysis [4]. All questions relative to the interrelated actors (i.e., connected with URN links) should be administered to both stakeholders.

6 Validation survey data analysis

Our main goal is to check whether stakeholders (survey participants) agree on the proposed GRL model. Errors and conflicts arise when we have major differences in the answers of the stakeholders. Therefore, such issues should be addressed and resolved.

As shown in Fig. 8, our model analysis strategy is based on the data collected from the attitudinal questions. The respondents' responses to each attitudinal question may be analyzed using: (1) descriptive statistics (in case there is only one group involved), (2) independent sample t test [17] (in case there are exactly two groups involved), or (3) one-way ANOVA [13] (in case there are at least three groups involved).
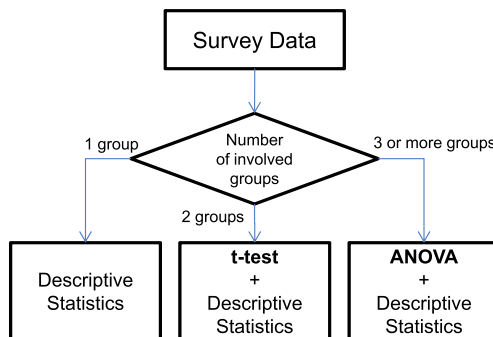
Fig. 7 URN Actor Links Example

Fig. 8 Selection of statistical analysis

A common question in the literature is whether it is legitimate to use Likert scale data in parametric statistical procedures that require interval data, such as t test [17] and ANOVA [13]. This topic is the subject of considerable disagreement [27, 32, 41]. A Likert scale is ordinal in nature, but when it is symmetric and when the items are equidistant, it behaves more like an interval-level measurement [32]. In addition, many researchers consider that parametric statistics are robust [33, 41] with respect to violations of such an assumption and parametric methods can be utilized without concern for "getting the wrong answer" [41].

We use the SPSS software (Release 16.0) [25] to generate statistical descriptives, to perform t test [17], and to run one-way ANOVA [13] analysis.

6.1 Descriptive statistics

Descriptive statistics describe quantitatively the main features of a collection of responses. Common measures include the mean, median, standard deviation (and/or variance), and 95 % confidence intervals for the dependent variables for each separate group as well as when all groups are combined. In our context (i.e., a 7-point symmetrical Likert scale), the group mean relative to a question would indicate whether the group members agree (e.g., mean is between 1 and 3) or disagree (e.g., mean is between 4 and 7) with the corresponding GRL sub-model. In addition, *skewness* and *kurtosis* provide some information concerning the distribution of scores. Positive skewness values indicate that scores are clustered to the left, while negative skewness values indicate a clustering of scores to the right. Positive kurtosis values indicate that the distribution is rather peaked (i.e., clustered in the center), while negative kurtosis values indicate a distribution that is relatively flat (too many cases in the extremes). Tabachnick and Fidell [48] suggest inspecting the shape of the distribution using a histogram. In this research, we will analyze histograms annotated with normal curve (see "Appendix" for examples).

6.2 Independent sample t test

The independent sample t test [17] evaluates the difference between the means of two independent group subjects. It evaluates whether the means of two independent groups are significantly different from each other, with respect to some continuous variable.

The independent sample t test null hypothesis can be written as follows:

$$H_0 : \mu_1 = \mu_2 \tag{1}$$

where μ_1 stands for the mean of the first group and group mean and μ_2 stands for the mean of the second group.

The first step of the independent sample t test is to run Levene's test [34] for equality of variances. This tests whether the variances of scores for the two groups are the same. The outcome of this test determines which of the t values that SPSS provides is the correct one to use (see Fig. 9):

- If the significance level (i.e., *Sig.*) is larger than 0.05 (a priori $\alpha = 0.05$), the first line of the t test table produced by SPSS, referring to the equal variances assumption, is selected.
- If the significance level of Levene's test is 0.05 or less, this means that the homogeneity of variances assumption is violated. Hence, the second line of the t test table should be used.

The second step consists on assessing differences between the groups. To find out whether there is a significant difference between the two involved groups, we refer to the *Sig. (2-tailed)*:

- If the value of the *Sig. (2-tailed)* is equal to or less than 0.05, there is a significant difference in the mean scores of the dependent variable of each group.
- If the value of the *Sig. (2-tailed)* is above 0.05, there is no significant difference between the two groups.

6.3 One-way ANOVA (analysis of variance)

One-way ANOVA [13] is a statistical technique that can be used to evaluate whether there are statistically significant
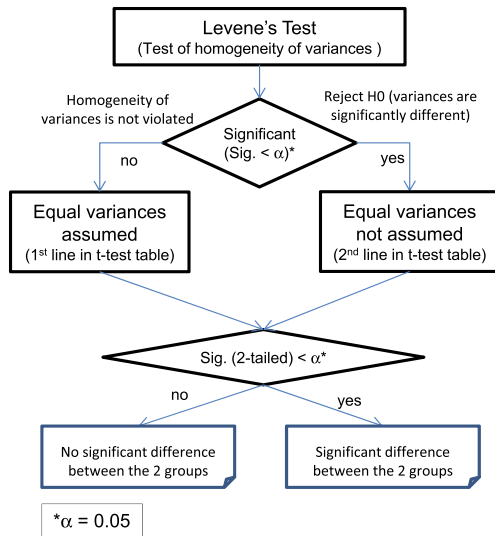
Fig. 9 Independent sample *t* test procedure

differences between the mean values across several population groups. This technique can be used only for numerical data.

More specifically, one-way ANOVA tests the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k \tag{2}$$

where μ is the group mean and k the number of groups. These variance components are then tested for statistical significance. If significant, then we reject the null hypothesis that there are no differences between means and accept the alternative hypothesis (i.e., $H_1 = \text{not } H_0$) where

the means (in the population) are different from each other [47].

The one-way ANOVA assumes the homogeneity of variances. Levene's F test [34] is used to test the assumption of homogeneity of variance. Levene's test uses the level of significance set of priori for the ANOVA (e.g., $\alpha = 0.05$) to test the assumption of homogeneity of variances. The following two sections discuss in detail the procedure in both cases, homogeneity and non-homogeneity of variances. Figure 10 illustrates the steps for each case.

6.3.1 The assumption of the homogeneity of variances is met

If the significance value (i.e., *sig.* in Levene's test [34]) is greater than 0.05 (the α level of significance), then the assumption of homogeneity of variances is met and we have to look for the ANOVA table. The ANOVA table shows the output of the one-way ANOVA and whether we have a statistically significant difference between our group means. If the significance level is greater than 0.05, then there is no statistically significant difference between the involved groups; otherwise, there is a statistically significant difference in the group means. However, the ANOVA table does not indicate which of the specific groups differed. This information can be found in the post hoc comparison table. Post hoc tests (also called multiple comparison tests or a posteriori comparisons) are tests of the statistically significant differences between group means calculated after having done ANOVA. The Tukey

Fig. 10 One-way ANOVA procedure
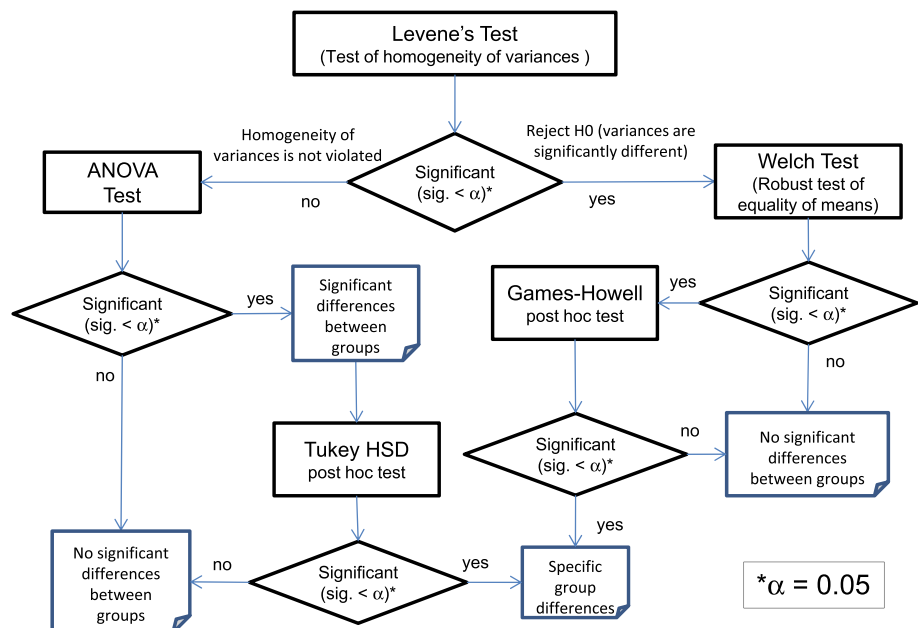
post hoc test [49] is generally the preferred test for conducting post hoc tests on a one-way ANOVA but there are many others. The Tukey's HSD (honest significant difference) test [49] is a single-step, multiple comparison procedure and statistical test. In order to apply Tukey's HSD test, the observations being tested should be independent and there should be an equal within-group variance across the groups associated with each mean in the test (i.e., homogeneity of variances).

6.3.2 The assumption of the homogeneity of variances is violated

If the equal variance assumption has been violated (e.g., if the significance *Sig.* for Levene's test is less than 0.05), we can use adjusted *F* statistics. Two types of adjustments are provided by the SPSS software: the Welch test [52] and the Brown–Forsythe test [7].

We adopt the Welch test [52] because it is more powerful and conservative than the Brown–Forsythe test [7]. If the *F* ratio (from the Welch test) is found to be significant (i.e., *Sig.* < α), the Games–Howell multi-comparison post hoc test [15] is used to find the statistically significant differences between the groups being studied. Otherwise, we can conclude that there are no differences between the involved groups.

7 Case study: fostering the university–alumni relationship

In this section, we apply our proposed approach to a GRL model (see Fig. 11) that describes how to foster the relationship between a university (King Fahd University of Petroleum and Minerals in Saudi Arabia) and its alumni. The case study involves four types of stakeholders: alumnus, professor, alumni department staff, and the university (administrators).

The realization of the goal *university–alumni relationship is fostered* and can be achieved through the satisfaction of three goals of three other actors (1) *enhanced industry–university collaboration*, (2) *serving university through alumni commitment*, and (3) *serving alumni through university commitment*.

The realization of the goal *serving the university through alumni commitment* can be achieved exclusively (with a non-negligible impact) through volunteering, mentoring current students, donating to the university, helping with student placement through internship offerings, facilitating research collaboration, and providing feedback on courses and syllabi.

The alumni department tasks (organizing networking events, providing access to the university learning facilities, providing free admissions to seminars, and providing discounts on short courses registration fees) contribute positively with different levels to the achievement of the goal *serving alumni through university commitment*. However, sharing alumni information with potential faculties and the use of SMS (short message services) for all communications, hurt the realization of that goal. These two tasks represent an invasion of alumni privacy.

The realization of the goal *enhanced industry–university collaboration* can be achieved (to some extent) through professor offering consulting services, professors contributing to industrial projects, and alumni facilitating research collaboration. In addition, by providing feedback on
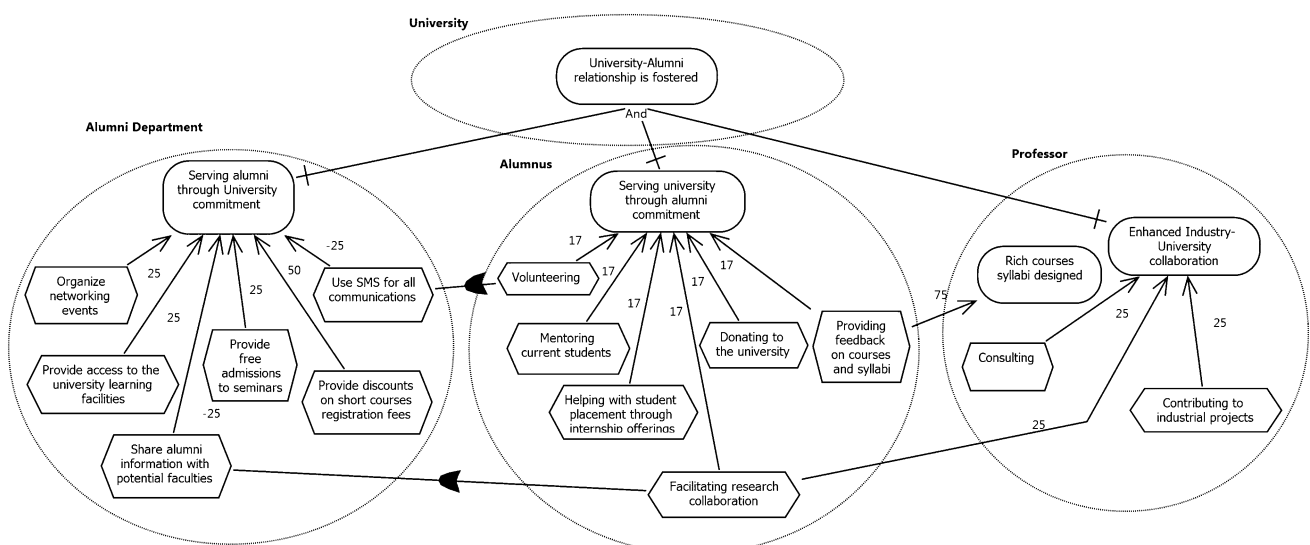

Fig. 11 GRL model describing the fostering of the university–alumni relationship

courses and syllabi, an alumnus helps the realization of the goal *rich course syllabi designed*.

The GRL model presents:

- Two explicit dependency relationships between actors *alumnus* and *alumni department*: (1) in order to volunteer, an alumnus depends on the alumni department to communicate all events through SMS, and (2) in order to facilitate research collaboration, an alumnus depends on the alumni department to share alumni information with potential faculties.
- Two implicit dependencies between actors *alumnus* and *professor*: (1) by facilitating research collaboration, an alumnus helps the realization of the goal *enhancing industry–university relationship*, and (2) by providing feedback on courses and syllabi, an alumnus helps the realization of the goal *rich course syllabi designed*.

7.1 Design of survey questions

Following the general guidelines introduced in Sect. 4, Table 5 presents the 25 questions that have been derived manually from the GRL model artifacts found in Fig. 11. The design of the questions was simple and straightforward, requiring little time and effort (about half a day). However, the questions' wording has been slightly adapted to improve the survey quality. A 7-point Likert scale was used for the answers, as suggested in Sect. 4.1.

Based on a dependency analysis of the GRL model, the designed survey has been administered to the participating stakeholders namely alumni, professors, the alumni department, and university administrators, as described in Table 5.

This application of the optimization guidelines of Sect. 5 results in smaller and more targeted surveys produced for each type of stakeholder, instead of giving the entire set of 25 questions to everyone: alumni (14 questions), professors (7 questions), the alumni department (10 questions), and university administrators (2 questions).

7.2 Survey administration and data collection

We have used Google Docs to create the survey, so we could conduct it at no cost. Customized invitations were emailed to potential participants. The survey was available to participants for five weeks. The participation rate was as follows: alumni (28 participants out of 52: 54 %), professors (20 participants out of 25: 80 %), alumni department (13 participants out of 14: 93 %), and university administrators (4 participants out of 10: 40 %). It is worth noting that most of the responses (65 %) were collected during the first week. All respondents' responses were confidentially stored using protected Google Docs spreadsheets. The

resulting spreadsheets were then imported into the SPSS [25] software in order6 to conduct data analysis.

7.3 Survey data analysis

We have used SPSS [25] to generate statistical descriptives and histograms, perform *t* test, and conduct one-way ANOVA. SPSS also generates several useful tables that will be used in the subsequent sections.

To model the four involved GRL actors, we use *group* as the dependent variable, with alumni coded as 1, professors coded as 2, alumni department coded as 3, and university administrators coded as 4.

7.3.1 Analysis of single group questions

Alumni Questions: Table 6a illustrates the statistical descriptives for the questions administered to the alumni group only (28 participants). It shows the mean, standard deviation, skewness, and kurtosis for each question. Based on the descriptives table and the analysis of the histograms in Fig. 12), we conclude that answers to questions Q1, Q2, Q4, Q5, Q6, Q11, and Q12 are positive (the mean values <3.05, skewness, and kurtosis are positive). Question Q3 has a mean of 4.11 and a negative kurtosis (meaning that the distribution is relatively flat) and skewness (meaning that a clustering is at the right side). We conclude that *the alumni do not agree on question Q3*. Hence, its corresponding contribution relationship (by donating to the university, an alumnus helps serving the university through alumni commitment) should be revised.

Professor Questions: Table 6b illustrates the statistical descriptives for the questions administered to the professor group (20 participants). Question Q15 has a negative skewness (i.e, clustering at the right side) value and a mean of 1.9. Question Q16 has a negative skewness (i.e., flat distribution) and a mean value of 1.7. Question Q17 has positive skewness and kurtosis values, and a mean of 2.1. The minimum and maximum values are between 1 and 3 for Q15 and Q16, and between 1 and 4 for Q17. A simple analysis of the frequency chart (i.e., histograms in Fig. 13 lead to the approval of the GRL artifacts relative to Q15, Q16, and Q17.

Alumni Department Questions: Table 6c illustrates the statistical descriptives for the questions administered to the alumni department group (13 participants). The ranges of answers for questions Q18, Q20, and Q22 is [1, 3], for question Q21 it is [1, 2], and for question Q19 it is [1, 4]. The mean values for the answers to Q18 to Q22 is between 1.85 and 2.23. The analysis of the frequencies (see Fig. 14) shows that the Alumni Department staff have responded positively to questions Q18, Q19, Q20, Q21, and Q22. Question Q23 has a mean value of 6 with a skewness of

Table 5 Set of questions generated for validating the GRL model of Fig. 11

Id	Questions	AQ	PQ	DQ	UQ
Q1	By *volunteering*, an alumnus helps the realization of the goal *serving the university through alumni commitment*	AQ1			
Q2	By *mentoring current students*, an alumnus helps the realization of the goal *serving the university through alumni commitment*	AQ2			
Q3	By *donating to the university*, an alumnus helps the realization of the goal *serving the university through alumni commitment*	AQ3			
Q4	By *helping with student placement through internship offerings*, an alumnus helps the realization of the goal *serving the university through alumni commitment*	AQ4			
Q5	By *providing feedback on courses and syllabi*, an alumnus helps the realization of the goal of *serving the university through alumni commitment*	AQ5			
Q6	By *facilitating research collaboration*, an alumnus helps the realization of the goal *serving the university through alumni commitment*	AQ6			
Q7	In order to perform *volunteering*, an alumnus depends on the alumni department to *communicate all events through SMS*	AQ7		DQ9	
Q8	In order to *facilitate research collaboration*, an alumnus depends on the alumni department to *share alumni information with potential faculties*	AQ8	PQ7	DQ10	
Q9	By *facilitating research collaboration*, an alumnus helps the realization of the goal *enhancing industry–university relationship*	AQ9	PQ3		
Q10	By *providing feedback on courses and syllabi*, an alumnus helps the realization of the goal *rich course syllabi designed*	AQ10	PQ4		
Q11	The following alumni tasks *volunteering*, *mentoring current students*, *donating to the university*, *helping with student placement through internship offerings*, *facilitating research collaboration*, and *providing feedback on courses and syllabi* help equally the realization of the goal *serving the university through alumni commitment*	AQ11			
Q12	The realization of the goal *serving the university through alumni commitment* can be achieved exclusively (with a non-negligible impact) through *volunteering*, *mentoring current students*, *donating to the university*, *helping with student placement through internship offerings*, *facilitating research collaboration*, and *providing feedback on courses and syllabi*	AQ12			
Q13	*Sharing alumni information with potential faculties* hurts the realization of the goal *serving alumni through university commitment*	AQ13	PQ6	DQ5	
Q14	The execution of *use SMS for all communications* hurts the realization of the goal *serving alumni through university commitment*	AQ14		DQ6	
Q15	Through *consulting*, a professor helps the realization of the goal *enhanced industry–university collaboration*		PQ1		
Q16	By *contributing to industrial projects*, a professor helps the realization of the goal *enhanced industry–university collaboration*		PQ2		
Q17	The realization of the goal *enhanced industry–university collaboration* can be achieved exclusively (with a non-negligible impact) through *professor offering consulting services*, *professors contributing to industrial projects*, and *alumni facilitating research collaboration*		PQ5		
Q18	By *providing access to the university learning facilities*, the alumni department helps the realization of the goal *serving alumni through university commitment*			DQ1	
Q19	By *providing free admissions to seminars*, the alumni department helps the realization of the goal *serving alumni through university commitment*			DQ2	
Q20	By *providing discounts on short courses registration fees*, the alumni department helps the realization of the goal *serving alumni through university commitment*			DQ3	
Q21	By *organizing networking events*, the alumni department helps the realization of the goal *serving alumni through university commitment*			DQ4	
Q22	The realization of the goal *serving alumni through university commitment* can be achieved exclusively (with a non-negligible impact) through *organizing networking events*, *providing access to the university learning facilities*, *providing free admissions to seminars*, and *providing discounts on short courses registration fees*			DQ7	
Q23	*Sharing alumni information with potential faculties* and *using SMS for all communications* have the same non-negligible negative impact on the realization of the goal *serving alumni through university commitment*			DQ8	

Table 5 continued

Id	Questions	AQ	PQ	DQ	UQ
Q24	The satisfaction of each of the following goals: (1) *serving university through alumni commitment* (2) *serving alumni through university commitment* (3) *enhanced industry–university collaboration* is necessary to achieve the goal *university–alumni relationship is fostered*				UQ1
Q25	The satisfaction of each of the following goals: (1) *serving university through alumni commitment* (2) *serving alumni through university commitment* (3) *enhanced industry–university collaboration* is sufficient to achieve the goal *university–alumni relationship is fostered*				UQ2

Italics refer to the model intentional elements

AQ Alumnus Question, *PQ* Professor Question, *DQ* Department Question, *UQ* University Question

Table 6 Descriptive statistics for the questions involving one group

	N	Mean	SD	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	SE	Statistic	SE
(a) Descriptive statistics for questions Q1–Q6, Q11, and Q12							
Q1	28	2.57	1.372	.765	.441	.078	.858
Q2	28	2.39	1.571	1.015	.441	.076	.858
Q3	28	4.11	2.283	−.202	.441	−1.481	.858
Q4	28	2.29	1.630	1.326	.441	.769	.858
Q5	28	1.86	1.145	1.254	.441	.791	.858
Q6	28	1.82	.819	.789	.441	.263	.858
Q11	28	2.64	1.193	.903	.441	1.182	.858
Q12	28	3.00	1.122	.508	.441	.779	.858
(b) Descriptive statistics for questions Q15–Q17							
Q15	20	1.90	.553	−.083	.512	.766	.992
Q16	20	1.70	.657	.396	.512	−.547	.992
Q17	20	2.10	.788	.531	.512	.490	.992
(c) Descriptive statistics for questions Q18–Q23							
Q18	13	2.00	.577	.000	.616	1.036	1.191
Q19	13	2.23	.725	1.156	.616	2.469	1.191
Q20	13	1.85	.689	.203	.616	−.496	1.191
Q21	13	1.69	.480	−.946	.616	−1.339	1.191
Q22	13	2.23	.599	−.065	.616	.051	1.191
Q23	13	6.00	.577	.000	.616	1.036	1.191
(d) Descriptive statistics for questions Q24 and Q25							
Q24	4	1.50	.577	.000	1.014	−6.000	2.619
Q25	4	2.25	.500	2.000	1.014	4.000	2.619

0.000 (symmetric distribution), a range between 5 and 7. We can conclude that *participants disagree on question Q23*.

University Administrators Questions: Table 6d shows mean values of 1.5 and 2.25 for Q24 and Q25, respectively. The answers range for Q24 is [1, 2], while the answers range for Q25 is [2, 3]. We can conclude that university administrators (4 participants) agree on both questions (Fig. 15).

7.3.2 Analysis of the questions involving two groups

In this section, we apply independent sample t test to the data collected from questions administered to two stakeholders.

Table 7a provides the mean and the standard deviations of alumnus and professor groups with respect to questions Q9 and Q10. Levene's test (see Table 7b) shows that the variances for the two groups (alumnus/professor) are not equal for both questions Q9 and Q10 (*Sig.* is 0.681 and 0.092, respectively). Hence, we have to refer to the value of the column labeled *Sig.(2-tailed)* on the second line of each question row. Based on the values of significance, we can conclude that there is no difference between both groups with respect to question Q9 (*Sig.* = 0.705 > 0.05), while there is a statistically significant difference with respect to question Q10 (*Sig.* = 0.000 < 0.05). We conclude that there is a conflict with respect to question Q10, between the

Table 7 Group statistics and independent sample *t* test for the questions involving two groups

	Group	N	Mean	Std. Deviation	Std. Error Mean
Q9	1	28	1.89	.916	.173
	2	20	1.80	.768	.172
Q10	1	28	2.14	1.239	.234
	2	20	5.35	1.843	.412

(a) Group statistics for questions Q9 and Q10

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
Q9	Equal variances assumed	.171	.681	.370	46	.713
	Equal variances not assumed			.381	44.742	.705
Q10	Equal variances assumed	2.956	.092	-7.217	46	.000
	Equal variances not assumed			-6.766	30.968	.000

(b) Independent samples t-test for Q9 and Q10

	Group	N	Mean	Std. Deviation	Std. Error Mean
Q7	1	28	2.96	1.815	.343
	3	13	2.00	.816	.226
Q14	1	28	4.57	1.989	.376
	3	13	6.08	.494	.137

(c) Group statistics for questions Q7 and Q14

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
Q7	Equal variances assumed	14.011	.001	1.822	39	.076
	Equal variances not assumed			2.346	38.997	.024
Q14	Equal variances assumed	33.331	.000	-2.674	39	.011
	Equal variances not assumed			-3.763	33.316	.001

(d) Independent samples t-test for Q7 and Q14

alumnus group (agrees with a mean of 2.14) and the professor group (disagrees with a mean of 5.35), that should be addressed.

Table 7c provides the means and the standard deviations of the alumnus and alumni department groups with respect to questions Q7 and Q14. Levene's test (see Table 7d) shows that the variances for the two groups (alumnus/alumni department) are equal for both questions Q7 and Q14 (*Sig.* is 0.001 and 0.000, respectively). Hence, we have to refer to the value of the column labeled *Sig.(2-tailed)* at the first line of each question row. Based on the values of significance, we can conclude that there is no difference between both groups with respect to question Q7 (*Sig.* = 0.076 > 0.05), while there is a statistically significant difference with respect to question Q14

(*Sig.* = 0.011 < 0.05). The alumni group is neutral with respect to Q14 (i.e., mean = 2.14), while the alumni department group disagree (i.e., mean = 6.08). We conclude that there is a slight conflict with respect to question Q14, between the alumnus group (kind of neutral with a mean of 4.57) and the alumni department group (disagreeing with a mean of 6.08), that should be addressed.

7.3.3 Analysis of questions involving three groups

Table 8a shows the group means and the standard deviation for each group for both questions Q8 and Q13. For question Q8, the *F* value for Levene's test (see Table 8b) is 1.381 with a *Sig.* value of 0.26. Because the *Sig.* value is greater than our alpha of 0.05, the assumption of

homogeneity of variances is not violated. Hence, we can proceed with ANOVA.

The ANOVA (see Table 8c) produces a *Sig.* value of 0.291, which is greater than 0.05. We can conclude that there is no differences between the three groups with respect to question Q8.

For question Q13, the *F* value for Levene's test (see Table 8b) is 5.171 with a *Sig.* value of 0.009. Because the *Sig.* value is less than our alpha of 0.05, we reject the null hypothesis for the homogeneity of variances and we proceed with the Welch test (see Table 8d). The *F* ratio is found to be significant with the Welch test since *Sig.* is 0.000 (less than 0.05). We reject the null hypothesis and proceed with the Games–Howell test [15] for post hoc comparison. The multiple comparisons table (see Table 8e) shows significant differences between the alumnus group and the professor/alumni department groups (with *Sig.* values of 0.002 and 0.000, respectively). No difference is found between the professor group and the alumni department group since *Sig.* is 0.081 > 0.05. We conclude that there is a conflict between the alumnus group (kind of neutral with a mean of 4.36) and the professor/alumni department groups (disagreeing with means of 5.75 and 6.46 respectively) that should be addressed.

For this case study, our validation approach has detected issues and conflicts with respect to several parts of the GRL model related to questions, which represent 20 % of the questions. These parts of the model hence require conflict resolution, and possibly re-validation, as suggested in our process (Fig. 2). Q3 is a positive contribution that can be reassessed by the alumnus group, whereas Q23 requires re-balancing the two negative contributions by the alumni department staff. The resolution of issues related to the positive and negative contributions covered by Q10, Q13, and Q14 requires the participation of two or three categories of stakeholders.

8 Discussion

8.1 General benefits of the approach

One important objective of using goal models in software engineering and in other domains is to enable stakeholders to *disagree sooner*, when modifications are not costly. It is also often easier to disagree about a compact goal model than to try to do the same on textual documents, whose length and stylistic or grammatical issues are often distracting.

The validation approach proposed here brings the following benefits, which allow people to agree and

Table 8 Analysis of the data collected from the questions involving three groups

		N	Mean	Std. Deviation	Std. Error
Q8	1	28	2.11	.956	.181
	2	20	2.30	1.418	.317
	3	13	1.69	.630	.175
	Total	61	2.08	1.085	.139
Q13	1	28	4.36	1.521	.287
	2	20	5.75	1.164	.260
	3	13	6.46	.660	.183
	Total	61	5.26	1.526	.195

(a) Descriptives for Q8 and Q13

	Levene Statistic	df1	df2	Sig.
Q8	1.381	2	58	.260
Q13	5.171	2	58	.009

(b) Levene's test for Q8 and Q13

Q8

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.942	2	1.471	1.261	.291
Within Groups	67.648	58	1.166		
Total	70.590	60			

(c) ANOVA table for Q8

Q13

	Statistic[a]	df1	df2	Sig.
Welch	18.798	2	37.869	.000

a. Asymptotically F distributed.

(d) Welch test for Q13

Q13
Games-Howell

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.
1	2	-1.393[*]	.388	.002
	3	-2.104[*]	.341	.000
2	1	1.393[*]	.388	.002
	3	-.712	.318	.081
3	1	2.104[*]	.341	.000
	2	.712	.318	.081

*. The mean difference is significant at the 0.05 level.

(e) Posthoc Games-Howell table for Q13

* The Mean difference is significant at the 0.05 level

[a] Asymptotically *F* distributed

especially to disagree sooner, in a more systematic way, at little cost:

1. Our approach offers a systematic way to cover the elements and relationships of a goal model through natural language questions readable by non-experts.
2. As our validation approach is survey based, it enables validation by a large number of stakeholders, who can be distributed geographically.

3. The survey administration can be done asynchronously: Not everyone needs to do it at the same time, and longer surveys do not have to be filled all at once either.

4. The data collection procedure avoids peer pressure since the survey is anonymous.

5. The number of questions is minimized by focusing on questions relevant to the context of each actor (internal elements and relationships, as well as immediate interactions with other actors).

6. Tables 1 and 2 provide vocabulary for all GRL constructs. We believe the generation of survey questions to be partially automatable in that context.

7. The generation of survey questions itself allows for reasoning about the model artifacts and provides insight into the model design rationale. An analyst is likely to detect issues in the model while deriving and inspecting the survey questions.

8.2 Threats to validity

Our approach and the case study we performed are subject to several limitations and threats to validity, categorized here according to three important types of threats identified by Wright et al. [53].

In terms of *construct validity*, there is some criticism with respect to the use of various parametric methods such as *t* test, ANOVA, and regression with (1) small sample sizes and (2) data that might not be normally distributed. However, Norman [41] has shown that parametric statistics are robust with respect to violations of these assumptions. In addition, GRL models for software applications often include an actor capturing the system and its intended functionalities (and how the latter contribute to fulfilling stakeholder objectives). Such system actor was not part of our case study, and one would need to determine who would be responsible to answer its associated questionnaire (e.g., developers). Note also that our case study did not focus on the resolution of the conflicts and other issues discovered through validation; whether there is a rapid convergence toward a resolution is outside the scope of this paper.

Regarding *internal validity*, there is a risk that participants do not answer seriously and meaningfully the survey questions. This could be detected in the future through partially redundant questions, at the cost of longer surveys. In our case study, there was no attempt to randomize participation (as this is a real situation, with real stakeholders) or train them properly on goal modeling and on GRL specifically. This might have introduced some bias, which might have influenced participant answers. Another threat may be related to the difficulty to enroll participants and its incurred cost. On the other hand, validation cannot happen without the participation of stakeholders.

As for *external validity*, the approach is currently tailored to GRL. Although GRL has many constructs that are common with other goal modeling languages, the approach and guidelines might not be generalizable to other such languages without substantial adaptation. In addition, feasibility was demonstrated only on one realistic case study (this paper), and on a simpler GRL model (with no actor) describing the introduction of a new security elective course in a university curriculum [18]. Generalization to other domains and to larger models remains to be investigated. Also, although these case studies offer a good coverage of GRL elements and relationships, some constructs have not been used yet (e.g., indicators and URN links). Future case studies should ensure a good coverage of such constructs.

9 Related work

The growing popularity of goal-oriented modeling, and its adoption by a large international community, led to the development of many goal-oriented analysis methodologies [3, 5, 9, 16, 21, 29]. The latter differ in their targeted notation and in their purpose. However, it is worth noting that most of these methodologies focus on the qualitative or/and quantitative evaluation of satisfaction levels of the goals and actors composing the model given some initial satisfaction levels [3, 9, 16, 22, 23].

Based on the *i** framework, Horkoff et al. [23] have developed an interactive (semi-automated), forward propagation algorithm with qualitative values. They have also explored an interactive backward propagation algorithm with quantitative values [22]. Amyot et al. [3] have proposed three algorithms (qualitative, quantitative, and hybrid) to automatically propagate satisfaction levels in a GRL model. Initial satisfaction levels for some of the intentional elements are provided in a strategy and then propagated, using a forward propagation mechanism, to the other intentional elements of the model through the various graph links. Giorgini et al. [16] have used an axiomatization approach to formalize goal models in TROPOS using four qualitative contribution levels ($-$, $--$, $+$, $++$). The authors have provided forward and backward propagation algorithms to detect three types of conflicts (weak, medium, and strong). However, there are several research contributions that target the correct syntactical construction of goal models, with the measurable satisfaction of generic properties. Several tools exist that can check the syntax and well-formedness of models in most goal-oriented modeling languages [2, 4].

There also exist techniques that aim to detect complexity or incompleteness issues by computing metrics on goal models, for example, in *i** [14] and in KAOS [11]. A

few approaches exploit discussions and argumentation theory for validating the consistency of vocabulary and ontologies used in goal models [29, 30, 37]. Jureta et al. [29] have proposed a question-based Goal Argumentation Method (GAM) to help clarify and detect any deficient argumentation within goal models. However, their approach considers neither survey administration nor statistical analysis. To the best of our knowledge, no empirical approach has been proposed to establish the validity of goal models with stakeholders.

There are even approaches that attempt to assess the usability and cognitive effectiveness of entire goal-oriented modeling languages, as in Moody's Physics of Notations applied to i* [38].

In his early work, Robinson [42] proposed an approach allowing for the identification and characterization of conflicts as differences at the goal level. The identified conflicts can then be resolved through negotiation [43].

In terms of participatory construction of goal models, there is recent work that aims to support stakeholders in reaching agreement on *contributions* in goal models. In particular, the analytical hierarchy process (AHP) technique [44] is used to take into consideration the opinions of many stakeholders, through surveys based on pairwise comparisons. Relative contribution weights have hence been computed with AHP for i* by Liaskos et al. [35], for the NFR framework by Kassab [31], for GRL by Akhigbe et al. [1], and for a proprietary goal modeling language by Vinay et al. [51]. Although and AHP-based approach is useful when constructing goal models, it does not prevent the need for validation. In addition, such approaches are quite limited in terms of the number of constructs they cover; they essentially target relative contributions (including correlations) where weights must be positive. Moreover, they neither cover negative contributions nor collections of contributions targeting a same intentional element and whose weights sum up to a value different from 100 % (i.e., over-contributions and under-contributions are not supported).

Our validation approach is original in its use of empirical data through surveys and of statistical analysis in the context of goal model validation, and it is complementary to the approaches discussed above.

10 Conclusions and future work

In this paper, we have proposed a novel participation-oriented validation approach for GRL models based on empirical data collection and analysis. We have proposed a procedure and vocabulary to derive survey questions from GRL models. To tackle the scalability issue, we have proposed optimization guidelines to allow for the assessment of sub-models instead of isolated GRL constructs. In addition, we have discussed the impact of conducting a dependency analysis of the intervening stakeholders on the reduction of the number of questions administered to each stakeholder group.

The collected data are then analyzed using proven statistical methods such as *t* test and *ANOVA (analysis of variance)* in order to detect conflicts between stakeholders, hence allowing them to "disagree sooner." Furthermore, our approach could guide the argumentation and justification of modeling choices during the construction of goal models.

We have illustrated and evaluated our approach through a realistic case study, leading to four types of questionnaires answered by a total of 65 participants. The results were discussed, the uniqueness of the approach was argued through a comparison with related work, and limitations and threats to the validity of this work were identified in the previous two chapters.

As part of our future work, we plan to develop our validation approach further to go beyond conflict detection to conflict *resolution*, and to study resolution convergence issues, if any. Ontologies could also be explored to support questionnaire-based validation and conflict detection rather than relying solely on the grammatical constructs of a language [12, 29]. The partial automation of the generation of questionnaires (with real or skeleton questions) from goal models themselves is also an interesting and critical challenge that can accelerate the usability and efficiency of the approach. Further empirical evidence should be collected on new case studies and usability experiments, which should also help cover constructs missing from our existing case studies. There is, finally, an opportunity to empirically evaluate the generation of questionnaires itself by assessing the validity of various alternatives to the currently proposed vocabulary.

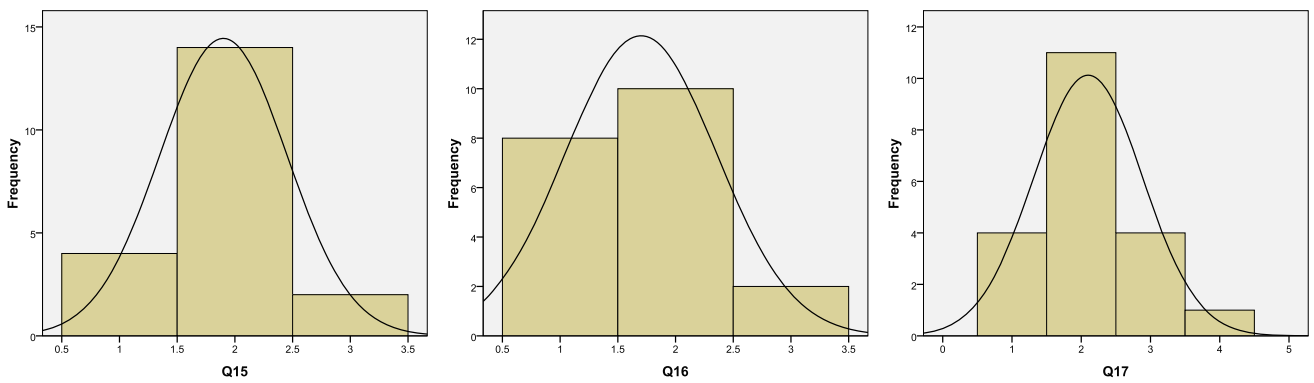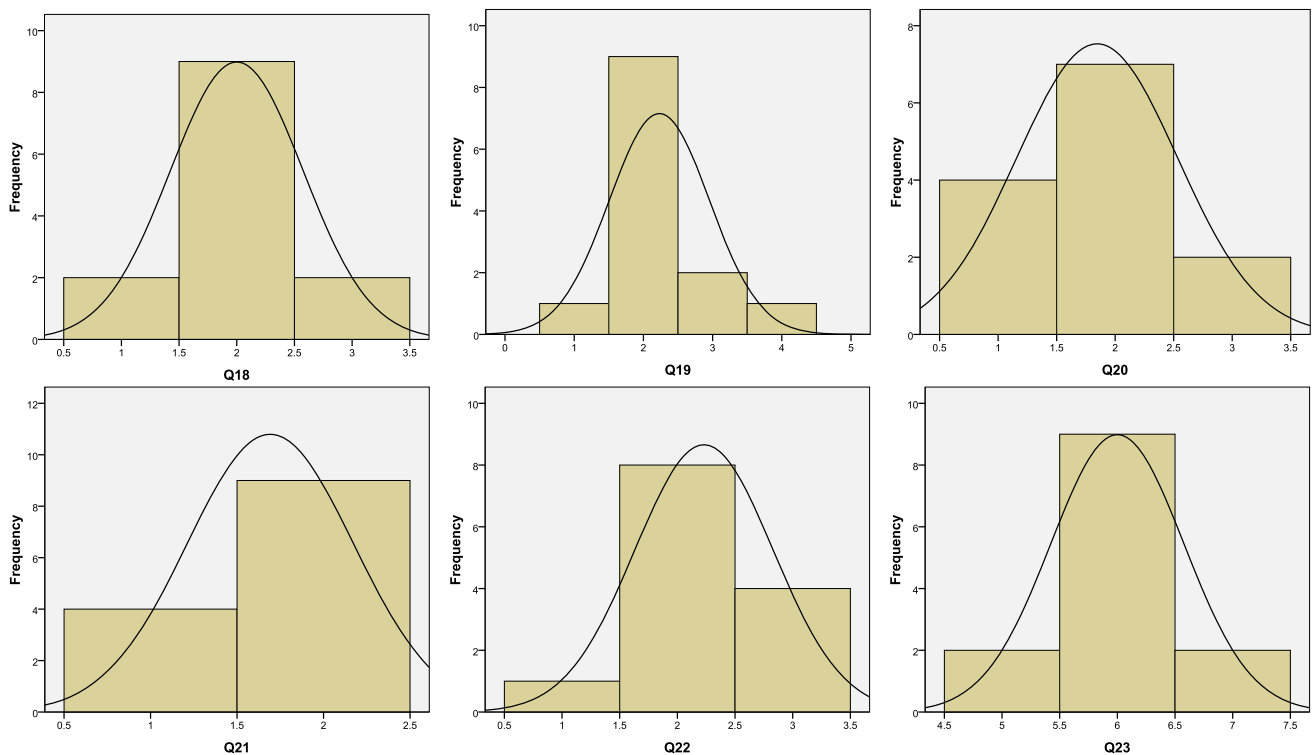## Appendix: SPSS generated histograms
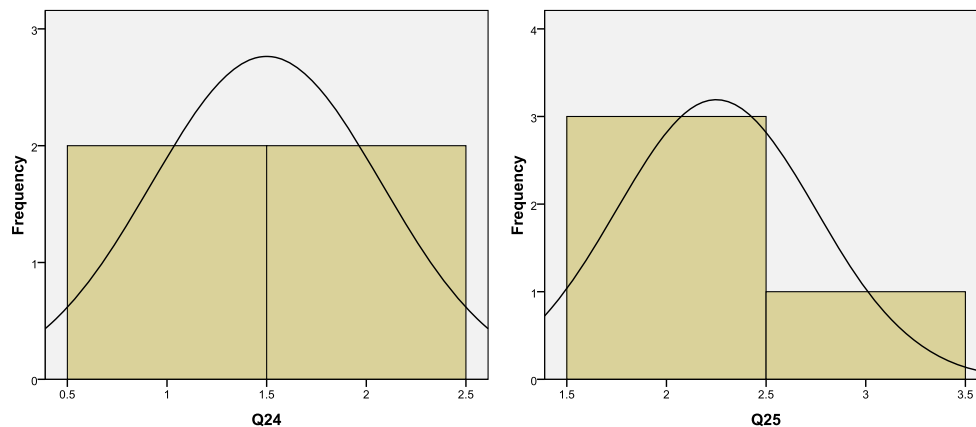
See Figs. 12, 13, 14 and 15.
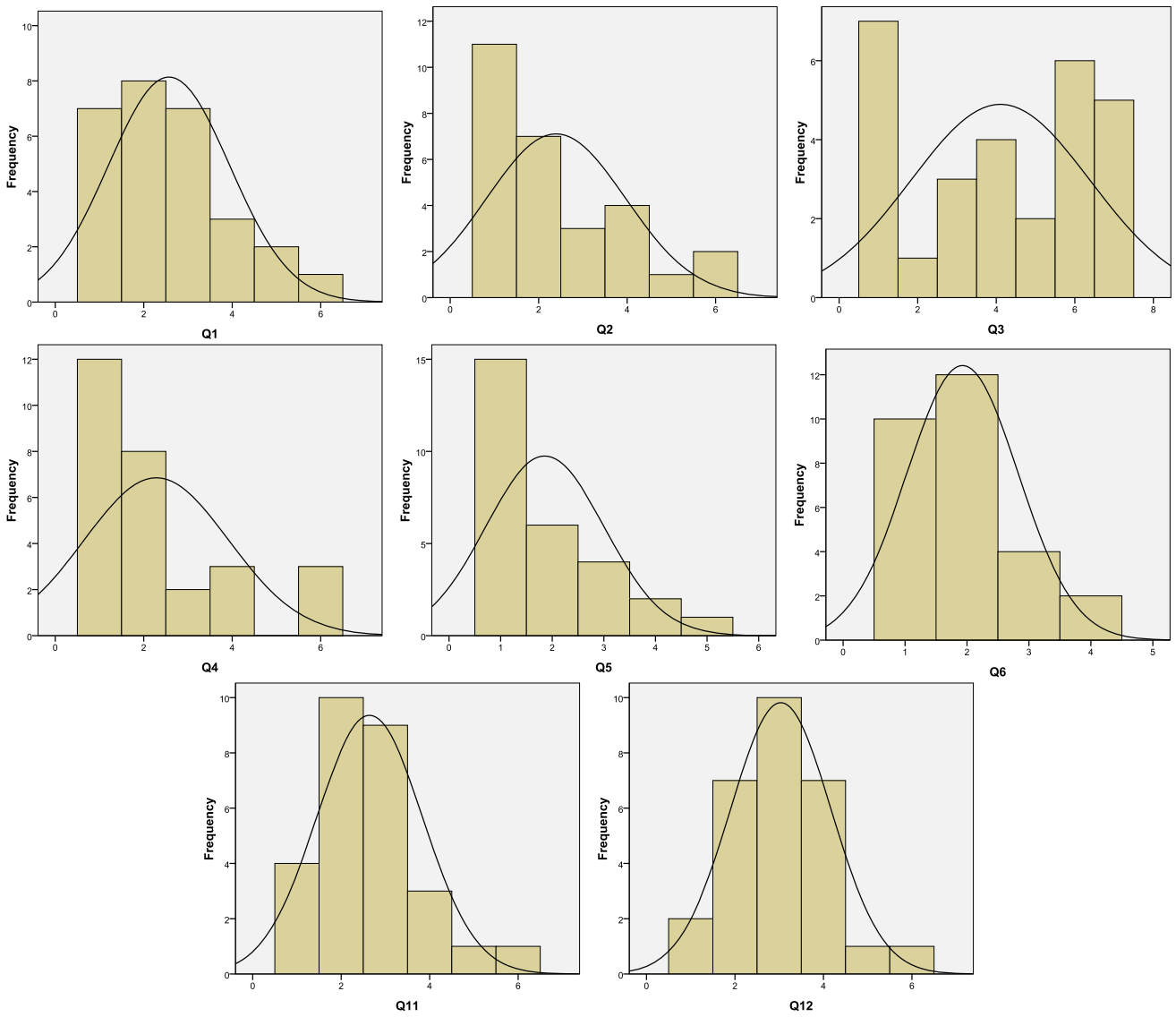
Fig. 12 Histograms for questions Q1 to Q6, Q11, and Q12 (alumnus), on a 1-to-7 Likert scale

Fig. 13 Histograms for questions Q15 to Q17 (professor), on a 1-to-7 Likert scale

Fig. 14 Histograms for questions Q18 to Q23 (alumni department), on a 1-to-7 Likert scale

Fig. 15 Histograms for questions Q24 and Q25 (university), on a 1-to-7 Likert scale

References

1. Akhigbe O, Alhaj M, Amyot D, Badreddin O, Braun E, Cartwright N, Richards G, Mussbacher G (2014) Creating quantitative goal models: governmental experience. In: 33rd international conference on conceptual modeling (ER'14), lecture notes in computer science, vol 8824. Springer, Berlin, pp 466–473

2. Almeida C, Goulão M, Araújo J A systematic comparison of i* modelling tools based on syntactic and well-formedness rules. In: Castro et al. (eds) Proceedings of the 6th international i* workshop 2013, Valencia, Spain, June 17–18, 2013, CEUR workshop proceedings, vol 978. CEUR-WS.org, pp 43–48

3. Amyot D, Ghanavati S, Horkoff J, Mussbacher G, Peyton L, Yu E (2010) Evaluating goal models within the goal-oriented requirement language. Int J Intell Syst 25:841–877. doi:10.1002/int.v25:8

4. Amyot D, Horkoff J, Gross D, Mussbacher G (2009) A lightweight GRL profile for i* modeling. In: Proceedings of the ER 2009 workshops (CoMoL, ETheCoM, FP-UML, MOST-ONISW, QoIS, RIGiM, SeCoGIS) on advances in conceptual modeling—challenging perspectives, ER '09, pp 254–264. Springer, Berlin. doi:10.1007/978-3-642-04947-7_31

5. Amyot D, Rashidi-Tabrizi R, Mussbacher G, Kealey J, Tremblay E, Horkoff J (2013) Improved GRL modeling and analysis with jUCMNav 5. In: Castro et al. (eds) Proceedings of the 6th

international *i** workshop 2013, Valencia, Spain, June 17–18, 2013, CEUR workshop proceedings, vol 978. CEUR-WS.org, pp 137–139

6. Ayala CP, Cares C, Carvallo JP, Grau G, Haya M, Salazar G, Franch X, Mayol E, Quer C (2005) A comparative analysis of i*-based agent-oriented modeling languages. In: Proceedings of the 17th international conference on software engineering and knowledge engineering (SEKE'2005), Taipei, Taiwan, Republic of China, July 14–16, pp 43–50

7. Brown MB, Forsythe AB (1974) Robust tests for the equality of variances. J Am Stat Assoc 69(346):364–367. doi:10.1080/01621459.1974.10482955

8. Castro J, Horkoff J, Maiden NAM, Yu ESK (eds) (2013) Proceedings of the 6th international *i** workshop 2013, Valencia, Spain, June 17–18, 2013, CEUR workshop proceedings, vol 978. CEUR-WS.org

9. Chung L, Nixon BA, Yu E, Mylopoulos J (1999) Non-functional requirements in software engineering. The Kluwer international series in software engineering. Kluwer Academic Publishers Group, Dordrecht

10. de Castro JB, Franch X, Mylopoulos J, Yu ESK (eds) (2011) Proceedings of the 5th international *i** workshop 2011, Trento, Italy, August 28–29, 2011, CEUR workshop proceedings, vol 766. CEUR-WS.org

11. Espada P, Goulão M, Araújo J (2013) A framework to evaluate complexity and completeness of KAOS goal models. In: Salinesi C, Norrie MC, Pastor O (eds) CAiSE. Lecture notes in computer science, vol 7908. Springer, Berlin, pp 562–577

12. Fernandes PCB, Guizzardi RSS, Guizzardi G (2011) Using goal modeling to capture competency questions in ontology-based systems. JIDM 2(3):527–540

13. Fisher R (1925) Statistical methods for research workers. Cosmo study guides. Cosmo Publications, New Delhi

14. Franch X (2009) A method for the definition of metrics over i* models. In: van Eck P, Gordijn J, Wieringa R (eds) CAiSE. Lecture notes in computer science, vol 5565. Springer, Berlin, pp 201–215

15. Games PA, Howell JF (1976) Pairwise multiple comparison procedures with unequal n's and/or variances: a monte carlo study. J Educ Behav Stat 1(2):113–125

16. Giorgini P, Mylopoulos J, Sebastiani R (2005) Goal-oriented requirements analysis and reasoning in the tropos methodology. Eng Appl Artif Intell 18:159–171. doi:10.1016/j.engappai.2004.11.017

17. Gosset WS (1908) The probable error of a mean. Biometrika 6(1):1–25 (Originally published under the pseudonym "Student")

18. Hassine J, Amyot D (2013) GRL model validation: a statistical approach. In: Haugen Ø, Reed R, Gotzhein R (eds) System analysis and modeling: theory and practice. Lecture notes in computer science, vol 7744. Springer, Berlin, pp 212–228. doi:10.1007/978-3-642-36757-1_13

19. Hesse-Biber SN, Leavy P (2006) The practice of qualitative research. SAGE, Beverley Hills

20. Horkoff J, Barone D, Jiang L, Yu E, Amyot D, Borgida A, Mylopoulos J (2013) Strategic business modeling—representation and reasoning. Softw Syst Model 13(3):1015–1041

21. Horkoff J, Yu ESK (2013) Comparison and evaluation of goal-oriented satisfaction analysis techniques. Requir Eng 18(3):199–222

22. Horkoff J, Yu ESK (2008) Qualitative, interactive, backward analysis of i* models. In: de Castro JB, Franch X, Perini A, Yu ESK (eds) Proceedings of the 3rd international i* workshop (iStar), Recife, Brazil, CEUR workshop proceedings, vol 322. CEUR-WS.org, pp 43–46

23. Horkoff J, Yu E, Liu L (2006) Analyzing trust in technology strategies. In: Proceedings of the 2006 international conference on privacy, security and trust: bridge the gap between PST technologies and business services, PST '06. ACM, New York, pp 9:1–9:12. doi:10.1145/1501434.1501446

24. Iarossi G (2006) The power of survey design: a user's guide for managing surveys, interpreting results, and influencing respondents. Stand Alone Series. World Bank http://books.google.tn/books?id=x964AAAAIAAJ

25. IBM (2012) SPSS software http://www-01.ibm.com/software/analytics/spss/

26. ITU-T (2012) Recommendation Z.151 (10/12) user requirements notation (URN) language definition, Geneva, Switzerland. http://www.itu.int/rec/T-REC-Z.151/en

27. Jamieson S (2004) Likert scales: how to (ab)use them. Med Educ 38(12):1217–1218. doi:10.1111/j.1365-2929.2004.02012.x

28. jUCMNav (2014) jUCMNav Project, v6.0.0 (tool, documentation, and meta-model). http://softwareengineering.ca/jucmnav

29. Jureta IJ, Faulkner S, Schobbens PY (2008) Clear justification of modeling decisions for goal-oriented requirements engineering. Requir Eng 13:87–115. doi:10.1007/s00766-007-0056-y

30. Jureta I, Mylopoulos J, Faulkner S (2009) Analysis of multi-party agreement in requirements validation. In: RE. IEEE Computer Society, pp 57–66

31. Kassab M (2013) An integrated approach of AHP and NFRs framework. In: Wieringa R, Nurcan S, Rolland C, Cavarero JL (eds) RCIS, pp 1–8. IEEE

32. Knapp TR (1990) Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nurs Res 39(2):121–123

33. Labovitz S (1967) Some observations on measurement and statistics. Social Forces 46(2):151–160. doi:10.2307/2574595

34. Levene H (1960) Robust tests for equality of variances. In: Olkin I (ed) Contributions to probability and statistics: essays in honor of Harold Hotelling. Stanford University Press, Palo Alto, pp 278–292

35. Liaskos S, Jalman R, Aranda J (2012) On eliciting contribution measures in goal models. In: Heimdahl MPE, Sawyer P (eds) RE. IEEE, pp 221–230

36. Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 140(140):1–55

37. Mirbel I, Villata S (2012) Enhancing goal-based requirements consistency: an argumentation-based approach. In: Fisher M, van der Torre L, Dastani M, Governatori G (eds) CLIMA. Lecture notes in computer science, vol 7486. Springer, Berlin, pp 110–127

38. Moody DL, Heymans P, Raimundas Matulevičius R (2010) Visual syntax does matter: improving the cognitive effectiveness of the i* visual notation. Requir Eng 15(2):141–175. doi:10.1007/s00766-010-0100-1

39. Munro S, Liaskos S, Aranda J (2011) The mysteries of goal decomposition. In: de Castro et al. (eds) Proceedings of the 5th international *i** workshop 2011, Trento, Italy, August 28–29, 2011, CEUR workshop proceedings, vol 766. CEUR-WS.org, pp. 49–54

40. Mussbacher G, Amyot D, Heymans P (2011) Eight deadly sins of GRL. In: de Castro et al. (eds) Proceedings of the 5th international *i** workshop 2011, Trento, Italy, August 28–29, 2011, CEUR workshop proceedings, vol 766. CEUR-WS.org, pp 2–7

41. Norman G (2010) Likert scales, levels of measurement and the "laws" of statistics. Adv Health Sci Educ 15(5):625–632. doi:10.1007/s10459-010-9222-y

42. Robinson WN (1989) Integrating multiple specifications using domain goals. In: Proceedings of the 5th international workshop on software specification and design, IWSSD '89. ACM, New York, pp 219–226. doi:10.1145/75199.75232

43. Robinson WN (1990) Negotiation behavior during requirements specification. In: Proceedings of the 12th international conference on software engineering, ICSE '90. IEEE Computer Society Press, Los Alamitos, pp. 268–276.http://dl.acm.org/citation.cfm?id=100296.100335

44. Saaty TL (1990) How to make a decision: the analytic hierarchy process. Eur J Oper Res 48(1):9–26. doi:10.1016/0377-2217(90)90057-I

45. Schuman H, Presser S (1981) Questions and answers in attitude surveys: experiments on question form, wording, and context. Academic Press, London

46. Shamsaei A, Amyot D, Pourshahid A, Yu E, Mussbacher G, Tawhid R, Braun E, Cartwright N (2013) An approach to specify and analyze goal model families. In: Haugen Ø, Reed R, Gotzhein R (eds) System analysis and modeling: theory and practice. Lecture notes in computer science, vol 7744. Springer, Berlin, pp 34–52. doi:10.1007/978-3-642-36757-1_13

47. StatSoft (2012) Electronic statistics textbook. Tulsa, OK: Statsoft http://www.statsoft.com/textbook/

48. Tabachnick BG, Fidell LS (2006) Using multivariate statistics, 5th edn. Allyn & Bacon Inc, Needham Heights

49. Tukey J (1977) Exploratory data analysis. Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company, Reading

50. van Lamsweerde A (2008) Requirements engineering: from craft to discipline. In: Harrold MJ, Murphy GC (eds) Proceedings of the 16th ACM SIGSOFT international symposium on foundations of software engineering (FSE 2008), Atlanta, GA. ACM, New York, pp 238–249

51. Vinay S, Aithal S, Sudhakara G (2012) A quantitative approach using goal-oriented requirements engineering methodology and analytic hierarchy process in selecting the best alternative. In: Kumar AM, Kumar TVS (eds) Proceedings of international conference on advances in computing, advances in intelligent systems and computing, vol 174. Springer, India, pp. 441–454. doi:10.1007/978-81-322-0740-5_54

52. Welch BL (1947) The generalization of 'student's' problem when several different population variances are involved. Biometrika 34(1–2):28–35. doi:10.1093/biomet/34.1-2.28

53. Wright HK, Kim M, Perry DE (2010) Validity concerns in software engineering research. In: Roman GC, Sullivan KJ (eds) FoSER. ACM, New York, pp 411–414

54. Yu ESK (1997) Towards modeling and reasoning support for early-phase requirements engineering. In: Proceedings of the 3rd IEEE international symposium on requirements engineering., RE '97IEEE computer society, Washington, DC, pp 226–235