ADVANCING TASK ELICITATION SYSTEMS – AN EXPERIMENTAL EVALUATION OF DESIGN PRINCIPLES

Completed Research Paper

Hendrik Meth
University of Mannheim
L 15, 1-6, Mannheim, Germany
meth@es.uni-mannheim.de

Ye Li
University of Mannheim
L 15, 1-6, Mannheim, Germany
yeli@es.uni-mannheim.de

Alexander Maedche
University of Mannheim
L 15, 1-6, Mannheim, Germany
maedche@es.uni-mannheim.de

Benjamin Mueller
University of Mannheim
L 15, 1-6, Mannheim, Germany
mueller@es.uni-mannheim.de

Abstract

In large IS development projects a huge number of natural language documents becomes available and needs to be analyzed and transformed into structured requirements. This elicitation process is known to be time-consuming and error-prone when performed manually by a requirements engineer. Thus, there is a clear demand for advanced support of the entire elicitation process. Our work focuses on providing automated and knowledge-based support of the task elicitation sub-process. Following a design science approach, design principles for task elicitation systems are conceptualized and instantiated in an artifact. We evaluate our design principles in a laboratory experiment and examine its external validity in a field setting. We contribute to the body of knowledge by explaining effects of the conceptualized and instantiated design principles. Specifically, our results show that the level of automation as well as the extent and origin of the knowledge used for the automation process affect task elicitation productivity.

Keywords: Requirements elicitation, task elicitation, natural language processing, design science, experiment, productivity

Introduction

The success of IS development highly depends on the accuracy of the requirements gathered from users and other stakeholders during the requirements elicitation process (Appan and Browne 2012; Hickey and Davis 2004). Approximately 80% of software requirements are recorded in natural language (Mich et al. 2004; Neill and Laplante 2003), in documents like interview transcripts, workshop memos, or narrative scenarios. Natural language is inherently powerful and expressive and can therefore be used to communicate between a broad range of stakeholders and users (Alkhader et al. 2006). Even though it appears to be a well suited mean to articulate and discuss requirements, severe problems emerge when using natural language in specification documents as they might be ambiguous, inconsistent and incomplete (Wilson et al. 1997). Furthermore, a direct interpretation of these documents by subsequent development tools is almost impossible (Alkhader et al. 2006). Therefore natural language requirements are usually transformed from initially informal statements into more consistent and unambiguous models (Tichy and Koerner 2010). Especially in large IS development projects, this is a challenging task as a huge number of natural language documents becomes available and needs to be analyzed. In these cases, manual requirements elicitation can become time-consuming, error-prone, and monotonous, especially if it has to be repeated multiple times when updates to previously existing documents become available (Ambriola and Gervasi 2006; Huffman Hayes et al. 2005). Given the fact that requirements engineers are scarce, valuable resources, a more efficient way to support their work seems imperative. In a study about current requirements elicitation practices, Mich et al. (2004) asked more than 150 software developers to name the two things in their job they would like to do more efficiently. The activity they named most frequently (46%) was “identify user requirements.” In the same study, participants were asked about the most useful thing to improve general day-to-day efficiency of this task: The majority (69%) chose “automation.”

Requirements elicitation is a broad, comprehensive process that delivers multiple, complementary views on the information system to be built. Within this extensive range, we focus our study on the elicitation of tasks to be supported by the software. The importance of this activity has been recognized across multiple IS domains (most importantly Requirements Engineering and Human Computer Interaction) and is a core activity in many IS development approaches (Robertson and Robertson 2006; Sharp et al. 2007). Depending on the field of research, different models are suggested to consistently capture the outcomes of this process. While the field of Human Computer Interaction (HCI) proposes the usage of task models (Paterno 2002; Tam et al. 1998), researchers in the Requirements Engineering field suggest similar models, for example structured scenarios (Pohl 2010; Robertson and Robertson 2006). Similar to general requirements elicitation, the manual elicitation of tasks from natural language documents is a tedious and error-prone activity, demanding for automation (Lemaigre et al. 2008). Various works presented approaches to provide this automation by the means of specific task elicitation systems (Brasser and Vander Linden 2002; Lemaigre et al. 2008; Tam et al. 1998). However, few efforts have been made to extend the existing body of knowledge understanding how these systems affect the individual performance of requirements engineers and which design principles contribute to this.

Accordingly, we suggest a study to systematize previous findings through the formulation of design principles and put them in a socio-technical context by evaluating productivity in an experiment. We thereby provide a basis to conceptually advance the class of task elicitation systems. Consequently, the goal of our work is to answer the question: *Which design principles of task elicitation systems improve task elicitation productivity over manual task elicitation?* Following a Design Science approach, we identify meta-requirements based on expert interviews and the existing body of knowledge. We then conceptualize design principles and instantiate them in an artifact. The artifact is used to measure effects of the identified design principles on task elicitation productivity in two experiments; one in a laboratory and one in a field setting. Through our work, we intend to contribute to the design theory body of knowledge by explaining the effects of different design principles of task elicitation systems on elicitation productivity. From a practical perspective, our study helps software vendors to improve the elicitation capabilities of their Requirements Engineering software packages and hereby reduce the current problems in task elicitation practice.

The remainder of the paper is organized in the following sections: The second section summarizes the foundations of our work and the state of the art. Subsequently, the conceptualization of design principles and the research model is presented. In the fourth section the evaluation methodology is depicted, followed by the evaluation results. The sixth section includes a discussion of the results, which is followed by a summary of limitations, future activities, and contributions of our work.

Foundations and Related Work

Requirements elicitation is the process of discovering requirements through direct interaction with stakeholders or analysis of documents or other sources of information (Ratchev et al. 2003). A core activity in this process is the identification of relevant tasks to be supported by the software, referred to as task elicitation (or sometimes task analysis) (Lemaigre et al. 2008; Paterno 2002). More specifically, task elicitation aims at capturing the interaction between user and system on a detailed level, differentiating between actors, actions, and objects (Tam et al. 1998). The results of this activity can be used to develop various types of formalizations (e.g., structured scenarios, task models, or domain models).

To support task elicitation, various systems – referred to as Task Elicitation Systems (TES) – have been proposed. These systems support requirements engineers through (partial) automation of the elicitation process. Starting from natural language documents containing task information, TES aim at deriving the described abstractions (e.g., task models). Automation is achieved by various algorithms (e.g., natural language processing techniques or information retrieval techniques) combined with the usage of a corresponding knowledge base. These knowledge bases contain a collection of potential task elements and a categorization of these elements. Categorizations can include an assignment of a task element to a specific task category (e.g., "activity"), domain (e.g., "purchasing") or word class (e.g., "noun") and serve as meta-information for the automation algorithm (Brasser and Vander Linden 2002; Lemaigre et al. 2008). The creation of knowledge is either initiated by an upload of existing knowledge to the system (referred to as "imported knowledge") or knowledge retrieval from documents (referred to as "retrieved knowledge") (Staab et al. 2001). In the case of TES, retrieved knowledge can be obtained from natural language documents containing requirements (or more specifically task information).

Various attempts have been made to improve requirements and task elicitation from natural language documents both in the field of Requirements Engineering as well as in the field of Human Computer Interaction (HCI).

In the Requirements Engineering domain, three corresponding research streams can be differentiated. The first research stream aims at the automatic *identification of abstractions* from natural language documents which will, for example, assist an analyst in gaining an understanding of an unfamiliar domain (Berry et al. 2012). In this context abstractions represent single words within the requirements document that form a representation of necessary domain knowledge (Gacitua et al. 2011). This domain knowledge can then be both a reference as well as a starting point during the subsequent elicitation of requirements and helps to avoid information overload and to overlook important aspects that might evolve into requirements (Berry et al. 2012). Artifacts that support abstraction identification through automatisms have been proposed by Gacitua et al. (2011), Goldin and Berry (1997), Kof (2004), and Rayson et al. (2000). In contrast to our work, however, these artifacts do not address the specifics of task elicitation. They focus on the identification of relevant terms to build domain knowledge rather than on the actual elicitation of relevant requirements or tasks and their transformation into models.

A second research stream within this domain aims at the automatic *identification and classification of requirements* within natural language documents. Though including parts of the processing necessary for the first research stream, the main goal here is not the creation of domain knowledge or domain understanding, but the actual elicitation of requirements. The artifacts presented by Cleland-Huang et al. (2007) and Casamayor et al. (2010) focus on the identification and classification of non-functional requirements. In addition to the optimization of the elicitation process itself, both works also describe more efficient ways to build up the knowledge base that is needed for the automatism. However, they specifically focus on non-functional requirements which are usually distinct statements within a natural language document that are not integrated in one consistent model; as they would be in the case of task elicitation. Similarly, the artifact presented by Kiyavitskaya and Zannone (2008) also focuses on a specific type of requirements (in this case security and privacy requirements). More specifically their approach supports the elicitation of these requirements from textual scenarios resulting in semi-structured, tabular templates. These templates structure the requirements from different viewpoints, highlighting for example different actors, risks and privacy aspects. Vlas and Robinson (2011) present an artifact that particularly addresses requirements identification and classification for open-source software development projects. Consequently, their approach is optimized to the discovery of requirements within chats, email, and forums, which are possible, but unusual sources for the elicitation of task models as they mix requirements and task information with social communications, code segments, or slang. In summary, the described works within this second research stream focus either on the elicitation of particular types of requirements (e.g., non-functional requirements) or on the utilization of specific data sources that are unrelated or at least unusual to task elicitation (e.g., chats, email). Furthermore, they result in distinct statements within a natural language document that are not integrated in one consistent model; as they would be in the case of task elicitation. An alternative approach to identify and classify requirements is presented by Kaiya and Saeki (2006). Based on already classified requirements and a domain ontology, their artifact detects incomplete or incorrect requirements and recommends extensions. In contrast to the goal of this paper, their work requires an initial (manual) identification and categorization before the artifact can be applied and, therefore, rather aims at the improvement of first elicitation results. This might be a subsequent activity of the process we are willing to support.

Third, adding a further step of processing, some attempts have been made to use automated analysis of natural language to *create requirements and design models*. While early works (Buchholz et al. 1995; Rolland and Proix 1992) primarily focused on the creation of static models (e.g., entity relationship diagrams), more recent concepts (Ambriola and Gervasi 2006) specifically include dynamic models (e.g., UML sequence diagrams). However, these cannot be applied to the context of task elicitation: the early works are restricted to static models, which cannot cover user interactions with the system. The latter work focuses on models describing the system-internal structure and behavior; rather than the interaction between user and system and requires the existence of formally specified requirements written in restricted natural language (Ambriola and Gervasi 2006). In contrast to all previously described approaches, in which a first automatic elicitation of requirements is followed by manual adaptions, the work of Shibaoka et al. (2007) proposes to proceed in reverse order: During the manual creation of a requirements model, the implemented artifact automatically proposes items from an ontology which might be appropriate in the current step of modeling. While their approach is a considerable alternative to existing works, it requires a significant manual effort (which is also shown in their first evaluation results). Therefore, it might not be an appropriate approach to pursue our goal to improve elicitation productivity.

In the HCI domain, several works specifically describe how to (partially) automate task elicitation with corresponding artifacts. Tam et al. (1998) developed a TES named U-TEL that enables designers to transform passages of a textual scenario into elements of three models: action names (relevant for task models), user classes (relevant for user models), and objects names (relevant for domain models). The allocation can be conducted either manually or automatically. Brasser and Vander Linden (2002) propose the use of natural language parsing to automatically extract task information from written task narratives. Their system extracts two kinds of information: domain information (i.e., actors and objects) and procedural information (e.g., "when the user saves a file, …"). Lemaigre et al. (2008) developed a tool for model elicitation from textual scenarios to conduct model-driven engineering of user interfaces. Their artifact employs manual classification, dictionary-based classification, and nearly natural language understanding based on semantic tagging and chunk extraction. It uses a more detailed classification scheme than former works (Brasser and Vander Linden 2002; Tam et al. 1998).

While the TES that evolved from the HCI domain aim at a similar target like our work, they have several shortcomings. To enable automated task elicitation, a knowledge base with relevant terms is required (Brasser and Vander Linden 2002; Lemaigre et al. 2008). All depicted works require building up a knowledge base in a separate, explicit process. The explicit creation of a knowledge base can be very time-consuming and therefore puts the added value of the automatism into question. Consequently, an investigation of more efficient ways to extend this knowledge could conceptually advance the class of TES.

Furthermore, although the (partial) automation of the task elicitation process predominantly aims at improving the productivity of requirements engineers, previous evaluations paid little attention to evaluating corresponding effects. Due to the ambiguity and inconsistency of natural language documents, results of automated elicitation in most cases require manual rework to correct mistakes of the automatism, adapt its findings, or add task elements which were overlooked (Cleland-Huang et al. 2007).

This raises the question if automation really improves overall productivity (in comparison to manual elicitation). Consequently, the mentioned works could be complemented with a study investigating whether the use of a corresponding system actually improves individual performance – or more specifically individual productivity – by comparing it to a manual approach.

Finally, while the studies described above include detailed descriptions of their specific implementations, an abstraction of the demands to be fulfilled by the system and the concepts addressing each of these demands is missing. This abstraction to "meta-requirements" and "meta-design" / "design principles" (Markus et al. 2002; Walls and Sawy 1992) and the mapping process between them has been intensively discussed in Design Science Research (Baskerville and Pries-Heje 2010; Gregor and Jones 2007). An according conceptualization enables a generalization of design approaches going beyond the description of specific solutions to specific problems. Applying this approach to TES, the theoretical contribution drawn from previous works can be extended substantially.

Our work is intended to address the depicted gaps by answering the research question we defined above. We thus aim at deriving design principles of TES and evaluating their effect on task elicitation productivity in comparison to manual task elicitation.

Conceptualization

To be able to shape and evaluate the design principles addressed in the research question, we follow a design science approach as suggested by Hevner et al. (2004). Our approach was guided by the general design cycle described by Kuechler and Vaishnavi (2008). Starting from meta-requirements, which were identified based on expert interviews and the existing body of knowledge, we derived design principles. These design principles were instantiated in an artifact designed in three consecutive design cycles over a time period of 12 months. The design included several qualitative evaluations and subsequent refinements of the design principles. This chapter presents the final conceptualization of the design principles and the research model for their evaluation.

Design Principles

According to Gorschek and Davis (2008), most companies target one of two goals when assessing the success of their requirements engineering activities: (1) improving the *process* itself, for example by measuring the time and/or resources consumed while requirements engineering is performed or benchmarking the process against a set of "best practices" or (2) improving the primary *produc*t of the requirements process through, for example, a measurement of the quality of the software requirements. In our work, we combine the process and product dimension in a productivity variable, representing an input-output ratio wherein the quality of the elicited tasks serves as the output part (numerator of the ratio) and the invested elicitation effort as the input part (denominator). As we specifically investigate the task elicitation process, we refer to this variable as task elicitation productivity.

As depicted in the introduction, manual task elicitation can be both time-consuming and error-prone. These two problems can be directly mapped to the described input-output ratio: While a reduction of the invested elicitation time decreases the denominator of the ratio (input), a reduction of the error-rate improves elicitation quality (output). Based on this observation we derive the following principle goal for TES that will be addressed by each of the following design principles:

Principle Goal: *Task Elicitation Systems (TES) should improve task elicitation productivity.*

During the elicitation of tasks from natural language documents, a text is analyzed to identify relevant words and assign them to task categories. This process can be decomposed into single steps which are repeatedly performed and follow specific rules (Brasser and Vander Linden 2002). Consequently, they can be translated into algorithms that can be executed by a computer, making it possible to free the requirements engineer from a major part of this activity by automating elicitation. In most cases, however, the ambiguity and inconsistency of natural language documents require that results of automated elicitation are manually reworked to correct mistakes of the automatism, adapt its findings, or add task elements that were overlooked. Consequently, the automatism needs to be complemented with functionality supporting manual elicitation (Kiyavitskaya and Zannone 2008; Lemaigre et al. 2008; Tam et al. 1998). Any manual adaptation of automatically elicited tasks represents an additional effort for the requirements engineer that may reduce overall productivity. To limit this effect, functionality for manual elicitation should provide a high level of usability and performance to enable efficient operations. In summary, we propose:

***DP1. Semi-Automatic Task Elicitation**: TES should support efficient automatic and manual task elicitation within natural language documents.*

As illustrated earlier, automated task elicitation requires an underlying knowledge base containing task elements and a categorization of these elements. Similarly to the task elicitation itself, the creation of a knowledge base can be a time-consuming task (Ambriola and Gervasi 2006). Looking at overall task elicitation productivity, these efforts have to be included and large additional expenses for knowledge creation should not put the added value of automatic elicitation into question. In our approach, we propose to solve this dilemma through supervised inclusion of elicited task elements in the knowledge base and the re-use of this knowledge in subsequent elicitations. Supervision is supported by a restriction of this functionality to a specific user role (expert users) and the calculation of probability values for the assignment of task elements to categories. The overall mechanism results in an automatic supplementation of the imported knowledge that has been initially loaded into the knowledge base with retrieved knowledge from the elicitation process. As shown in the related work section, existing works concentrate on either building up requirements / task knowledge from natural language documents (Gacitua et al. 2011; Goldin and Berry 1997; Kof 2004; Rayson et al. 2000) or using existing task knowledge to support the elicitation itself (Brasser and Vander Linden 2002; Lemaigre et al. 2008; Tam et al. 1998). In our concept, these two approaches are combined in a closed loop to reduce knowledge creation efforts. In the elicitation process itself, retrieved knowledge can have two positive effects: Additionally to the mere extension of the knowledge base, it can also add domain-specificity. Documents that originate from the same domain share specific task elements, which are not included in general imported knowledge (Lemaigre et al. 2008) (e.g., the data field "frequent flyer number" in the domain "traveling"). Similarly, specific writing styles or standards for single projects or entire organizations can result in needs to extend imported knowledge (Cleland-Huang et al. 2007). Feeding back the results of the elicitation process into the knowledge base allows leveraging the task information within the documents and thereby improving the outcome of the automation algorithm. As a result, we propose:

***DP2. Usage of imported and retrieved knowledge:** TES should use both manually imported and automatically retrieved knowledge during automatic elicitation.*

In the final step of the conceptualization, the identified design principles are mapped to design features. Design features correspond to specific artifact capabilities, for example the algorithm chosen for automatic elicitation[1].

Research Model

As suggested by Gregor and Jones (2007) we formulate testable hypotheses to be able to evaluate our design. More specifically, our research model strives to measure effects of alternative combinations of the depicted design principles on multiple dependent variables.

As introduced earlier we conceptualize *task elicitation productivity* as an input-output ratio wherein the quality of the elicited tasks serves as the output part (numerator of the ratio) and the invested elicitation effort as the input part (denominator). We use this ratio as our dependent variable. To evaluate the quality of automatically elicited requirements, precision and recall are common measures (Casamayor et al. 2010; Cleland-Huang et al. 2007; Gacitua et al. 2011) that can be applied to task elicitation as well (Brasser and Vander Linden 2002). They are calculated by comparing participants' task elicitation outputs with expert solutions similar to the studies done by Kiyavitskaya and Zannone (2008) and Vlas and Robinson (2011). *Recall* can be seen as a measure of completeness, comparing the number of correctly identified task elements with the total number of task elements existing in a document.

[1] The concrete implementation of the artifact is described in Meth et al. (2012).

Precision represents a measure of correctness and is calculated as the proportion of correctly identified task elements in comparison to the number of **identified** task elements in a document.

The input factor *task elicitation effort* can be measured by the time required to conduct the elicitation task, that is, transforming an unstructured input document to a set of formally specified tasks. As the evaluation was based on an experiment with a fixed time schedule, task elicitation effort was also fixed and only the differences in recall and precision (i.e., the quality of the elicited tasks) were measured. Consequently the evaluation measured productivity in a fixed time period, similar to the studies done by Diehl and Stroebe (1991) and Gallupe and McKeen (1990).

The conceptualization of the independent variable is directly linked to the design principles of the artifact. Both design principles can be switched on and off resulting in different TES configurations that can be evaluated separately. For example, semi-automatic task elicitation (DP1) would be switched on, while the usage of retrieved knowledge (DP2) would be switched off. While DP1 can be switched on independently from DP2, DP2 can only be activated when DP1 is switched on. This resulted in the three *TES configurations* depicted in Table 1.

Table 1. TES configurations

TES configuration	Design Principle Activation	
	DP1	DP2
(1) Manual elicitation		
(2) Semi-automatic elicitation with imported knowledge	**X**	
(3) Semi-automatic elicitation with imported and retrieved knowledge	**X**	**X**

We expect various effects of the design principles of our artifact on task elicitation productivity.

Expected productivity effects of DP1 (Semi-Automatic Task Elicitation) as measured by recall in a fixed time period: Process automation is a well-known mechanism to improve productivity both for IT supported processes as well as non-IT supported processes (Atkinson and Kuhne 2003; Jämsä-Jounela 2007). In the case of automated task elicitation, it can be expected that productivity (measured by recall in a fixed time period) will similarly improve, as an algorithm can automatically identify a large percentage of task elements without spending the requirements engineer's time during the analysis (Cleland-Huang et al. 2007; Kiyavitskaya and Zannone 2008; Vlas and Robinson 2011).

The recall using a semi-automatic TES can be seen as a sum of the automatism's initial recall and the recall resulting from subsequent manual adaptions and extensions. These subsequent manual activities are comparable to using a purely manual TES: a requirements engineer needs to read a natural language text document, mark passages containing task elements and assign task element categories to them. Therefore we do not expect a significant recall difference between semi-automatic and manual TES from these manual activities. In contrast, the initial recall provided by the automatism will remain and can be expected to have a significant effect. We therefore hypothesize that:

> *H1: In a fixed time period, the use of TES that support semi-automatic task elicitation with imported knowledge will result in higher recall than the use of TES that only support manual task elicitation.*

This hypotheses is additionally supported by Information Processing Theory (Miller 1956), explaining that human information processing is restricted by cognitive limitations. Information systems supporting elicitation activities through automation can help to overcome or at least reduce these limitations resulting in a larger amount of identified task elements in a fixed time period.

Expected productivity effects of DP2 (Usage of imported and retrieved knowledge) as measured by recall in a fixed time period: As described above, automated task elicitation requires a knowledge base containing task elements and a categorization of these elements. Each automatically elicited task can be traced back to a specific entry in this knowledge base. Accordingly, a more elaborate and extensive knowledge base can generally be expected to result in a higher percentage of elicited tasks and therefore a reduction of manual efforts (Cleland-Huang et al. 2007). Additionally to the size of the knowledge base, the domain-specificity of the knowledge plays an important role in the task elicitation process (Brasser and Vander Linden 2002; Tam et al. 1998). Generally, a higher degree of domain-specificity is expected to deliver better elicitation results (Lemaigre et al. 2008), for example by including domain-specific task elements (like "physician" or "nurse") additionally to domain-independent ones (like "manager" or "worker"). As depicted earlier, we propose to use two sources of knowledge to fill the knowledge base. Additionally to manually imported knowledge, which is commonly used in existing TES (Brasser and Vander Linden 2002; Lemaigre et al. 2008), the content of the knowledge base can be extended by automatically retrieved knowledge originating from documents that have been processed before. As described in the conceptualization of DP2, this should increase both size and domain-specificity of the knowledge base. Therefore we hypothesize that:

> *H2: In a fixed time period, the use of TES that support semi-automatic task elicitation with imported and retrieved knowledge will result in higher recall than the use of TES that only support semi-automatic task elicitation with imported knowledge.*

Expected productivity effects of DP1 and DP2 as measured by precision in a fixed time period: As depicted earlier, both recall and precision determine requirements quality and therefore are of utmost importance for the overall requirements elicitation process. However, in *automated* requirements elicitation from natural language documents, recall is significantly more important than precision, as it is a much simpler activity for a requirements engineer to evaluate a set of candidate requirements and reject the unwanted ones than it is to browse through an entire document looking for entirely missed ones (Cleland-Huang et al. 2007). The same argument is used by Berry et al. (2012) who state that requirements engineering tools that treat natural language documents "should be tuned to favour recall over precision because errors of commission are generally easier to correct than errors of omission" (Berry et al. 2012, p.213). Because of that, the design principles of the artifact primarily address an improvement of the recall rate and do not target precision improvements.

Furthermore, while the recall rate is predominantly determined by the automatism's ability to find as many relevant words and text passages as possible, the precision rate is strongly linked to the quality of the decision-making following the identification of a word/text passage. A significant precision improvement could therefore only be realized if the algorithm would provide better decision-making capabilities than a human being doing manual task elicitation, which we do not expect. Therefore, we hypothesize that:

> *H3: In a fixed time period, the use of manual TES, TES that support semi-automatic task elicitation with imported knowledge, and TES that support semi-automatic task elicitation with imported and retrieved knowledge does not result in significant differences in precision.*

Figure 1 summarizes our hypotheses in a comprehensive research model.

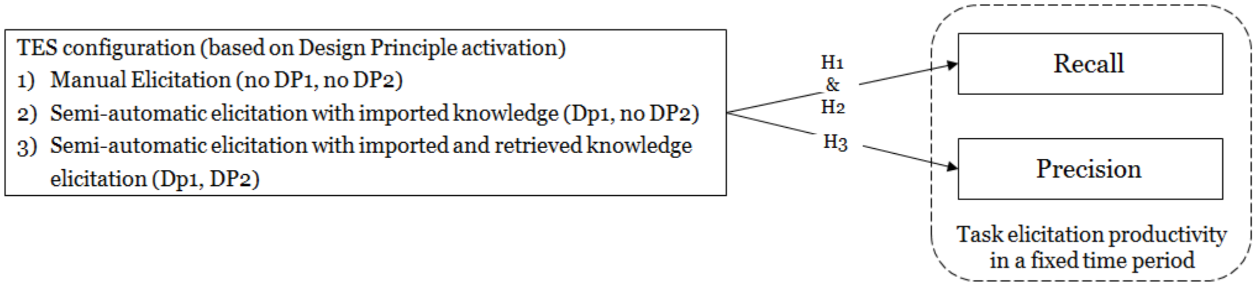
TES configuration (based on Design Principle activation)
1) Manual Elicitation (no DP1, no DP2)
2) Semi-automatic elicitation with imported knowledge (Dp1, no DP2)
3) Semi-automatic elicitation with imported and retrieved knowledge elicitation (Dp1, DP2)
H1 & H2
H3
Recall
Precision
Task elicitation productivity in a fixed time period

Figure 1. Research Model

Evaluation Methodology

Previous works on automated requirements elicitation have chosen both qualitative and quantitative evaluation approaches. In the former group, some evaluated the corresponding artifact by a mere demonstration, for instance through an application to a real-world example without data collection and analysis (Kof 2004; Lemaigre et al. 2008). While these evaluations are typically performed in a laboratory setting, other works have been evaluated in a field setting, applying case study methodology (Ambriola 2006). Apart from these qualitative assessments, a significant number of artifacts have been evaluated quantitatively using simulations (Casamayor et al. 2010; Cleland-Huang et al. 2007; Gacitua et al. 2011). In simulations, the artifact's output is usually compared to a "gold standard," a set of requirements, which is elicited manually by an expert or a group of experts. Finally, controlled experiments represent another possible quantitative evaluation approach. Following this approach, the performance of an experiment participant supported by an artifact is compared with the performance of a participant devoid of its support. Few previous works on automated requirements elicitation have chosen experimental evaluations (Kaiya and Saeki 2006; Shibaoka et al. 2007). While promising in terms of methodology, the limited sample size used in the examples quoted above seems to warrant that future experimental evaluation should draw upon a larger sample.

Following Hevner and Chatterjee's (2010) suggestion, we evaluated the artifact with a controlled experiment to rigorously test the effect of two design principles (DP1, DP2) on task elicitation productivity. The overall experiment consisted of a laboratory experiment and a field experiment. First, we evaluated the artifact in a laboratory setting with student participants. By using a laboratory experiment, we can accurately adjust the design principles and measure their impacts on task elicitation productivity while controlling for potential influential factors (e.g., task elicitation knowledge, motivation); by using student participants, we can obtain a relatively large sample size with reasonable efforts and achieve adequate statistical power (Gallupe and McKeen 1990).

Second, to evaluate the generalizability of findings from the student participants, we carried out the same experiment with a small sample of requirements engineers in a field setting. By comparing the behavioral patterns of the two groups of participants, we can evaluate the external validity of the results from the laboratory setting. It should be noted that we did not intent to merge the two samples to test the hypotheses, but only used the results of the small sample of requirements engineers as an examination of the student sample's external validity. All conclusions from the experiment should be reliably drawn from the relatively large sample of students.

We used a single factor within-subject design for both the laboratory experiment and the field experiment to increase statistical power for each experimental setting and reduce error variance introduced by individual differences (Hill and Lewicki 2007). The within-subject factor is the TES configuration. This factor has three levels: manual elicitation, semi-automatic elicitation with imported knowledge, and semi-automatic elicitation with imported and retrieved knowledge.

Pilot Test

A pilot test was conducted to estimate the necessary sample size and appropriate length of the interview transcripts used in the experimental tasks. The same single factor within-subject design was applied in the pilot test as in the main experiment, and three graduate students participated in the pilot test. The results indicated the lowest correlation among the repeated measures was 0.35. Calculated with G*Power 3 (Faul et al. 2007), to detect a medium effect (f= 0.25) at the significance level of 0.05 with a sufficient statistical power (about 0.80) (Cohen 1988, pp. 273 - pp.288), the sample size should be at least 35. Thus, we set the sample size for the laboratory experiment to be 40 to detect a medium effect on recall and on precision.

Within the experimental time for each task (5 minutes), the maximum amount of words that the participants processed was 247, 277, and 328 for manual task elicitation, semi-automatic task elicitation with imported knowledge, and semi-automatic task elicitation with imported and retrieved knowledge respectively. Accordingly, we set the length of the interview transcripts used in the main experiment to be of 325 words. With this length, most of the participants are expected not to be able to completely process all the text within the experimental time, but they can achieve their optimal recall and precision while working at their normal pace. A very small number of participants might be extraordinarily fast in task elicitation and be able to complete the first round of task elicitation within the experimental time, allowing them to further improve recall and precision in the remaining time by checking the first round results. We did not set up the interview transcript to be of a length that no participant could possibly complete the first round of task elicitation, because we wanted to limit the impact of the automatically elicited task elements on the achieved recall and precision. Participants should be able to read and check the automatically elicited task elements within the task time, which aligns to the application situation in practice.

Participants and Experimental Procedures

According to the sample size calculation, 40 participants were recruited for the laboratory experiment. The participants were graduate students enrolled in a master level IS course in a public university with an average age of 25.4 years (*SD*=2.07). Thirty-two of the participants were male and eight of them were female. Participants were evenly assigned to six time slots on three experimental days, with 6 or 7 participants per time slot.

The experiment was carried out in a multimedia classroom in the university. A lecturer of the IS course introduced the experiment as an exercise for a course-related assignment with the objectives of understanding different requirements categories relevant for Business Intelligence and learning how to use a web application to perform task elicitation from text documents. No participant had access to the TES before the experiment and all participants were naive to the purpose of the experiment. To teach how to perform task elicitation and how to use the web application to perform task elicitation, the lecturer presented a tutorial video to the participants. Then, the participants were asked to fill in a brief questionnaire about their demographic information and how often they performed task elicitation previously. Measured on a five point Likert scale (1-Never, 5-Very frequently), the task elicitation experience yielded an average value of 1.78 (*SD*=0.85), which indicated that the student participants had low prior experience in task elicitation.

Next, the lecturer guided the participants through a training session to make them familiar with task elicitation. The participants were asked to perform task elicitation using an interview transcript about requirements of a train reservation application for smartphones. In the first five minutes, they conducted task elicitation manually; in the next five minutes, they performed task elicitation within the same transcripts again but with a few automatically elicited task elements at the beginning. Afterwards the lecturer presented the expert task elicitation results for the transcript and answered any question raised by the participants. Then the participants were allowed for a five-minute break.

After the break, the lecturer asked the participants to practice their task elicitation skills with a different set of interview transcripts, which contained three transcripts about requirements of a car sharing application for smartphones. By design, task elicitation within the three interview transcripts was supported with three different TES configurations. To compensate for learning and fatigue effects in the within-subject design, the presentation order of the three TES configurations was fully counterbalanced across the participants, yielding a total of six orders; the participants were randomly assigned into one of the six orders of TES configurations. For each interview transcript, the participants were given five minutes to perform the task elicitation. Then they were asked to report their task elicitation knowledge (e.g., According to your experience in the last task, the meanings of different categories were clear/unclear) on a 7-point semantic differential scale (3 items, adapted from Alba 1983) and their task motivation (e.g., I put a lot of effort into coming up with the best possible solution) on a 7-point Likert scale (3 items, adapted from Maynard and Hakel 1997). When all the participants finished the questions, they were instructed to switch to the next interview transcript and start task elicitation on it.

In the field experiment, participants were five requirements engineers (targeted users of the TES) recruited from a large high-tech company. The practitioner sample consisted of three males and two females with an average age of 34.8 (*SD*=3.56) and an average experience of 5.0 years (*SD*=5.83) and 3.6 years (*SD*= 1.14) in requirements engineering and task elicitation respectively.

The participants in the field experiment followed similar experimental procedures as the ones in the laboratory experiment, with a few necessary modifications. Firstly, the participants were randomly assigned into one of the orders of the TES configurations; since only five participants were involved in the field experiment and each participant got a different order of the TES configurations through randomization, thus five among the six orders of the TES configurations were covered in the field experiment. Secondly, the purpose of the study was introduced as "to get experts' opinions on future design of TES"; no participant had access to the TES before the experimental tasks and the participants were unaware of the real purpose of the experiment. The participants were told to work at their normal working pace in different tasks. All the other procedures in the field experiment were the same as the ones in the laboratory experiment.

Experimental Tasks and Materials

In the experiment, participants were instructed to perform task elicitation within interview transcripts with different TES configurations. To set up the experimental tasks, we followed three steps: choose a task domain, gather interview transcripts, and set up the semi-automatic task elicitation.

A task domain determines the area of knowledge that participants and the semi-automatic TES rely on in order to elicit task elements. Some task domains require specialized knowledge and expertise (e.g., computer aided design), while the others only require routine knowledge that can be easily acquired in ordinary life (e.g., online shopping). In the experiment, we chose "travel management" as the task domain, since this domain does not require specialized knowledge, and the student participants would be able to elicit task elements with their routine knowledge.

Next, in order to get users' requirements through interviews, two smartphone applications were specified within the "travel management" domain: a train reservation application and a car sharing application. We carried out interviews with 12 potential end-users to gather their requirements on the two applications; each interview lasted 5-10 minutes and it was transcribed into transcripts after the interview. In the experiment, participants were provided a training session to get used to the TES before the experimental tasks. To reduce the practice effect, we specified interview transcripts on different smartphone applications for the training and for the experimental tasks respectively. In the training, a short transcript about requirements of the train reservation application was provided (238 words); in the experimental tasks, transcripts about requirements of the car sharing application were used. For the transcripts used in the experimental tasks, we controlled on the length, readability, and the distribution of task elements. Each transcript was edited to contain 325 words without sacrificing the integrity and meaningfulness of the interview content. Examined by the Flesch-Kincaid score, the three transcripts have similar and high readability (*M*=75.1, *SD*=3.50), which indicates that all the transcripts were highly readable for university students at master level (Kincaid et al. 1975). To examine the distribution of the task elements in the transcripts, two task elicitation experts analyzed the transcripts independently; their task elicitation results were compared and any inconsistency was discussed and resolved. The converged expert solutions showed that the three transcripts contained a relatively equal amount of task elements (*M*=70.3, *SD*=2.09) and that the task elements were evenly distributed across the complete text of each transcript.

Finally, we set up the semi-automatic task elicitation within the "travel management" task domain. To prepare for the semi-automatic task elicitation with imported knowledge, we extracted relevant task elements from different data sources: for the role category, we extracted a list of pronouns from Oxford Dictionary; for the activity category, we extracted a list of action verbs from Hart (2004); for the data category, we used the master data from a SAP Travel Management application (SAP 2012); for the non-functional category, we extracted usability goals and design behaviors from Sharp et al. (2007). The extracted lists were imported to the TES and an automatic task elicitation was run to generate semi-automatic task elements in the selected interview transcripts for the experimental task. The resulting average recall and precision was 54.0% (*SD*=9.4%) and 79.0% (*SD*=6.9%) respectively.

In contrast to imported knowledge, retrieved knowledge does not require additional efforts to be acquired. Retrieved knowledge for a task domain is acquired automatically by the TES when users perform task elicitation on any text document within the specific task domain. To acquire the retrieved knowledge for the "travel management" task domain, one task elicitation expert performed task elicitation with the TES on a set of interview transcripts about the train reservation application. The choice of the transcripts ensured that knowledge was retrieved within the same task domain (travel management), but for an application different from the car sharing application used in the experimental tasks, which made the knowledge retrieving process closely aligned to the real situation in practice. With imported and retrieved knowledge, we achieved an average recall of 75.0% (*SD*=4.2%) and an average precision of

75.0% (*SD*=4.0%) after running the automatic task elicitation on the three interview transcripts for the car sharing application.

In the experimental tasks, the order of the three interview transcripts was randomized across the participants. Automatic task elicitation with either “imported knowledge” or “imported and retrieved knowledge” was performed before the participants started the task elicitation. Therefore, in the experimental conditions “manual task elicitation,” “semi-automatic task elicitation with imported knowledge,” and “semi-automatic task elicitation with imported and retrieved knowledge,” the participants started with a transcript containing “no automatically elicited task elements,” “automatically elicited task elements based on imported knowledge,” and “automatically elicited task elements based on imported and retrieved knowledge” respectively[2].

Measurements of the Dependent Variables

As described in the research model, task elicitation productivity was measured by the achieved quality within a fixed time frame. We evaluated participants’ task elicitation quality with two variables: recall and precision. Following the approach by Kiyavitskaya and Zannone (2008) and Vlas and Robinson (2011), we obtained recall and precision by comparing participants’ task elicitation outputs with the expert solutions. Within a text document, if a participant identified a text segment as one task element, no matter in which requirements category the participant classified this task element, it was counted as one “identified task element.” If the participant identified a text segment as one task element and assigned it to a requirements category in the same way as shown in the expert solution, this task element was counted as one “correctly identified task element.” As shown in Table 2, a participant’s achieved recall for a text document was calculated as a ratio of the number of correctly identified task elements by the participant to the total number of task elements contained in this text document according to the expert solution. A participant’s achieved precision for a text document was calculated as a ratio of the number of correctly identified task elements by the participant to the total number of identified task elements by the participant. To reduce the bias introduced by document analysts, two task elicitation experts analyzed 10% of the participants’ outputs independently and achieved an inter-rater reliability of 98.97%; afterwards, the two experts spilt the remaining outputs and analyzed them separately.

Table 2. Measurements of the Dependent Variables

Variable	Explanation
Recall	$\frac{\text{Number of correctly identified task elements by the participant}}{\text{Total number of task elements in the expert solution}}$
Precision	$\frac{\text{Number of correctly identified task elements by the participant}}{\text{Total number of }\textbf{identified}\text{ task elements by the participant}}$

Data Analysis and Results

All the data analysis was conducted using SPSS for Windows Version 16.0. First, the data obtained from the laboratory experiment was examined and used to test the hypotheses. Then, as an estimation of the external validity of the laboratory experiment, the data from the field experiment was analyzed and compared with the data from the laboratory experiment.

[2] The detailed design of the TES and the mechanisms of knowledge importing and retrieving was described in Meth et al. (2012).

Preliminary Analysis

Prior studies suggest that individuals' task related knowledge and motivation during the task process affect their performance in the task (Maynard and Hakel 1997). In the experiment, participants performed three task elicitations with the support of different TES configurations in a consecutive manner, their task elicitation knowledge might increase over time due to the learning effect, and their motivation of successfully completing the task elicitation might decrease over time due to the fatigue effect (Seltman 2012). To compensate for the learning and the fatigue effect, we counterbalanced the presentation order of the TES configurations. To further examine the learning and fatigue effects on the dependent variables, after each experimental task, we measured participants' task elicitation knowledge and their task motivation with two scales. Evaluated with Cronbach's α, the two scales achieved acceptable internal reliability (task elicitation knowledge: α= .81; motivation: α= .71) (Nunnally and Bernstein 1994) and thus the average scores on the scales were taken as the measurements for the task elicitation knowledge and for the task motivation respectively.

We examined whether the task elicitation knowledge and motivation should be taken as covariates in the hypotheses testing with repeated measures of analysis of covariance (RMANCOVA). Before the RMANCOVA was conducted, outliers and violations of statistical assumptions were examined (Tabachnick and Fidell 2007, p. 201). First, all variables were screened for univariate outliers. Variables with standardized scores exceeding 3.29 (p< .001, two-tailed test) are potential univariate outliers (Tabachnick and Fidell 2007, p. 73; Wang and Benbasat 2007). One case was identified as a potential outlier, of which the standardized score was -3.81. However, no data entry or sampling errors were found in the raw data. We decided to retain the case and follow Tabachnick and Fidell's approach (2007, p. 77) to change the value on this variable to be one unit smaller than the next most extreme value; in this way, the impact of the outlier was reduced. After the change, the data were screened again and no further univariate outlier was found. The investigation of Mahalanobis distances among all the cases did not detect any multivariate outlier (Tabachnick and Fidell 2007, p. 74). As such, all cases were retained for the further analysis. Next, data were checked for violations of statistical assumptions of RMANCOVA. The normality of all dependent variables was examined by skewness and kurtosis as well as visually on the P-P plot and Q-Q plot; no violation of data normality was detected in the dependent variables. Homogeneity of variance among DVs and covariates across the three experimental conditions was evaluated by Levene statistics at a probability level of 0.01; no violation of equal variance of different experimental groups was detected. Finally, the homogeneity of regression between the DVs and covariates across different experimental conditions was checked, and no significant interaction between the covariates and the independent variable in the prediction of the dependent variables was detected. Hence, the assumption of homogeneity of regression was also satisfied. With univariate RMANCOVAs, we examined the covariate effects of task elicitation knowledge and motivation separately on each dependent variable (recall, precision); at the significance level of 0.05, no significant covariate effect was detected on task elicitation knowledge (DV: recall, F= 1.80, p= .18; DV: precision, F=0.23, p= .63) or on task motivation (DV: recall, F=0.28, p= .60; DV: precision, F=2.81, p= .10). Therefore, according to the principle of parsimony (Tabachnick and Fidell 2007, p. 212), we did not include any covariate in the hypotheses testing.

Table 3 presents the means and standard deviations of the dependent variables in different experimental conditions for the laboratory experiment and the field experiment respectively. For the manual task elicitation, the practitioner sample appeared to achieve a relatively lower recall than the student sample; the reasons could be the students were more motivated and concentrated during the experimental task than the practitioners, or the small sample of practitioners might not be evenly distributed on both sides of the true value of the population mean. In hypotheses testing, only the data from the laboratory experiment was used to achieve a sufficient power and get reliable conclusions.

As explained in the research model, task elicitation recall and precision are conceptually independent variables. The hypotheses predict that the TES configurations exert effects on recall and precision in different directions, thus hypotheses on recall and precision should be tested separately with univariate repeated measures of analysis of variance (RMANOVA) (Huberty and Morris 1989). Multivariate analysis was not conducted because multivariate analysis is mainly used to test an overall effect of a treatment on a set of interrelated outcome variables (Huberty and Morris 1989), which is not the focus of this study.

Table 3. Means and Standard Deviations of Recall and Precision for Different TES Configurations

	Manual		Semi-automatic with imported knowledge		Semi-automatic with imported and retrieved knowledge	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Lab experiment (student participants, *N*=40)						
Recall	50.7%	12.0%	69.8%	9.8%	79.5%	8.0%
Precision	71.0%	8.5%	72.0%	6.7%	73.2%	6.5%
Field experiment (practitioner participants, *N*=5)						
Recall	37.6%	12.9%	68.6%	6.0%	77.8%	3.9%
Precision	70.1%	14.5%	72.7%	3.5%	68.5%	5.3%

Hypotheses Testing

With the data from the laboratory experiment, we performed RMANOVA to examine the impacts of the design principles on task elicitation recall and on precision respectively.

As shown in Table 4, participants' recall was significantly influenced by the TES configurations at the significance level of 0.05. To test hypothesis 1 and hypothesis 2, we performed pairwise comparisons on the main effects of TES configurations; a Bonferroni correction was applied to control on the family-wise error rate (Vasey and Thayer 1987). The multiple comparisons results are shown in Table 5. All the pairwise comparisons were significant at the level of 0.05: participants using semi-automatic task elicitation with imported knowledge achieved significantly higher recall than using manual task elicitation, and using semi-automatic task elicitation with imported and retrieved knowledge achieved significantly higher recall than using semi-automatic task elicitation with imported knowledge only. Thus, hypothesis 1 and hypothesis 2 are supported.

Hypothesis 3 was also supported by the RMANOVA on precision (see Table 4): no significant difference in precision across the three TES configurations was detected in the experiment.

Table 4. Results of RMANOVA for Recall and Precision

DV	Source	*DF*	*MS*	*F*	*p*	η^2	Cohen's *f*
Recall	TES Config.	2	0.861	129.76	< .001	.77	1.82
	Error	78	0.007				
Precision	TES Config.	2	0.005	1.36	.263	.03	0.19
	Error	78	0.004				

Table 5. Results of Pairwise Comparisons for Recall

Pair comparison		Mean difference	*p**	95% confidence interval*	
				Lower	Upper
Semi-automatic with imported knowledge	Manual	19.2%	< .001	14.4%	23.9%
Semi-automatic with imported and retrieved knowledge	Semi-automatic with imported knowledge	9.7%	< .001	5.8%	13.6%

* Bonferroni corrections are applied for multiple comparisons

External Validity Evaluation

In the previous section, we tested the hypotheses with the data obtained from student participants in a laboratory setting. Since we wish to generalize the results to requirements engineers who carry out task elicitation activities in workplaces, external validity is a concern for our laboratory experiment with students. However, prior studies suggest that causal relationships are more generalizable across populations than specific characteristics (Pedhazur and Schmelkin 1991), which indicates that the causal relationships between the design principles of TES and improved task elicitation productivity may remain across different samples.

We performed a RMANOVA on recall to compare the effects of different TES configurations. The result showed a significant difference on participants' recall when the TES configuration varied (*F* (2, 8) = 31.74, *p*< .001, η2= .89, *f*=2.82). The pairwise comparisons with Bonferroni corrections indicated that semi-automatic task elicitation with imported knowledge outperformed manual task elicitation on recall (mean difference= 31.0, *p*= .007, 95% CI [13.4%, 48.7%]), but no significant difference was detected between semi-automatic task elicitation with imported knowledge and semi-automatic task elicitation with imported and retrieved knowledge (mean difference= 9.1%, *p*= .301, 95% CI [-7.8%, 26.1%]). When analyzed with a more powerful paired t-test, the difference between the two semi-automatic TES configurations was marginally significant (*t*(4) = 2.13, *p* = .100, 95% CI [-2.8%, 21.0%], *d*= 0.95). The observed effect size was classified as a large effect according to Cohen (1988, pp. 273 - pp.288); thus, the insignificant result might stem from the very small sample size used in the field experiment. We conducted a post-hoc power analysis with G*Power 3 (Faul et al. 2007). The result showed that to detect this effect size (*d*=0.95) with paired t-test, a sufficient power (e.g., 0.80) can be achieved by adding 7 more participants into the practitioner sample, resulting in a total sample size of 12. As we expected, no significant difference was detected on precision with the practitioner sample analyzed by RMANOVA (*F* (2, 8)= 0.34, *p*= .723, η2= .08, *f*=0.29).

In addition, we performed a RMANOVA with the pooled data from the laboratory and the field experiment and specified "role" as a between-subject factor to differentiate the student sample and the practitioner sample. Not surprisingly, at the significance level of 0.05, TES configurations demonstrated the same significant effects on recall (*F* (2, 86) = 84.78, *p*< .001) and no effects on precision (*F* (2, 86)= 0.39, *p*= .682); neither a main effect of role nor an interaction effect between the TES configurations and role was detected on recall and precision.

The above analyses did not reveal evidence that the practitioner sample demonstrated a different behavioral pattern on recall and precision when using the different TES configurations compared with the student sample; there was no evidence showing that the conclusions drawn from the laboratory experiment could not be generalized to practitioners in a field setting. However, due to the small size of the practitioner sample used in the field experiment, the results have to be treated with caution.

Discussion

The evaluation aimed to measure the effects of different design principles of TES on task elicitation productivity in comparison to manual task elicitation. More specifically, we investigated how semi-automatic task elicitation (DP1) and the combined usage of imported and retrieved knowledge (DP2) affect recall and precision of the elicited tasks in a fixed time period.

Concerning DP1, we found that the use of semi-automatic task elicitation significantly improved task elicitation recall over participants' prior task elicitation knowledge and motivation. Different explanation patterns can be applied to this result. First, the automation process provided the participants with an initial set of elicited tasks that already represented a substantial recall (54.0%). Therefore, in comparison to the manual elicitation task, in which participants started with an unprocessed document, a higher final recall could be assumed, provided that participants trust the suggestions of the automatism. The increase of recall from the initial 54.0% (provided by the automatism) to the final 69.8% indicates that participants trusted the recommendations of the automatism sufficiently enough to let them use at least a part of their time to increase recall through manual elicitation of additional tasks (rather than using the entire time to correct potential mistakes of the automatism).

As expected, DP1 did not significantly affect precision. The automatism resulted in an initial precision of 79.0% using imported knowledge, which is comparable to the average precision value achieved during the manual elicitation task (71.0%). Manual adaptations and extensions that were made during the experiment task slightly reduced the initial precision, resulting in a value of 72.0%. This value is between the precision values of the automatism and the value of manual elicitation, which reflects the semi-automatic nature of the task.

Concerning DP2, we found that the additional use of extracted knowledge further improved task elicitation recall. A possible explanation for this effect is that a more extensive, domain-specific knowledge base results in a higher initial recall provided by the automatism. This assumption could be confirmed, as the initial recall of the automatism rose from 54.0% to 75.0% through the activation of DP2. To assess the generalizability of these results, it is important to revisit the corresponding preconditions of our findings. The extension of the knowledge base through extracted knowledge resulted from a previous, manual elicitation by a domain expert. This manual elicitation was based on different documents and a different application context than the experiment itself, but referred to a similar domain (travelling). These quality pre-conditions (extension of knowledge done by an expert and using documents of the same domain) enabled the automatism to elicit tasks with significantly increased recall and with a precision comparable to manual elicitation. Consequently, the final result also showed this recall/precision pattern. To achieve comparable results in a field setting, it is therefore important to enforce the described quality pre-conditions, which can be supported by the TES itself (e.g., through specific expert user roles and the mandatory assignment of documents to domains) or by organizational enforcement (e.g., through recurrent, mandatory quality checks of the knowledge base).

Similar to DP1, DP2 did not significantly affect precision. The automatism resulted in an initial precision of 75.0% using imported knowledge, which is again comparable to the average precision value achieved during the manual elicitation task (71.0%) and therefore can be explained analogously.

Based on the previously introduced definition, task elicitation productivity is the quality of the elicited tasks divided by the invested elicitation effort. Since we measured the invested elicitation effort with time and kept it constant in the experiment, the results support that the deployment of the two design principles can improve task elicitation productivity. Alternatively, the invested elicitation effort can also be measured by the frequency of keystrokes and mouse clicks, which is often termed as physical effort (Tamir et al. 2008). We installed a screen capture tool on participants' computers that automatically captured their keystrokes and mouse clicks during the experiment. Tested with RMANOVA, the frequency of keystrokes and mouse clicks was significantly different across different TES configurations (F (2, 78) = 50.15, p< .001, η2= .56, f=1.14). The pairwise comparisons with Bonferroni corrections showed that the use of semi-automatic task elicitation with imported knowledge significantly reduced the frequency of keystrokes and mouse clicks from an average of 251.2 (*SD*=62.61) to an average of 185.0 (*SD*=55.94) (mean difference= 66.2, p< .001, 95% CI [42.2, 90.2]). The use of semi-automatic task elicitation with imported and retrieved knowledge further reduced the frequency of keystrokes and mouse clicks to an average of 157.1 (*SD*=50.58) (mean difference= 28.0, p= .013, 95% CI [4.9, 51.0]). The invested elicitation effort measured by frequency of keystrokes and mouse clicks was reduced by 37.5% with the deployment of the two design principles. Consequently, task elicitation productivity measured by recall per keystroke or mouse click was significantly improved by the use of semi-automatic task elicitation with imported knowledge (mean difference=0.20%, p< .001, 95% CI [0.15%, 0.25%]) and further improved by the use of semi-automatic task elicitation with imported and retrieved knowledge (mean difference=0.21%, p=0.025, 95% CI [0.02%, 0.39%]). This finding confirms the improvement of task elicitation productivity by the deployment of the two design principles and provides support for reduction of physical efforts by the design principles. Overall, to achieve a certain level of quality of the elicited tasks, participants with the semi-automatic TES require shorter time and invest lower physical effort.

Conclusion

We presented our design science research project focusing on the evaluation of design principles for a software artifact aiming to improve the individual performance of requirements engineers in the task elicitation process. More specifically, we investigated the effect of semi-automatic task elicitation and the combined usage of imported and retrieved knowledge on task elicitation productivity. Based on a conceptualization of design principles, we developed an artifact implementing these principles and a corresponding research model measuring their effect on elicitation productivity. In the subsequent evaluation, we compared the productivity effects caused by the activation and de-activation of design principles in a laboratory and a field experiment. In the results of this evaluation, we find strong evidence that the design principles of our artifact are having a positive impact on task elicitation productivity.

In order to adequately interpret the implications of our findings, the following limitations of our study need to be considered. The laboratory experiment sessions were conducted with master IS students, not with experts, which limits the external validity of our findings. However, the replication of the experiment with a small group of experts showed evidence, that the same results pattern which has been observed in the laboratory setting can be expected in a field setting as well. A further limitation can be seen in the analysis of the experiment text data, which was based on manual document analysis. Although this analysis was thoroughly conducted, manual analysis is error-prone and can reduce reliability. However, the fact that results were analyzed by two researchers independently and with a high inter-rater reliability (98.97% in the documents which were coded twice) gives us confidence that this did not affect our findings.

There are many possible extensions to this work. The level of initial precision of automatically generated task elicitations is fixed at a relatively high level in the experiment. It is interesting to see how the initial precision will influence participants' perception and productivity. By varying the initial precision, we can examine to which extend semi-automatic task elicitation will still lead to improved recall and not negatively influence participants' resulting precision. Furthermore, a replication of our study in a different domain could add interesting insights. In the experiment, we adopted "traveling" as the task domain, which is reusable to a wide range of applications. When the domain becomes highly specific and dynamic, domain-specific knowledge becomes scarce and cannot easily be acquired and imported into the TES. In this case, the TES might be less useful since many task elements need to be manually established and might not be reused in further task elicitation. It is interesting to explore across multiple task domains which factors influence the perceived usefulness of the deployment of the design principles. Similarly, since we adopted a commonly understandable domain in the experiment, it is not observable how different levels of domain knowledge might influence participants' perception and productivity. Future research could use a more sophisticated domain and differentiate participants according to their domain knowledge, specifically examining the moderating effects of participants' domain knowledge on the relationships between design principles and task elicitation productivity.

From a theoretical perspective, our study provides the following contributions. First, it extends the design theory body of knowledge by the exploration and evaluation of design principles for TES. The setup of both the conceptualization and evaluation centered on the abstraction of specific design features to generic design principles. We are confident that this abstraction allows us to generalize our findings from the specific artifact to the class of TES. The prescriptive theoretical findings of our study may guide future research in designing efficient TES. Second, as described earlier, TES should improve requirements engineers' productivity in the corresponding process to provide an added value in comparison to manual task elicitation. Our study complements existing research on TES, investigating if this productivity improvement can actually be observed. Finally, beyond the topical aspects of our paper, we hope to also contribute to the ongoing methodological discussion in the design science context. Based on the conceptualization of design principles, we designed and conducted an experimental evaluation that allows quantifying the effects of each principle on a dependent variable. Going beyond an assessment of the artifact's overall effect this procedure allows precise inference from the evaluation back to the design process. We hope that this approach can inform other design researchers in the evaluation of their artifacts and the underlying design principles.

From a practical point of view, our study addresses practitioners' call to support requirements elicitation in general and task elicitation in specific through an IS-based automation. It thus helps to make the process less time-consuming, error-prone, and monotonous. Furthermore, our study helps software vendors to improve their Requirements Engineering software packages with regard to automated requirements and task elicitation capabilities. While support for manual requirements elicitation has been incorporated to selected commercial software packages (e.g., IBM Rational Doors), automated elicitation support is still scarce. The design features that were mapped to design principles in the conceptualization part of our study may inspire future commercial implementations.

References

Alba, J. 1983. “The effects of product knowledge on the comprehension, retention, and evaluation of product information,” Advances in consumer research (10), pp. 577-580.

Alkhader , Y., Hudaib, A., and Hammo, B. 2006. “Experimenting With Extracting Software Requirements Using NLP Approach,” in Proceedings of International Conference on Information and Automation, pp. 349-354.

Ambriola, V., and Gervasi, V. 2006. “On the Systematic Analysis of Natural Language Requirements with CIRCE,” Automated Software Engineering (13:1), pp. 107-167.

Appan, R., and Browne, G. J. 2012. “The Impact of Analyst-Induced Misinformation on the Requirements Elicitation Process,” MIS Quarterly (36:1), pp. 85-106.

Atkinson, C., and Kuhne, T. 2003. “Model-driven development: a metamodeling foundation,” IEEE Software (20:5), pp. 36-41.

Baskerville, R., and Pries-Heje, J. 2010. “Explanatory Design Theory,” Business & Information Systems Engineering (2:5), pp. 271-282.

Berry, D., Gacitua, R., Sawyer, P., and Tjong, S. F. 2012. “The Case for Dumb Requirements Engineering Tools,” in Requirements Engineering: Foundation for Software Quality, pp. 211-217.

Brasser, M., and Vander Linden, K. 2002. “Automatically eliciting task models from written task narratives,” in Proceedings of the 4th International Conference on Computer-Aided Design of User Interfaces, pp. 1-6.

Buchholz, E., Cyriaks, H., Dusterhoft, A., Mehlan, H., and Thalheim, B. 1995. “Applying a Natural Language Dialogue Tool for Designing Databases,” in Proceedings of the First International Workshop on Applications of Natural Language to Databases, pp. 1-16.

Casamayor, A., Godoy, D., and Campo, M. 2010. “Identification of non-functional requirements in textual specifications: A semi-supervised learning approach,” Information and Software Technology (52:4), pp. 436-445.

Cleland-Huang, J., Settimi, R., Zou, X., and Solc, P. 2007. “Automated classification of non-functional requirements,” Requirements Engineering (12:2), pp. 103-120.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences, (2nd ed,)Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Diehl, M., and Stroebe, W. 1991. “Productivity loss in idea-generating groups: Tracking down the blocking effect,” Journal of personality and social psychology (61:3) American Psychological Association, pp. 392-403.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. 2007. “G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences.,” Behavior research methods (39:2), pp. 175–91.

Gacitua, R., Sawyer, P., and Gervasi, V. 2011. “Relevance-based abstraction identification: technique and evaluation,” Requirements Engineering (16:3), pp. 251-265.

Gallupe, R. B., and McKeen, J. D. 1990. “Enhancing Computer-Mediated Communication: An experimental investigation into the use of a Group Decision Support System for face-to-face versus remote meetings,” Information & Management (18:1), pp. 1-13.

Goldin, L., and Berry, D. M. 1997. “AbstFinder, A Prototype Natural Language Text Abstraction Finder for Use in Requirements Elicitation,” Automated Software Engineering (4:4), pp. 375-412.

Gorschek, T., and Davis, A. M. 2008. “Requirements engineering: In search of the dependent variables,” Information and Software Technology (50:1-2), pp. 67-75.

Gregor, S., and Jones, D. 2007. “The Anatomy of a Design Theory,” Journal of the Association for Information Systems (8:5), pp. 312-335.

Hart, A. 2004. 801 Action Verbs For Communicators: Position Yourself First With Action Verbs For Journalists, Speakers, Educators, Students, Resume-writers, Editors &Travelers, Lincoln: iUniverse, Inc.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. “Design Science in Information Systems Research,” MIS Quarterly (28:1), pp. 75-105.

Hevner, A., and Chatterjee, S. 2010. Design Research in Information Systems, Information Systems Research (Vol. 22) Boston, MA: Springer US.

Hickey, A. M., and Davis, A. M. 2004. “A unified model of requirements elicitation,” Journal of Management Information Systems (20:4), pp. 65-84.

Hill, T., and Lewicki, P. 2007. STATISTICS: Methods and Applications, Tulsa, OK: StatSoft.

Huberty, C. J., and Morris, J. D. 1989. "Multivariate analysis versus multiple univariate analyses.," Psychological Bulletin (105:2), pp. 302–308.

Huffman Hayes, J., Dekhtyar, A., and Sundaram, S. 2005. "Text Mining for Software Engineering: How Analyst Feedback Impacts Final Results," ACM SIGSOFT Software Engineering Notes (30:4), pp. 1-5.

Jämsä-Jounela, S.-L. 2007. "Future trends in process automation," Annual Reviews in Control (31:2), pp. 211-220.

Kaiya, H., and Saeki, M. 2006. "Using Domain Ontology as Domain Knowledge for Requirements Elicitation," In Proceedings of the 14th IEEE International Requirements Engineering Conference (RE'06)Ieee, pp. 189-198.

Kincaid, J. P., Fishburn, R. P., Rogers, R. L., and Chissom, B. S. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel Research Branch Report 8-75 Millington, Tennessee: National Technical Information Service.

Kiyavitskaya, N., and Zannone, N. 2008. "Requirements model generation to support requirements elicitation: the Secure Tropos experience," Automated Software Engineering (15:2), pp. 149-173.

Kof, L. 2004. "Natural Language Processing for Requirements Engineering: Applicability to Large Requirements Documents," in Proceedings of the 19th International Conference on Automated Software Engineering.

Kuechler, B., and Vaishnavi, V. 2008. "On theory development in design science research: anatomy of a research project," European Journal of Information Systems (17:5), pp. 489–504.

Lemaigre, C., García, J. G., and Vanderdonckt, J. 2008. "Interface Model Elicitation from Textual Scenarios," in Proceedings of the Human-Computer Interaction Symposium (Vol. 272), pp. 53-66.

Markus, M. L., Majchrzak, A., and Gasser, L. 2002. "A design theory for systems that support emergent knowledge processes," MIS Quarterly (26:3), pp. 179-212.

Maynard, D. C., and Hakel, M. D. 1997. "Effects of objective and subjective task complexity on performance," Human Performance (10:4), pp. 303-330.

Meth, H., Maedche, A., and Einoeder, M. 2012. "Exploring design principles of task elicitation systems for unrestricted natural language documents," Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems - EICS '12. New York, New York, USA: ACM Press, pp. 205 - 210.

Mich, L., Franch, M., and Novi Inverardi, P. L. 2004. "Market research for requirements analysis using linguistic tools," Requirements Engineering (9:1), pp. 40-56.

Miller, G. A. 1956. "The magical number seven, plus or minus two: Some limits on our capacity for processing information," The Psychological Review (63:2), pp. 81 - 97.

Neill, C. J., and Laplante, P. A. 2003. "Requirements Engineering: The State of the Practice," IEEE Software (20:6), pp. 40-45.

Nunnally, J. C., and Bernstein, I. 1994. Psychometric Theory, (3rd ed,) New York: McGraw-Hill.

Paterno, F. 2002. "Task Models in Interactive Software Systems," in Handbook of Software Engineering and Knowledge Engineering Vol 1 Fundamentals, S. K. Chang (ed.), (Vol. 1)World Scientific, pp. 1-19.

Pedhazur, E., and Schmelkin, L. 1991. Measurement, design, and analysis: An integrated approach, Lawrence Erlbaum Associates, Inc.

Pohl, K. 2010. Requirements Engineering: Fundamentals, Principles, and Techniques, Springer.

Ratchev, S. M., Urwin, E., Muller, D., Pawar, K. S., and Moulek, I. 2003. "Knowledge based requirement engineering for one-of-a-kind complex systems," Knowledge Based Systems (16:1), pp. 1-5.

Rayson, P., Garside, R., and Sawyer, P. 2000. "Assisting requirements engineering with semantic document analysis," in Proceedings of the RIAO, pp. 1363-1371.

Robertson, S., and Robertson, J. 2006. Mastering the Requirements Process, Pearson Education.

Rolland, C., and Proix, C. 1992. "A natural language approach for requirements engineering," in Advanced Information Systems Engineering, pp. 257-277.

SAP. 2012. "Travel Management (FI-TV)," SAP AG, (http://help.sap.com/printdocu/core/print46c/en/data/pdf/FITVPLAN/FITVGENERIC.pdf; accessed February 1, 2012)

Seltman, H. J. 2012. "Within-Subjects Designs," In Experimental Design and Analysis, pp. 339–356.

Sharp, H., Rogers, Y., and Preece, J. 2007. Interaction design: beyond human-computer interaction, Chichester: John Willey & Sons Ltd.

Shibaoka, M., Kaiya, H., and Saeki, M. 2007. "GOORE: Goal-Oriented and Ontology Driven Requirements Elicitation Method," In Advances in Conceptual Modeling – Foundations and ApplicationsBerlin / Heidelberg: Springer, pp. 225-234.

Staab, S., Studer, R., Schnurr, H. P., and Sure, Y. 2001. "Knowledge processes and ontologies," IEEE Intelligent Systems (16:1), pp. 26-34.

Tabachnick, B. G., and Fidell, L. S. 2007. Using Multivariate Statistics, (S. Hartman, ed.) (5th ed,)Pearson Education, Inc.

Tam, R. C.-man, Maulsby, D., and Puerta, A. R. 1998. "U-TEL: A Tool for Eliciting User Task Models from Domain Experts," in Proceedings of the 3rd international conference on Intelligent user interfaces, pp. 77-80.

Tamir, D., Marcos, S., and Mueller, C. J. 2008. "An Effort and Time Based Measure of Usability," in Proceedings of the 6th international workshop on Software quality, pp. 47-52.

Tichy, W. F., and Koerner, S. J. 2010. "Text to Software Developing Tools to Close the Gaps in Software Engineering Categories and Subject Descriptors," in Proceedings of the FSE/SDP workshop on Future of software engineering research - FoSER '10, pp. 379-383.

Vasey, M., and Thayer, J. 1987. "The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution," Psychophysiology (24:4), pp. 479–486.

Vlas, R., and Robinson, W. N. 2011. "A Rule-Based Natural Language Technique for Requirements Discovery and Classification in Open-Source Software Development Projects Related research," in Proceedings of the 44th Hawaii International Conference on System Sciences, pp. 1-10.

Walls, J. G., and Sawy, O. A. E. 1992. "Building an Information System Design Theory for Vigilant EIS," Information Systems Research (3:1), pp. 36-60.

Wang, W., and Benbasat, I. 2007. "Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs," Journal of Management Information Systems (23:4), pp. 217–246.

Wilson, W. M., Rosenberg, L. H., and Hyatt, L. E. 1997. "Automated Analysis of Requirement Specifications," in Proceedings of the 19th International Conference on Software engineering, pp. 161-171.