Empir Software Eng (2017) 22:2001–2049
DOI 10.1007/s10664-016-9475-z

Group versus individual use of power-only EPMcreate as a creativity enhancement technique for requirements elicitation

Victoria Sakhnini[1] · Luisa Mich[2] · Daniel M. Berry[1]

Published online: 7 January 2017

Abstract Creativity is often needed in requirements elicitation, i.e., generating ideas for requirements, and therefore, techniques to enhance creativity are believed to be useful. How does the size of a group using the Power-Only EPMcreate (POEPMcreate) creativity enhancement technique affect the group's and each member of the group's effectiveness in generating requirement ideas? This paper describes an experiment in which individuals and two-person and four-person groups used POEPMcreate to generate ideas for requirements for enhancing a high school's public Web site. The data of this experiment combined with the data of two previous experiments involving two-person and four-person groups using POEPMcreate show that, similar to what has been observed for brainstorming, the size of a group using POEPMcreate does affect the number of raw and new requirement ideas generated by the group and by the average member of the group. The data allow concluding that a two-person group using POEPMcreate generates more raw and new requirement ideas, both per group and per group member or individual, than does a four-person group and than does an individual. This conclusion is partially corroborated by qualitative data gathered from a survey of professional business or requirements analysts about group sizes and creativity enhancement techniques.

Communicated by: Tony Gorschek

See the section titled "Compliance with Ethical Standards", just before the references, for a statement about previous publication of parts of this paper's contents.

✉ Victoria Sakhnini
vsakhnin@uwaterloo.ca

Luisa Mich
luisa.mich@unitn.it

Daniel M. Berry
dberry@uwaterloo.ca

[1] Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, N2L 3G1 Canada

[2] Department of Industrial Engineering, University of Trento, 38122 Trento, Italy

1 Introduction

Creativity plays an important role throughout the development of software-intensive computer-based systems (CBSs) (Glass 1995; Glass and DeMarco 2006). Many have stressed the point that requirements engineering (RE), in general, and requirements elicitation, in specific, are fundamentally creative activities (Gause and Weinberg 1989, 1990; Maiden and Gizikis 2001; Robertson and Maiden 2002; Maiden et al. 2004), even to the point of declaring that requirements may need to be invented (Goguen 1993, 1994; Robertson 2002; Maiden et al. 2006), particularly when the very understanding of the problem to be solved is among what must be invented, i.e., the problem is wicked (Rittel and Webber 1973). Thus, creativity plays an important role in RE for CBSs.

Accordingly, creativity in RE is manifested mainly during the generation of ideas for requirements, i.e., mainly during requirements elicitation. Indeed, most of the literature about creativity in RE cited in Section 12 is focused on requirements elicitation. Therefore, any requirements elicitation technique is likely to be improved by creativity, and, conversely, any way to be more creative should be helpful in requirements elicitation.

Creativity is very hard to define because it plays roles in so many fields, technical as well as non-technical, and each field has its own definition (Runco 2007). Nevertheless, creativity, in general, is the ability of an individual or a group to think of new and useful ideas (Runco 2007), while solving a problem (Rickards 1974; Leigh 1983).

Many creativity enhancement techniques (CETs), e.g., brainstorming (Osborn 1953); Six Thinking Hats (de Bono 1985); Creative Problem Solving (Parnes 1992); and the Creative Pause Technique (de Bono 1993), have been developed to help people be more creative. Some of these techniques have been applied to RE (Aurum and Martin 1998; Maiden et al. 2004; Lemos et al. 2012; Berntsson Svensson et al. 2015), generally, as mentioned, as part of requirements elicitation. Applying a CET to requirements elicitation may end up being indistinguishable from a requirements elicitation method.

See Section 12, about related work, for deeper discussions about creativity and creativity in RE.

1.1 The Purpose of this Paper

The present paper is the third in a series of papers (Mich et al. 2005; Sakhnini et al. 2012) that empirically explore the use and the effectiveness of the *EPMcreate (Elementary Pragmatic Model Creative Requirements Engineering [A] TEchnique)* CET and one optimization to help in generating requirement ideas. The optimization is called *Power-Only EPMcreate (POEPMcreate)* and has only one quarter of the steps that EPMcreate has.

The feasibility of applying EPMcreate and POEPMcreate to help idea generation in requirements elicitation was established by earlier experiments (Mich et al. 2005, 2006; Sakhnini et al. 2012). The results of these experiments confirmed that:

1. EPMcreate helps generate more ideas and more new ideas for requirements than does brainstorming.
2. POEPMcreate helps generate more ideas and more new ideas for requirements than does each of EPMcreate and brainstorming.

The facts that POEPMcreate is more effective than EPMcreate in fostering requirement idea generation and that POEPMcreate has fewer steps than EPMcreate allows us to use POEPMcreate exclusively and to focus our research attention on POEPMcreate.

In each experiment, the size of groups doing the different CETs was held constant at 2 or 4 when comparing groups using two different CETs. Therefore, group size could be discounted as a cause of the observed differences in the quantity and quality of the requirement ideas generated. However, an informal comparison for each EPMcreate CET, of the quantities of the requirement ideas generated for the different group sizes, yielded some surprises.

While on average, an EPMcreate group of size 4 generated more ideas than an EPMcreate group of size 2, the number of ideas generated by a group of size 4 was less than double the number of ideas by a group of size 2. Thus, on average, a member of an EPMcreate group of size 2 generated more ideas than a member of an EPMcreate group of size 4! The surprise was even more pronounced for POEPMcreate groups. On average, a POEPMcreate group of size 4 generated just fewer ideas than a POEPMcreate group of size 2, and on average, a member of a POEPMcreate group of size 2 generated more ideas than a member of a POEPMcreate group of size 4.

Several researchers (Furnham and Yazdanpanahi 1958; Taylor et al. 1958; Dennis and Valacich 1993; Aurum and Martin 1998; Isaksen and Gaulin 2005; Dornburg et al. 2008; Ocker 2010; Kohn and Smith 2011) have noticed similar and stronger phenomena for brainstorming, that smaller groups are more effective per person than larger groups, and that beyond that, individuals are the most effective. The similarity between brainstorming and EPMcreate and POEPMcreate for at least group sizes 2 and 4 naturally raises the question of whether the similarity extends to individuals. That is, are individuals more effective than groups when using EPMcreate and POEPMcreate?

Given that we are focusing on POEPMcreate as more efficient than EPMcreate, it is time to explore empirically the effect of group size on the effectiveness of POEPMcreate and to consider POEPMcreate conducted by individuals. We thus posed the following research question:

> In POEPMcreate, how does the number of members of an elicitation group affect the quantity and quality of requirement ideas generated by the group and by each member?

This question was answered by conducting an experiment in the context of eliciting requirements for a high school's Web site, the same site that was used in two previous experiments (Sakhnini et al. 2012) and then combining the data from the three experiments. The results of the combined experiments showed, in essential confirmation of the informal observations, that a POEPMcreate group of size 2 is more effective at idea generation, both per group and per person, than is a group of size 4 or an individual. However, to our surprise, a whole group of size 4 is more effective than an individual, but a person in a group of size 4 is less effective than an individual. Thus, there is a difference between brainstorming and at least POEPMcreate, that is explored in the last paragraphs of Section 10.

1.2 The Rest of this Paper

In the rest of this paper, Section 2 describes the EPMcreate technique and the POEPMcreate optimization. Section 3 describes the general experimental design, including its hypotheses and its steps. Section 4 gives the particulars that distinguish the three specific instantiations of the general experimental design. Section 5 gives the data gathered from all three experiments. Section 6 discusses problems with the gathered data, including whether it is legitimate to combine the data from the three experiments into one analysis. Section 7 explains how multivariate regressions are used for the present analysis. Section 8 gives the results of the regressions, and determines whether the hypotheses are supported. Section 9 discusses threats to the validity of the conclusions and how they are or can be mitigated. Section 10 speculates about optimal group sizes for POEPMcreate and proposes a general theory applicable to any CET. Section 11 describes the results of a survey conducted to obtain qualitative triangulation for the results and speculation. Section 12 summarizes the related work, and Section 13 concludes the paper.

2 The POEPMcreate Technique

Since POEPMcreate is an optimization of EPMcreate, describing POEPMcreate requires first describing EPMcreate. The explanation of EPMcreate given here is abbreviated to what is necessary to understand this paper. A fuller description of EPMcreate can be found in our earlier publications (Mich et al. 2005; Sakhnini et al. 2012).

2.1 EPMcreate

EPMcreate supports idea generation by focusing the search for ideas on only one logical combination of two stakeholders' viewpoints at a time. Sixteen such combinations are possible, each corresponding to one of the Boolean functions, fi for $0 \leq i \leq 15$, of two variables. These functions are given in Table 1. In this table, "Vn" means "Stakeholder n's Viewpoint" and "fi" means "boolean function i". The bits in each column fi form the binary encoding for i when they are read from top to bottom. These functions are $f0 = 0$, $f1 = V1 \wedge V2$, $f2 = V1 \wedge \neg V2$, $f3 = V1$, $f4 = \neg V1 \wedge V2$, $f5 = V2, \ldots, f8 = \neg V1 \wedge \neg V2, \ldots$, and $f15 = 1$. These sixteen functions are used to specify how the viewpoints of stakeholders SH1 and SH2 are combined in the sixteen steps of the EPMcreate procedure described in the next subsection. The interpretations of *some* of these functions in terms of combining the viewpoints of stakeholders SH1 and SH2 are:

$f0 =$ 0, represents the analyst's looking for ideas that disagree with everything, independently of both SH1's viewpoint and SH2's viewpoint, i.e., looking for nothing.

$f1 =$ SH1 ∧ SH2, represents the analyst's looking for ideas that agree with SH1's viewpoint and with SH2's viewpoint.

$f2 =$ SH1 ∧ ¬SH2, represents the analyst's looking for ideas that agree with SH1's viewpoint but disagree with SH2's viewpoint.

$f3 =$ SH1, represents the analyst's looking for ideas that agree with SH1's viewpoint completely, independently of SH2's viewpoint.

$f4 =$ ¬SH1 ∧ SH2, represents the analyst's looking for ideas that agree with SH2's viewpoint but disagree with SH1's viewpoint.

Table 1 Table of the 16 combinations of two viewpoints

$V1$	$V2$	$f0$	$f1$	$f2$	$f3$	$f4$	$f5$	$f6$	$f7$	$f8$	$f9$	$f10$	$f11$	$f12$	$f13$	$f14$	$f15$
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

$f5 =$ SH2, represents the analyst's looking for ideas that agree with SH2's viewpoint completely, independently of SH1's viewpoint.

$f8 =$ ¬SH1 ∧ ¬SH2, represents the analyst's looking for ideas that disagree with SH1's viewpoint and with SH2's viewpoint.

$f10 =$ ¬SH2, represents the analyst's looking for ideas that disagree with SH2's viewpoint completely, independently of SH1's viewpoint.

$f15 =$ 1, represents the analyst's looking for ideas that agree with everything, independently of both SH1's viewpoint and SH2's viewpoint.

For example, the principal stakeholders for a school's Web site, such as used in the experiments, are students, parents, and teachers. Suppose SH1 is the students and SH2 is the parents. Then, an example of

$f1 =$ SH1 ∧ SH2, a requirement idea that agrees with SH1's viewpoint and with SH2's viewpoint, is making vacation days for a full academic year visible to all, starting six months in advance of the start of the academic year;

$f2 =$ SH1 ∧ ¬SH2, a requirement idea that agrees with SH1's viewpoint but disagrees with SH2's viewpoint, is making future homework assignments visible to only students;

$f4 =$ ¬SH1 ∧ SH2, a requirement idea that agrees with SH2's viewpoint but disagrees with SH1's viewpoint, is informing a parent of his or her child's absence from school; and

$f8 =$ ¬SH1 ∧ ¬SH2, a requirement idea that disagrees with SH1's viewpoint and with SH2's viewpoint, is making evaluations of teachers visible to only teachers.

If there are more than two types of stakeholders, the technique can be applied several times, for each relevant pair of stakeholder types, up to $\binom{n}{2}$ times for n stakeholders.

2.2 EPMcreate in Practice

EPMcreate can be applied whenever ideas need to be generated, e.g., at any time that one might apply a CET, such as brainstorming. When a lead requirements analyst (leader) adopts EPMcreate as the CET for eliciting requirements for a CBS under consideration, she first chooses two kinds of stakeholders, SH1 and SH2, usually users of the CBS with different roles, as those whose viewpoints will be used to drive the application of EPMcreate. She may ask the CBS's analysts for assistance in this choice. She then convenes a group of these analysts. Figure 1 contains a diagram that the leader shows the analysts as part of her explanation of EPMcreate. In this diagram, the two ellipses represent two different stakeholders' viewpoints. Thus, for example, the intersection region represents the stakeholders' shared viewpoints.

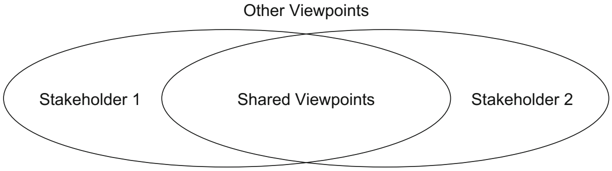


Fig. 1 Venn Diagram of Two Stakeholders' Viewpoints

While showing the diagram of Fig. 1, the leader tells all convened analysts, Today, we are going to generate requirement ideas for the CBS S in 16 idea generation steps. In all the steps, you will be pretending to think from the viewpoints of two particular stakeholders of S, SH1 and SH2.

- In Step 0, you will blank out your minds ($f0 = 0$).
- In Step 1, you will try to come up with ideas for problem solutions that are needed by both **SH1 and SH2** ($f1 = V1 \wedge V2$).
- In Step 2, you will try to come up with ideas for problem solutions that are needed by **SH1 but not** by **SH2** ($f2 = V1 \wedge \neg V2$).
- In Step 3, you will try to come up with ideas for problem solutions that are needed by **SH1** without concern as to whether they are needed by SH2 ($f3 = V1$).
- In Step 4, you will try to come up with ideas for problem solutions that are needed by **SH2 but not** by **SH1** ($f4 = \neg V1 \wedge V2$).
- In Step 5, you will try to come up with ideas for problem solutions that are needed by **SH2** without concern as to whether they are needed by SH1 ($f5 = V2$).

  ...
- In Step 8, you will try to come up with ideas for problem solutions that are needed **neither** by **SH1 nor** by **SH2**, but are needed by other stakeholders ($f8 = \neg V1 \wedge \neg V2$).

  ...
- In Step 10, you will try to come up with ideas for problem solutions that are not needed by **SH2** without concern as to whether they are needed by SH1.

  ...
- In Step 15, you will try to come up with ideas for problem solutions without concern as to whether they are needed by either SH1 or SH2 ($f15 = 1$).

Note that each Step i is based on the Boolean function fi.

In the event that the leader believes that more than two stakeholders' viewpoints should be considered, she will convene more EPMcreate sessions, one for each pair of stakeholder viewpoints she believes to be useful. Her experience tells her how to identify subsets of stakeholders and stakeholder pairings that will yield the most new ideas for the fewest pairs. For example, marketing suggests taking into account users' profiles for creating market segments. Each such profile usually has different requirements.

The choice of the stakeholders is straightforward for some types of systems, e.g., for an e-learning platform, the clear stakeholders are the students and the teachers. In other cases, the choice could be strategic, e.g. for a tourism destination Web site, the chosen viewpoints could correspond to targeted market segments. On the other hand, when the requirements for a stakeholder are already known or are irrelevant, e.g., for an e-learning platform, if the requirements for the owning university are already known, it is not necessary to chose this stakeholder for any session of EPMcreate.

Note that EPMcreate, like brainstorming, has as its goal, generation of as many ideas as possible, as input to a later stage in which ideas are discussed, pruned, enhanced, combined, etc., according to an almost standard creativity process model (Zhou 2016), to arrive at a final set of quality ideas. While it is hoped that all the generated ideas will be innovative and useful, no EPMcreate group is to slow down to filter out useless, non-innovative, incorrect ideas. Besides, it has been observed that, at least in brainstorming, quality follows quantity (Osborn 1953).

2.3 Power only EPMcreate

The optimization of EPMcreate under study in this paper is called "Power-Only EPMcreate (POEPMcreate)", because it does only the four steps, described above, whose names contain the powers of two, namely Step 1, Step 2, Step 4, and Step 8.

This optimization, which does only four of the sixteen original steps, was theorized, and later demonstrated (Sakhnini et al. 2012), to be more effective than the full EPMcreate, because as illustrated by Fig. 2, the Boolean function of each of the power-of-two steps corresponds to exactly one of the four regions of Fig. 1. Thus, the four power-of-two steps are sufficient to cover the entire space of potential ideas, and the other twelve steps just repeat the coverage. This coverage happens because these four regions are the four atoms of the 16-element free Boolean algebra that is generated from the two stakeholder viewpoints (Preparata and Yeh 1973; Givant and Halmos 2009).

3 Experimental Design

As mentioned, the effectiveness of POEPMcreate as a CET and as an *improvement* over EPMcreate was established by two experiments, (Sakhnini et al. 2012).

The present paper describes

1. a new experiment, Experiment 3, which follows exactly the design used in Experiments 1 and 2, and
2. an analysis of the combined data of Experiments 1, 2, and 3

to answer the research question mentioned in Section 1.1, which is to determine how the number of members of an elicitation group, using POEPMcreate as a CET, affects the quantity and quality of requirement ideas generated by the group and by each member.

The purpose of Experiment 3 was to add groups and individuals performing CETs in order to have collections of enough groups of different sizes and of enough individuals that an analysis to answer the research question could be conducted. We knew from the conduct of Experiments 1 and 2 that recruiting subjects was difficult, and therefore, our most precious commodity was subjects. We had already proved that POEPMcreate was enough of an improvement over EPMcreate by all measures (Sakhnini et al. 2012), that anyone considering using an EPMcreate-like CET would naturally use POEPMcreate from

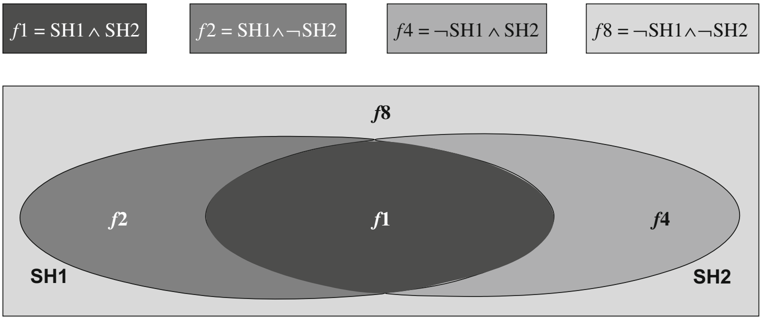


Fig. 2 The Four Steps of the Optimization and the Four Regions of the Venn Diagram

the beginning. Therefore, we decided to focus Experiment 3 on POEPMcreate, having all of its subjects be in groups or be individuals doing POEPMcreate.

The rest of this section describes the experimental design that was used in all three experiments. This description ignores the fact that the first two experiments tested both EPMcreate and POEPMcreate and talks as if the only CET tested were POEPMcreate. The reader that is interested in the results of the analysis for the EPMcreate data should consult a technical report by the same authors (Sakhnini et al. 2016).

Section 4, following this one, describes the details that are particular to specific experiments. Note that all decisions about the experimental design were made during the conduct of Experiment 1, based on what had been learned in the earliest experiment conducted by us (Mich et al. 2005). To allow combining the data of Experiments 1, 2, and 3, it was necessary to maintain these decisions in the conduct of Experiments 2 and 3.

3.1 Research Question and Hypotheses

Recall that the research question to be answered by the research described in this paper was:

> In POEPMcreate, how does the number of members of an elicitation group affect the quantity and quality of requirement ideas generated by the group and by each member?

As indicated in Section 1.1, we had an idea of what the answer would be. However, there were good reasons to believe that smaller groups are more effective, and there were good reasons to believe that larger groups are more effective. Supporting that smaller groups are more effective:

- with fewer members in a group, each member has more opportunities to speak up with ideas, and
- with fewer members in a group, the group's management overhead decreases, leaving more time for each member of the group to generate ideas.

Supporting that larger groups are more effective:

- with more members in a group, the group has more brain power with which to generate ideas, and
- with more members in a group, the group has more of the synergy that is supposed to increase the whole group's idea generation (Osborn 1953; von Bertalanaffy 1976).

As for individuals conducting EPMcreate, Mich et al. (2010) report support for a hypothesis that EPMcreate can be used effectively by individuals, as well as groups, to help generate requirement ideas. While an individual appears to generate fewer ideas in a time span than a group of four, there were not enough data to correlate group size with the number of ideas generated.

Therefore, we thought it is best to test only null hypotheses that address the research question:

H1 In POEPMcreate, the number of members of an elicitation group has no effect on the quantity and quality of the requirement ideas generated by the group.

H2 In POEPMcreate, the number of members of an elicitation group has no effect on the quantity and quality of the requirement ideas generated on average by each member of the group.

3.2 Context of Experiments

All groups participated in the experiments for the same amount of time. Each group was to generate, using its CET, EPMcreate or POEPMcreate, ideas for requirements for an improved version of one existing Web site, that of a Canadian high school with information directed to students, parents, teachers, and administrators (Administrator 2010). This site was chosen for its cost-free, password-free availability, lack of intellectual property restrictions, and the fact that as educators, the authors could be considered domain experts. We decided that the two types of stakeholders whose viewpoints would be adopted by the groups were students and parents.

3.3 Measuring the Effectiveness of a CET

The effectiveness of an individual or a group using a CET is normally measured by two numbers about the requirement ideas generated by the individual or group when using the CET (Dean et al. 2006; Jones et al. 2008; Nguyen and Shanks 2009),

1. the quantity of the generated requirement ideas, i.e., the raw number of requirement ideas generated, and
2. the quality of the generated requirement ideas, i.e, the number of high quality requirement ideas generated.

Counting raw requirement ideas is straightforward. The subjects were instructed to write each idea on one line in Microsoft Word. About 90 % of these ideas are in the form of one complete sentence or a bullet item phrase describing a feature. Of the remaining 10 % of these ideas, about 95 % are at most two sentences. In other words, almost all the requirement ideas written down express what is called an atomic requirement (Salzer and Levin 2004). Finally, counting raw ideas is a valid measure of the effectiveness of a CET, since, as mentioned in Section 2.2, an EPMcreate session's goal is to generate as many ideas as possible.

Measuring the quality of a requirement idea is not so straightforward, as this measure depends on the definition of creativity being used. The main problem is that there is no universally agreed-upon definition of creativity. All definitions agree that a creative idea is a new or novel idea (Maiden et al. 2004; Dean et al. 2006; Jones et al. 2008; Nguyen and Shanks 2009; Conboy et al. 2009). Beyond newness, there is no universal agreement. Many definitions stress also usefulness (Jones et al. 2008; Nguyen and Shanks 2009). Other characteristics that have been mentioned include applicability, effectiveness, implementability, non-obviousness, originality, relevance, realizability, specificity, thoroughness, and workability (Mich et al. 2005; Dean et al. 2006).

To be safe, we decided to stick to the newness common denominator and to use only newness of a requirement idea as the measure of the requirement idea's quality. We are not the first to do so. For example, in an empirical evaluation of a method to invent creative requirement ideas, Zachos and Maiden used only novelty, as measured by dissimilarity to existing features, as the measure of a requirement idea's quality (Zachos and Maiden 2008). Not just in RE, measuring quality of ideas is subject to debate. For example, Briggs et al. (1997) observe that "Evaluating idea quality can be a grueling, expensive, and uncertain task. Some studies do not address idea quality [*citations in the original*], while others argue that the existing empirical evidence precludes the necessity for going to the expense and effort of measuring idea quality." Even as Briggs et al. conclude that "researchers must continue to measure the effects of their brainstorming treatments on idea quality" because there are factors other than quantity that affect the quality of ideas, they admit that “the empirical record [on the subject] is equivocal”.

Finally, when we were conducting Experiments 1 and 2, we did classify ideas for both newness and realizability. However, we noticed that the strongest correlation between the two experts’ classifications was in the newness classification. So, we ended up using only newness as the measure of an idea’s quality for Experiments 1 and 2 (Sakhnini et al. 2012). Combining the results of multiple experiments requires following this decision in all experiments.

A requirement idea is classified as “new” if and only if it describes a feature that is not already in the Web site for which requirement ideas are being generated. The school Web site used in the experiments publishes the teachers’ e-mail addresses. Thus, the idea “Allow sending e-mail to a teacher.” is not a new idea, since anyone can use the published e-mail address to send e-mail to a teacher. However, “Hide the teachers’ e-mail addresses and provide an in-site way to send secure messages to teachers.” would be a new idea.

To evaluate the newness of the requirement ideas in any experiment, each of two domain experts, namely the first and third authors of this paper, independently classified each idea as to whether or not it is new. In order to reduce the chances that the authors’ desired results might affect the newness evaluation, we merged the requirement ideas generated by all the groups into one file. We then sorted the ideas alphabetically to produce the list of ideas to be evaluated, making it impossible for any evaluator to see which group or individual, with its known group size (and known CET), generated any idea being evaluated. After each evaluator had assigned a newness to each idea, the assignments were copied to the original idea files, in order to be able to evaluate the newness of the requirement ideas of each group or individual separately. The actual ideas evaluated for the experiments and their evaluations are found in the package of experimental materials available for download in the folder named “IdeasFiles” (Sakhnini et al. 2016).

3.4 Refining Hypotheses into Subhypotheses

The two hypotheses H1 and H2 may be refined into four subhypotheses, each one taking the number of either raw or new requirement ideas produced either by the whole group or on average by a member of the group.[1] The four subhypotheses are, therefore:

H1:

HPTR: The number of members of an elicitation group using **POEPMcreate** has no effect on the **total number of requirement ideas per group** of **raw** requirement ideas generated.

HPTN: The number of members of an elicitation group using **POEPMcreate** has no effect on the **total number of requirement ideas per group** of **new** requirement ideas generated.

[1] The general form of a subhypothesis is:

“The number of members of an elicitation group using $\left\{\begin{array}{l} E: \ \text{EPMcreate} \\ P: \ \text{POEPMcreate} \end{array}\right\}$ has no effect on the $\left\{\begin{array}{l} T: \ \text{total number of requirement ideas per group} \\ A: \ \text{average number of requirement ideas per group member} \end{array}\right\}$ of $\left\{\begin{array}{l} R: \ \text{raw} \\ N: \ \text{new} \end{array}\right\}$ requirement ideas generated.” The name of any subhypothesis is “H” followed by concatenation of the labels designating the choices made to construct the subhypothesis. Each label is the first letter of the phrase that it labels.

H2:

- **HPAR:** The number of members of an elicitation group using **POEPMcreate** has no effect on the **average number of requirement ideas per group member** of **raw** requirement ideas generated.
- **HPAN:** The number of members of an elicitation group using **POEPMcreate** has no effect on the **average number of requirement ideas per group member** of **new** requirement ideas generated.

3.5 Steps of an Experiment

To simplify the descriptions of the experiments and of the analysis of their data, an individual working alone to generate requirement ideas is called “a one-person group”.

Each experiment consisted of four steps. Steps 1 and 2 were done in one 50-minute meeting for each subject, and Steps 3 and 4 were done in several multi-group sessions with four-person, two-person, and one-person groups in attendance. The steps and their approximate times were:

Step 1: 20 minutes for each subject to read and sign an informed-consent form and to fill out a general information form that allowed us to know his or her background: The form included questions about his or her age, gender, native language, computer science (CS) courses, qualifications related to CS, employment history in CS, and knowledge of the CETs: brainstorming, EPMcreate, and POEPMcreate. For Experiments 2 and 3, we made sure before any subject began, that he or she had not been a subject in any previous experiment. We lost one subject in this checking.

Step 2: 30 minutes for each subject to take an adult version of Frank Williams’s Creativity Assessment Packet (Williams and Taylor 1966), hereinafter called the *Williams test* to measure the subject’s individual[2] creativity.

Step 3: 10 minutes for us to deliver to all groups an explanation about the experiment and POEPMcreate, the CET that they were to use. The explanation of POEPMcreate was basically a recitation of the second paragraph of Section 2.2 of this paper, but using only Function Steps 1, 2, 4, and 8, accompanied by showings of Figs. 1 and 2.

Step 4: 120 minutes for each group to carry out its requirements elicitation session using POEPMcreate. Each group was provided with two computers: one with which to access the Web site that the group was to improve, and the other with which to write the requirement ideas generated by the group. The typical one-person group used only one of the computers to which it had access.

The materials for conducting the experiment are available for downloading (Sakhnini et al. 2016).

3.6 Recruiting and Assigning Subjects into Balanced Groups

For each experiment, we recruited subjects from upper-division undergraduate and graduate students in the various software engineering programs at the University of Waterloo. In the

[2] The phrase “individual creativity” is a technical term from the creativity assessment field that means *natural, unassisted, original creativity of the individual* and not just individual as opposed to group creativity (Kaufman and Sternberg 2006).

recruiting advertisement, delivered verbally, electronically, or by poster, we offered each subject an honorarium of $20.00 (Canadian). Nevertheless, despite all of the advertising and recruiting we did, it was extremely difficult to convince people to be subjects, and we had to find ways to maximize the value of each subject that we did find. Altogether over the three experiments, for POEMPcreate, we had 45 subjects, 24 in four-person groups, 16 in two-person groups, and 5 in one-person groups.

The Williams test was administered to each subject to measure his or her individual creativity. The subjects' test scores were originally to be used to ensure that any observed differences in the numbers of requirement ideas were not due to differences in the individual creativity of the subjects. Instead, in order to avoid having to interpret specific scores during analysis, we used the subjects' Williams test scores to form groups that were a priori as balanced as possible by their members' computer science knowledge, work experience, and individual creativity scores.

To make it *even possible* to form groups, we ignored gender and age in creating the groups because it would have been very difficult to balance these factors while balancing the other factors. In any case, we did not believe that these factors are relevant, and even if they are, they are probably less relevant than the ones we did consider. As expected, none of the subjects had heard about any form of EPMcreate, even though all had heard about brainstorming.

4 Experiment-Specific Details

This section describes those details about the design and conduct of the experiments that are different in each experiment.

4.1 Focus of Experiment 3

Experiments 1 and 2 had addressed other research questions about EPMcreate and POEPMcreate using data from four two-person and two four-person groups for each CET. We had no data points for individuals' uses of these CETs in these experiments. For sure, Experiment 3 had to focus on individuals' use of these CETs. Experiment 2 had established that POEPMcreate is more effective and requires fewer steps than EPMcreate. So, we decided to focus the experimentation on POEPMcreate. To conserve the precious resource of volunteer subjects and to get the maximum bang from each subject buck, we decided to have all groups in Experiment 3 use POEPMcreate.

Table 2 Characteristics of POEPMcreate Groups of Experiment 1 and Their Subjects (Sakhnini et al. 2012)

Group	# of subjects per group	# Males	# Females	# native in English	# not native in English	# taken $\geq$ 10 CS courses	# taken 3–5 CS courses	# worked professionally	# not worked professionally	Average age	Average Williams test score
1	4	3	1	1	3	2	2	2	2	25.5	70.66
2	4	2	2	2	2	2	2	3	1	23.8	71.00

Table 3 Characteristics of POEPMcreate Groups of Experiment 2 and Their Subjects (Sakhnini et al. 2012)

Group	# of subjects per group	# Males	# Females	# native in English	# not native in English	# taken ≥ 10 CS courses	# taken 3–5 CS courses	# worked professionally	# not worked professionally	Average age	Average Williams test score
E	2	1	1	0	2	0	2	2	0	30.5	75.5
F	2	2	0	0	2	0	2	0	2	25	80.5
G	2	0	2	0	2	1	1	1	1	24	72.5
H	2	2	0	0	2	2	0	2	0	26	63

4.2 Conduct and Demographics of Experiments 1 and 2

Experiment 1 was conducted in 4 sessions during the third week of November 2009. Experiment 2 was conducted in 6 sessions during the second week of March 2010. The demographic properties of the POEPMcreate groups in Experiments 1 and 2, obtained from the data gathered during Steps 1 and 2 of those experiments, are shown in Tables 2 and 3 (distilled from the paper about Experiments 1 and 2 (Sakhnini et al. 2012)). In these and other tables, the value under "# worked professionally" for a group is the number of members of the group whose answers to the employment history question in Step 1 indicated that they had worked for CS-related jobs for pay, e.g., in cooperative education[3] terms.

The average Williams test scores for the two POEPMcreate groups in Experiment 1 were 70.66 and 71.00 out of a possible 100, and the average of the average Williams test score for the two POEPMcreate groups was 70.83. The average Williams test scores for the four POEPMcreate groups in Experiment 2 were in the range from 63 to 80.5, and the average of the average Williams test score for the four POEPMcreate groups was 72.875. There was no way to form groups with closer average Williams test scores without severely unbalancing them in other factors.

4.3 Conduct and Demographics of Experiment 3

Experiment 3 was conducted in two rounds. Its first round was conducted in one Step 4 session on 9 June 2010, and its second round was conducted in several sessions in October 2012.[4] In the first round of Experiment 3, only 15 students replied to the call for subjects, and of these, 13 ended up being subjects in the experiment. These 13 subjects were distributed into four two-person groups, G1–G4, and five one-person groups, G5–G9, as shown in the first nine lines of Table 4. For the second round of Experiment 3, 18 students replied

[3] In cooperative education at the University of Waterloo, each student works for pay, over his or her four years, one term per year, in an off-campus job, generally in his or her area of study. A CS or SE student typically works in a computing-related job, often in software development. In some cases, the student ends up getting a permanent job at one of his or her co-op employers.

[4] We had to wait at least a year and then until the Fall term between rounds to get a large enough crop of new potential subjects, a.k.a. new students, who had never participated in any of our experiments.

Table 4 Characteristics of Groups of Experiment 3 and Their Subjects

Group	# of subjects per group	# Males	# Females	# native in English	# not native in English	# taken ≥ 10 CS courses	# taken 3–5 CS courses	# worked profes-sionally	# not worked profession-ally	Average age	Average Williams test score
G1	2	2	0	1	1	2	0	2	0	27.5	60.5
G2	2	1	1	1	1	2	0	2	0	26.5	76.5
G3	2	2	0	1	1	2	0	1	1	32.5	78
G4	2	2	0	1	1	2	0	2	0	26.5	86.5
G5	1	0	1	0	1	1	0	1	0	41	68
G6	1	1	0	0	1	1	0	1	0	25	72
G7	1	1	0	0	1	1	0	1	0	33	73
G8	1	0	1	0	1	1	0	0	1	21	79
G9	1	1	0	1	0	1	0	1	0	26	85
G10	4	4	0	2	2	2	2	2	2	22.5	73
G11	4	4	0	2	2	2	1	2	2	17.25	76.75
G12	4	4	0	1	3	3	1	2	1	22.5	75.5
G13	4	3	1	1	3	4	0	4	0	25.75	77.75

to the call for subjects, and of these, 16 ended up being subjects in the experiment. These 16 subjects were distributed into four four-person groups, G10–G13, as shown in the last four lines of Table 4. Recall that all Experiment 3 groups used POEPMcreate.

Table 4 shows also the demographic data gleaned from Steps 1 and 2. As in Experiments 1 and 2, we used these data about each subject from Steps 1 and 2 in order to create homogeneous groups with nearly equivalent spreads of CS knowledge, English fluency, work experience, and individual creativity. Table 4 shows also that despite that the average Williams test scores for the thirteen groups in the experiment were in a wide range from 60.5 to 86.5 out of a possible 100, the average Williams test scores for the 4 two-person groups was 75.375, the average of the Williams test score for the five one-person groups was 75.4, and the average of the Williams test score for the four four-person groups was 75.75. Thus, the groups were well balanced with respect to their average Williams test scores.

5 Data Obtained from the Three Experiments

Table 5 shows all the data collected from the three experiments. Each row whose first column does not say "Avg" is about one of each group that participated in one of the three experiments. The structure of the table is explained columnwise and then rowwise.

The first two columns give a group's characteristics:

1. its experiment number and
2. the number of members in it.

The next four columns, under the collective header "Original", give the data gathered and calculated from the experiment. (The next four columns, under the collective header

Table 5 Generated and Scaled, Raw and New, Requirement Ideas, Per Group and Per Member, for POEPMcreate Groups in Three Experiments

		Original				Scaled			
		PTR	PAR	PTN	PAN	PTR	PAR	PTN	PAN
Exp #	Group Size	# Raw Ideas Generated by Group	Average # Raw Ideas per Member	# New Ideas Generated by Group	Average # New Ideas per Member	# Raw Ideas Generated by Group	Average # Raw Ideas per Member	# New Ideas Generated by Group	Average # New Ideas per Member
1	4	74	18.5	70.5	17.625	39.96	9.99	38.07	9.5175
1	4	76	19	70.5	17.625	41.04	10.26	38.07	9.5175
3	4	40	10	36.5	9.125	40	10	36.5	9.125
3	4	40	10	35.5	8.875	40	10	35.5	8.875
3	4	44	11	36	9	44	11	36	9
3	4	38	9.5	28	7	38	9.5	28	7
Avg	4	52	13	46.17	11.54	40.5	10.125	35.3567	8.8392
2	2	40	20	32.5	16.25	53.2	26.6	43.23	21.613
2	2	42	21	32	16	55.86	27.93	42.56	21.28
2	2	45	22.5	36	18	59.85	29.93	47.88	23.94
2	2	63	31.5	51.5	25.75	83.79	41.9	68.5	34.248
3	2	66	33	46	23	66	33	46	23
3	2	30	15	20.5	10.25	30	15	20.5	10.25
3	2	90	45	68.5	34.25	90	45	68.5	34.25
3	2	67	33.5	57.5	28.75	67	33.5	57.5	28.75
Avg	2	55.38	27.69	43.06	21.53	63.2125	31.6075	49.3338	20.7914
3	1	27	27	19.5	19.5	27	27	19.5	19.5
3	1	30	30	29	29	30	30	29	29
3	1	18	18	18	18	18	18	18	18
3	1	18	18	17	17	18	18	17	17
3	1	27	27	15.5	15.5	27	27	15.5	15.5
Avg	1	24	24	19.8	19.8	24	24	19.8	19.8

"Scaled", are data whose need is explained in Section 8.3 and which are to be ignored for now.) Under (each of) "Original" (and "Scaled"), each of the four columns gives data that figure in deciding support for the subhypothesis for which the last three letters of its name, i.e., "PTR", "PAR", "PTN", and "PAN", matches the header of the column. Under the three-letter header of a column is a description of the data displayed in the column. We use this three-letter column header as the name of the dependent variable whose values appear under the header. With this naming convention, the name of the dependent variable that is relevant to a subhypothesis appears as the last three letters of the subhypothesis's name, e.g., the values of the dependent variable PTR are relevant to the HPTR subhypothesis.

Rowwise, there are three sections, separated by blank rows, each about one size of group doing POEPMcreate. At the end of each section, comes a row whose first column says "Avg", that gives data-column-by-data-column, the average of the section's data for the column.

Figures 3 and 4 plot the “Original” data of Table 5. Specifically, Fig. 3 shows a graph plotting the numbers of raw and new requirement ideas generated by entire groups in all three experiments while Fig. 4 shows a graph plotting the numbers of raw and new requirement ideas generated on average by each member of groups, again from all three experiments. From these graphs, it is already apparent that a two-person group outperforms a four-person group. In both graphs, many of the bars for two-person groups are taller than most of the bars for four-person groups.

The next section considers problems about the gathered data that make their analysis difficult.

6 Data Problems

This section discusses problems with three aspects about the conduct of the experiments, threats to construct validity of the conclusions, which are mitigated by the introduction of additional independent variables.

6.1 Validity of Combining Data from Experiments 1, 2, and 3

Experiment 3’s design and conduct were identical to those of Experiments 1 and 2 (Sakhnini et al. 2012). Each group in Experiments 1 and 2 and in both rounds of Experiment 3 participated in a Step 4 (viz. Section 3.5) session for the same amount of time so that the resources for all groups in the experiment would be the same. Each such group generated requirement ideas for the same Web site. Of course, the real-life Web site had undergone content but not structural changes during the interludes between runs of the experiments. We believe that the structure of the site, i.e., the types of the data present, e.g., the school calendar, and their relationships with each other, should have an effect on idea generation while the contents, e.g., the time and dates of specific events and students and teachers involved, should have no effect on idea generation. The sole differences between experiments were in the number of subjects, the number of groups, and the number of subjects per group. Since each of these differing numbers is an independent variable of the hypotheses, we expect that we are

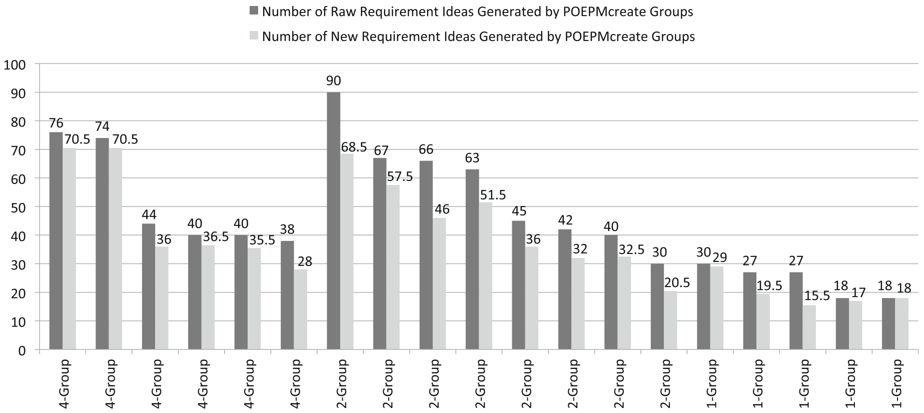


Fig. 3 Numbers of Raw and New Requirements Ideas Generated by POEPMcreate Groups

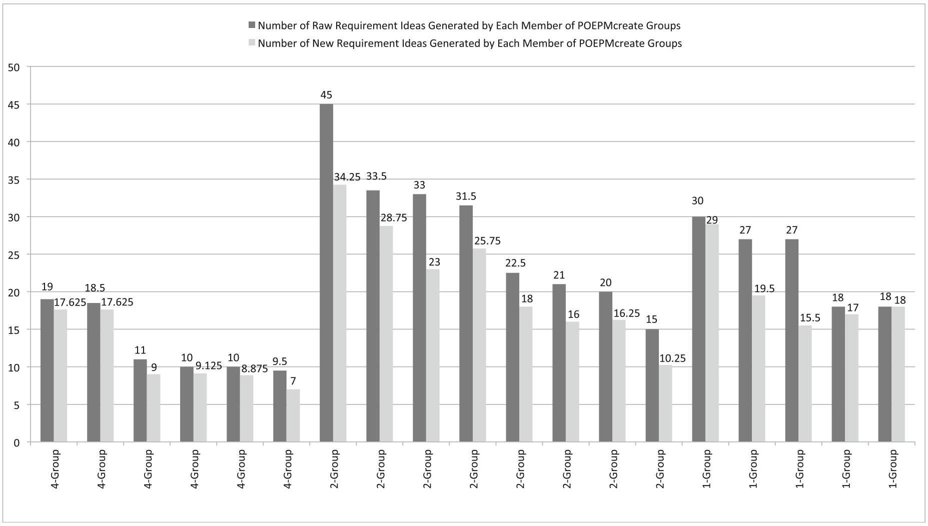


Fig. 4 Numbers of Raw and New Requirements Ideas Generated by Each Member of POEPMcreate Groups

able to use the data of all three experiments to address and test the two null hypotheses H1 and H2.

Of course, it will be necessary to test whether the expectations are borne out. We do that testing by letting the experiment number be an independent variable and seeing if it has a significant effect on the dependent variables. This independent variable is nominal and has values $E1$, $E2$, and $E3$, denoting Experiments 1, 2, and 3, respectively.

6.2 Did Balancing the Creativity Scores of the Groups Work?

Differences in the subjects' individual creativity could affect the results. In an effort to avoid this effect, within each experiment, we distributed the available subjects into groups with approximately the same average Williams test scores. That is, in each experiment, we balanced the groups by their average Williams test scores. It will be necessary to test whether this balancing worked as expected. Even if within an experiment, the balancing worked, there is no guarantee that the balancing worked, and worked uniformly, across the three experiments.

As a matter of fact, the average Williams test score, out of 100, for the subjects

– in Experiment 1 was 70.81,
– in Experiment 2 was 73.44, and
– in Experiment 3 was 75.58.

A single-factor analysis of variance (ANOVA) test shows that there is no significant difference between these averages.

Nevertheless, to dispel any doubt about whether balancing worked and worked uniformly over the three experiments, we let a group's average Williams test score be an independent variable, and we see if it has a significant effect on the dependent variables. Note that any difference among experiments in the way a group's average Williams test score affects the dependent variables will be reflected also in the test of the effect of the experiment numbers

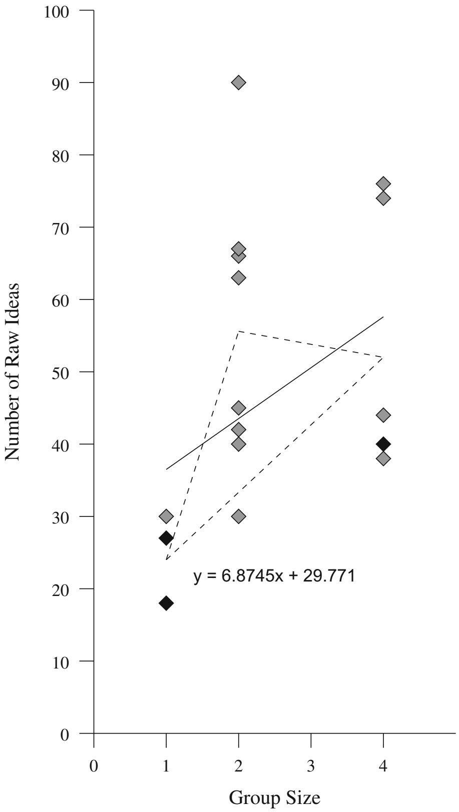


Fig. 5 Linear Regression and Point-by-Point Linear Regression for PTR Data

on the dependent variables. The average creativity test score independent variable, crt, is numerical and has values in the range of 0 through 100.

6.3 Treatment of Group Sizes

Initially, we had treated the group size independent variable as a numerical variable that takes on three values, 1, 2, and 4, in the experiments. As a numerical variable, the value 4 is twice the value 2, which in turn is twice the value 1. Doing a linear regression of the dependent variables on group size carries the assumption that the dependent variables *are linearly related to group size*. That assumption is just not borne out, because the dependent variables proved *not* to be linearly related to group size. For example, Fig. 5 shows with a solid line and a formula, the linear regression for the number of raw requirement ideas generated per group as a function of a numerical group size. It shows with dashed lines, the linear regressions for the three possible pairings of group sizes. The overall regression is very different from the regression for each of the pairings of group sizes. Therefore, we decided to treat group size as a nominal variable, with three values, $s1$, $s2$, and $s4$ and to do regressions for each pair of group size values.

7 Multivariate Regressions

We used multivariate regressions (Berry and Sanders 2000) to compute the coefficients of the effect on dependent variables of changes in the independent variables and to compute their statistical significance.

7.1 Regressions for POEPMcreate

For the POEPMcreate data, we decided to run for each group's value of each dependent variable, PTR, PTN, PAR, and PAN, a regression of it

- on the group's value for the independent numerical variable, crt, its average Williams test score, which is in the range [0 – 100], and
- on each of the following changes in the independent nominal variables: $E1 \rightarrow E2$, $E2 \rightarrow E3$, $E1 \rightarrow E3$. $s1 \rightarrow s2$, $s2 \rightarrow s4$, and $s1 \rightarrow s4$,

As a result, there are $4 \times 7 = 28$ regressions and their P-values to consider. Twelve of these 28 are to answer the research question by testing 4 subhypotheses over 3 pairs of group-size changes, by testing whether 3 pairs of group-size changes have effects on the 4 dependent variables. Another twelve of these are for testing whether 3 pairs of experiment number changes have effects on the 4 dependent variables. Finally, another four of these are for testing whether the groups' average creativity scores have effects on the 4 dependent variables. Thus, each regression and its P-value is

- for testing a previously established subhypothesis arising from the research question
- for testing for the existence of a previously noted unwanted effect.

Therefore, despite the large number of regressions and P-values, we are not fishing for hypotheses.

7.2 Regression Calculations

For the regression of a dependent variable on a numerical independent variable, the coefficient is the slope of the regression line that passes near the dependent variable's data points. The slope expresses the expected change in the dependent variable's value as a result of a change in the independent variable's value. For the regression of a dependent variable on the changes in values of a nominal independent variable, the coefficient is the expected change in the dependent variable's value as a result of the given change in the value of the independent variable. For both kinds of variables, the statistical significance of the expected change is given as a P-value, which needs to be less than $\alpha = 0.05$. To do the regressions, we used MS Excel 2010 and the Analysis ToolPak. The table generated for a regression by the tool pack shows at least its coefficient and its P-value.

8 Regression Results

Tables 6, 7, 8 and 9 give the results of the regression calculations of all the dependent variables on all the independent variables or changes thereof. As with Table 5, there are columns for original data and columns for so-called scaled data whose need and use are

Table 6 Regressions for PTR — POEPMcreate, Total per group, Raw ideas

Independent Variable	Original		Rescaled	
or Change Thereof	Coefficient	P-value	Coefficient	P-value
crt	−0.291580199	0.567169434	−0.391747566	0.463615259
$E1 \rightarrow E2$	−49.54437592	0.005243091	0.873029109	0.955823622
$E2 \rightarrow E3$	16.47895050	0.103490179	1.054368914	0.916227542
$E1 \rightarrow E3$	−33.06542542	0.014190472	1.927398022	0.876946771
$s1 \rightarrow s2$	39.24271050	0.000674015	39.24020631	0.000956873
$s2 \rightarrow s4$	−22.64065743	0.030533628	−22.60309466	0.037318200
$s1 \rightarrow s4$	16.60205307	0.083354149	16.63711165	0.095221507

described in Section 8.3. The present section concerns the coefficients and P-values of the original data that are in columns under the header "Original".

8.1 Checking for Possible Threats

As hoped for, a group's average Williams test score, crt, has no effect on any dependent variable. The coefficients for the numerical variable crt for PTR, PTN, PAR, and PAN are −0.29, −0.18, −0.15, and −0.13, respectively, all being very close to 0, and they are not significant, with P-values of 0.57, 0.67, 0.56, and 0.56, respectively, all greater than 0.05.

Table 10 summarizes the effects of changes in experiment number on the subhypothesis dependent variables. The table's structure is described first columnwise and then rowwise.

The first column gives the compared experiments, $E1 \rightarrow E2$, $E2 \rightarrow E3$, and $E1 \rightarrow E3$. The last four columns concern subhypothesis dependent variables, as indicated by the headers of the four columns and superheaders of pairs and the quadruple of columns.

Rowwise, the table is divided into two sections separated by the thickest horizontal rule:

1. the header section and
2. the results section.

In the header section, the headers for the first column is a simple description. The headers for the last four columns bear careful explanation. Reading from top to bottom, the topmost header says that the four columns are about subhypotheses. The first two of these four

Table 7 Regressions for PTN — POEPMcreate, Total per group, New ideas

Independent Variable	Original		Rescaled	
or Change Thereof	Coefficient	P-value	Coefficient	P-value
crt	−0.181949813	0.676068493	−0.272127704	0.552595134
$E1 \rightarrow E2$	−46.18468145	0.002990157	−0.993950955	0.941572020
$E2 \rightarrow E3$	10.57987453	0.212390119	−1.737180739	0.840465056
$E1 \rightarrow E3$	−35.60480692	0.003522347	−2.731131694	0.798884257
$s1 \rightarrow s2$	28.32045125	0.002506902	28.31819681	0.003489221
$s2 \rightarrow s4$	−14.05676882	0.102344896	−14.02295211	0.118392283
$s1 \rightarrow s4$	14.26368243	0.082728980	14.29524470	0.095635913

Table 8 Regressions for PAR — POEPMcreate, Average per group member, Raw ideas

Independent Variable	Original		Rescaled	
or Change Thereof	Coefficient	P-value	Coefficient	P-value
crt	−0.159017017	0.561493124	−0.209117916	0.463874233
$E1 \rightarrow E2$	−16.11517882	0.063748945	0.471065356	0.955373226
$E2 \rightarrow E3$	8.272542542	0.125856207	0.557794789	0.917022922
$E1 \rightarrow E3$	−7.842636277	0.233647540	1.028860144	0.877018927
$s1 \rightarrow s2$	7.621024575	0.133062912	7.619772052	0.146788221
$s2 \rightarrow s4$	−21.44036862	0.000902492	−21.42158078	0.001219004
$s1 \rightarrow s4$	−13.81934404	0.012284331	−13.80180873	0.015187802

columns are about H1, which is about whole groups, and the last two of these four columns are about H2, which is about group members on average.

H1 has two subhypotheses, each about its own dependent variable:

- HPTR, about numbers of raw requirement ideas generated by whole groups, the dependent variable PTR, and
- HPTN, about numbers of new requirement ideas generated by whole groups, the dependent variable PTN.

Also H2 has two subhypotheses:

- HPAR, about average numbers of raw requirement ideas generated by group members, the dependent variable PAR, and
- HPAN, about average numbers of new requirement ideas generated by group members, the dependent variable PAN.

The results section has three rows, one for each of the three regressions on changes in the nominal experiment independent variables, $E1 \rightarrow E2$, $E2 \rightarrow E3$, and $E1 \rightarrow E3$. The information that is at the intersection of

- the row for the change in one nominal independent variable, CIV, e.g., $E1 \rightarrow E2$, and
- the column for one subhypothesis, e.g., HPTR,

Table 9 Regressions for PAN — POEPMcreate, Average per group member, New ideas

Independent Variable	Original		Rescaled	
or Change Thereof	Coefficient	P-value	Coefficient	P-value
crt	−0.138556740	0.556947638	−0.183625140	0.455529535
$E1 \rightarrow E2$	−13.85219269	0.064181892	0.634622839	0.930277344
$E2 \rightarrow E3$	5.408891850	0.236198545	−0.748687149	0.871102781
$E1 \rightarrow E3$	−8.443300839	0.142395998	−0.114064310	0.984105503
$s1 \rightarrow s2$	4.259036081	0.317338860	4.257909371	0.335171702
$s2 \rightarrow s4$	−15.51054122	0.003281254	−15.49364037	0.004273012
$s1 \rightarrow s4$	−11.25150514	0.016642691	−11.23573120	0.020376945

Table 10 Summary of the Effects of Changes in Experiment Number on the Subhypothesis Dependent Variables

Compared Experiments (E)	Subhypotheses			
	H1		H2	
	# Raw Requirement Ideas Generated by	# New Requirement Ideas Generated by	# Raw Requirement Ideas Generated by	# New Requirement Ideas Generated by
	Whole Group		Group Member	
	PTR	PTN	PAR	PAN
$E1 \rightarrow E2$	↓ 49.54 **	↓ 46.18 **	↓ 16.12	↓ 13.85
$E2 \rightarrow E3$	↑ 16.48	↑ 10.58	↑ 8.27	↑ 5.41
$E1 \rightarrow E3$	↓ 33.06 *	↓ 35.60 **	↓ 7.84	↓ 8.44

is about a regression of the dependent variable PXY, from subhypothesis HP$XY$, on CIV, e.g., PTR, from HPTR, on $E1 \rightarrow E2$, and has a triple reporting the result of this regression:

1. an arrow to report the direction of the coefficient of regression, with ↑ for positive and ↓ for negative; and
2. a numeral to indicate the magnitude of the coefficient of the regression.
3. zero to three asterisks, reporting the strength of statistical significance of the result of the regression;[5]

Table 10 shows that some, but not all experiment number changes have significant effects on some but not all dependent variables.[6] These significant results are troublesome, because they say that despite all of our efforts to ensure that the experiments were run identically, there are measurable differences between the experiments. In particular, for each dependent variable, PTR, PTN, PAR, and PAN, Experiment 1 values are greater than each of Experiment 2 and Experiment 3 values. For the per-group dependent variables, PTR and PTN, these differences are significant. For the per-team-member dependent variables PAR and PAN, whose values are one half or one quarter of those of the corresponding per-group variable, the differences are not significant, even though they *are* real. In no case, is the difference between Experiment 2 and Experiment 3 values significant.

Examination of Tables 2 through 4 reveals no sustained differences between the characteristics of the groups in the three experiments that would account for the observed significant differences in the numbers of requirement ideas generated by groups in the different experiments. The only difference we can think of, not apparent in the table, is that all participants in Experiment 1 were graduate students taking a graduate seminar titled “Advanced Topics in Requirements Engineering” that was focusing, by the students’ topic choices, on empirical studies in requirements engineering. Perhaps the participants in

[5] The number of asterisks indicates the order of magnitude of the deciding P-value, i.e., the number of 0s after the decimal point before the first non-0 digit. Also, a single asterisk is shown *only* if the P-value is less than 0.05.

[6] Thus, there *is* an interaction between the independent variables experiment number and group size in their effects on the dependent variables PTR, PTN, PAR, and PAN. Section 8.3 deals with this interaction.

Experiment 1 had more intrinsic motivation to do well than did the paid participants in the other experiments. Their greater intrinsic motivation might have led to their being more effective in generating more raw requirement ideas than were participants in the other experiments.

Regardless of the reason for the observed differences as a result of differences in experiment number, it is necessary to ascertain that these observed differences are not affecting the final results. Section 8.3 describes this ascertaining.

8.2 Checking Support for Subhypotheses

Table 11 summarizes the effects of changes in group size on the the subhypothesis dependent variables. The table's structure is similar to that of 10.

The differences in the two tables' structures are in the results section. The results section of Table 11 is divided into three two-row subsections, one for each of the three regressions on changes in the nominal group size independent variables, $s1 \rightarrow s2$, $s2 \rightarrow s4$, and $s1 \rightarrow s4$. The information that is at the intersection of

- the two rows for the change in one nominal independent variable, CIV, e.g., $s1 \rightarrow s2$, and
- the column for one subhypothesis, e.g., HPTR,

is about a regression of the dependent variable PXY, from subhypothesis HP$XY$, on CIV, e.g., PTR, from HPTR, on $s1 \rightarrow s2$, and has two rows:

1. The first row consists of a triple not enclosed in parentheses, reporting the result of the regression.
2. The second row consists of a triple enclosed in parentheses, reporting the result of a regression on the *scaled* version of the values used for the first-row regression (which is to be ignored until Section 8.3).

Table 11 Summary of the Effects of Changes in Group Size on the Subhypothesis Dependent Variables

	Subhypotheses			
	H1		H2	
Compared Group Sizes (s)	# Raw Requirement Ideas Generated by	# New Requirement Ideas Generated by	# Raw Requirement Ideas Generated by	# New Requirement Ideas Generated by
	Whole Group		Group Member	
	PTR	PTN	PAR	PAN
$s1 \rightarrow s2$	↑ 39.24 ***	↑ 28.32 **	↑ 7.62	↑ 4.26
	(↑ 39.24 ***)	(↑ 28.32 **)	(↑ 7.62)	(↑ 4.32)
$s2 \rightarrow s4$	↓ 22.64 *	↓ 14.06	↓ 21.44 ***	↓ 15.51 **
	(↓ 22.60 *)	(↓ 14.02)	(↓ 21.44 **)	(↓ 15.49 **)
$s1 \rightarrow s4$	↑ 16.60	↑ 14.26	↓ 13.82 *	↓ 11.25 *
	(↑ 16.64)	(↑ 14.30)	(↓ 13.80 *)	(↓ 11.24 *)

The triple has the same structure as in Table 10.

Examination of the non-parenthesized triples in the results section of Table 11 shows that some, but not all group size changes have significant effects on some but not all dependent variables. In particular, for H1 about total requirement ideas generated by whole groups:

- A group of size 2 generates about 39 and 28 more raw and new requirement ideas, respectively, than does a group of size 1, and each difference is very significant.
- A group of size 4 generates about 23 and 14 *fewer* raw and new requirement ideas, respectively, than does a group of size 2, and this difference is significant for only the raw ideas.
- A group of size 4 generates about 17 and 14 more raw and new requirement ideas, respectively, than does a group of size 1, and neither difference is significant.

Thus, rejection of HPTR is supported significantly for $s1 \rightarrow s2$ and $s2 \rightarrow s4$ and is only suggested for $s1 \rightarrow s4$. Rejection of HPTN is supported significantly for $s1 \rightarrow s2$ and is only suggested for $s2 \rightarrow s4$ and $s1 \rightarrow s4$. For POEPMcreate, group size makes a difference in the numbers of raw and new requirement ideas generated by groups.

The directions of these rejections were as expected in four of these six cases: For HPTR and $s1 \rightarrow s2$, HPTR and $s1 \rightarrow s4$, HPTN and $s1 \rightarrow s2$, and HPTN and $s1 \rightarrow s4$, the more group members, the more raw and new ideas are generated. However, for HPTR and $s2 \rightarrow s4$ and HPTN and $s2 \rightarrow s4$, the more group members, the *fewer* raw and new ideas are generated.

We were very surprised to see that a group of size 4 generates *fewer* raw and new ideas than does a group of size 2. This surprise suggests that perhaps the larger group-management overhead in a larger group is decreasing the larger group's effectiveness in requirement idea generation. As mentioned in Section 1.1, this phenomenon has been observed in brainstorming (Dornburg et al. 2008; Dennis and Valacich 1993; Furnham and Yazdanpanahi 1958; Taylor et al. 1958; Isaksen and Gaulin 2005; Aurum and Martin 1998; Kohn and Smith 2011; Ocker 2010). Ocker (2010, pg. 62) puts it quite simply, "Osborn's claims that traditional face-to-face brainstorming groups produce more and better ideas than the same number of people working alone have been refuted time and again." Section 10 explores this issue thoroughly. In the meantime, the analysis of the data continues in order to gather evidence for a conclusion.

This surprise, for sure, says that the four subhypotheses that we designed the experiments to test are not fine enough; actual group sizes make a difference, as there is not an overall uni-directional tendency. So it will be necessary, as is done in Table 11, to include the relevant group sizes in the statement of a subhypothesis. This necessity strengthens the arguments of Section 7.1 that we needed to do regressions on each pair of group sizes and that we were not fishing for hypotheses.

For H2 about requirement ideas generated by average members of groups:

- Per member, a group of size 2 generates about 8 and 4 more raw and new requirement ideas, respectively, than does a group of size 1, and neither difference is significant.
- Per member, a group of size 4 generates about 21 and 16 *fewer* raw and new requirement ideas, respectively, than does a group of size 2, and each difference is very significant.
- Per member, a group of size 4 generates about 14 and 11 *fewer* raw and new requirement ideas, respectively, than does a group of size 1, and each difference is significant.

Thus, rejection of HPAR is supported significantly for $s2 \rightarrow s4$ and $s1 \rightarrow s4$ and is only suggested for $s1 \rightarrow s2$. Rejection of HPAN is supported significantly for $s2 \rightarrow s4$ and $s1 \rightarrow s4$ and is only suggested for $s1 \rightarrow s2$. For POEPMcreate, group size makes a difference in the average numbers of raw and new requirement ideas generated by group members.

The directions of these rejections were as expected in only two of these six cases: For HPAR and $s1 \rightarrow s2$ and HPAN and $s1 \rightarrow s2$, the more group members, the more raw and new ideas are generated per group member. However, for HPAR and $s2 \rightarrow s4$, HPAR and $s1 \rightarrow s4$, HPAN and $s2 \rightarrow s4$, and HPAN and $s1 \rightarrow s4$, the more group members, the *fewer* raw and new ideas are generated per group member.

After the very surprising results for H1, perhaps the results for H2 are not so surprising. Per group member, the larger of two groups tend to generate fewer raw and new requirement ideas than the smaller group. For the group sizes tested, only a group of size 2 generates more raw and new requirement ideas per group member than a group of size 1. These results strengthen the suggestion that maybe the larger group-management overhead in a larger group is decreasing the larger group's effectiveness in requirement idea generation. In these results, the only exception to this suggestion is the case in which the smaller group is of size 1. A group of size 1 is not really a group, and it cannot suffer any group-management overhead. Perhaps in going from an individual to a group of size 2, the drag from the group-management overhead is small enough that it is dominated by the synergy of a group, which cannot exist in an individual. This conclusion is consistent with Fred Brooks's observation that group communication overhead grows quadratically with an increasing number of group members (Brooks 1995). At group size 2, the group-management overhead is smaller than at group size 4; so synergy dominates overhead at group size 2, but is dominated by overhead at group size 4.

8.3 Scaling POEPMcreate Data and New POEPMcreate Results

In order to show that the observed unwanted effect of the experiment number on the dependent variables is not affecting subhypothesis decisions and thus the final results, we decided to try scaling the dependent variables by amounts that eliminate the effect of the experiment number nominal variable. That is, we wanted the regression on changes in the experiment number nominal variable to end up with coefficients near zero and with P-values that are greater than 0.05. To do this scaling, we needed to find one experiment that had groups of all three sizes. That experiment is Experiment 3. Experiment 3 was then taken as the *base experiment*. Then, for each pair of experiments involving the base experiment (i.e., ($E1$, $E3$) and ($E2$, $E3$)), we had to find one group size that was used in the two experiments, and then use as the scaling factor between the two experiments, the ratios of the average numbers of raw requirement ideas generated by groups of that size in the two experiments.

For the difference between Experiment 1 and the base experiment, Experiment 3, we saw that there are two groups of size 4 that did POEPMcreate in Experiment 1 and four groups of size 4 that did POEPMcreate in Experiment 3. The PTR values for the two Experiment 1 groups were 74 and 76 for an average of 75. The PTR values for the four Experiment 3 groups were 40, 40, 44, and 38, for an average of 40.5. Therefore, to scale Experiment 1's dependent variable values to be comparable to those for Experiment 3, we needed to multiply each Experiment 1 dependent variable by $\frac{40.5}{75} = 0.54$.

For the difference between Experiment 2 and the base experiment, Experiment 3, we saw that there are four groups of size 2 that did POEPMcreate in Experiment 2 and four groups of size 2 that did POEPMcreate in Experiment 3. The PTR values for the four Experiment 2 groups were 40, 42, 45, and 63 for an average of 47.5. The PTR values for the four Experiment 3 groups were 66, 30, 90, and 67, for an average of 63.25. Therefore, to scale Experiment 2's dependent variable values to be comparable to those for Experiment 3, we needed to multiply each Experiment 2 dependent variable by $\frac{63.25}{47.5} = 1.33$.

Each value of the dependent variables, PTR, PTN, PAR, and PAN, was rescaled by the proper value:

- If the value was obtained in Experiment 1, it was multiplied by 0.54.
- If the value was obtained in Experiment 2, it was multiplied by 1.33.
- If the value was obtained in Experiment 3, it was left alone.

The resulting scaled values are shown in the four columns of Table 5 that are under the header "Scaled", in the same notation used for the corresponding four columns that are under the header "Original".

Then, all of the regressions from Section 7.1 were run again with the scaled values. The results of these regressions are actually in the two rightmost numerical columns of Tables 6 through 9, the columns under the header "Rescaled".

Again, as hoped for, a group's average Williams test score, crt, has no effect on any dependent variable. The coefficients for the numerical variable crt for PTR, PTN, PAR, and PAN are -0.39, -0.27, -0.21, and -0.18, respectively, all being very close to 0, and they are not significant, with P-values of 0.46, 0.55, 0.46, and 0.46, respectively, all greater than 0.05.

Most importantly, now, changes in the nominal experiment value have no significant effect on any dependent variable.

1. The coefficients for the change $E1 \rightarrow E2$ for PTR, PTN, PAR, and PAN are 0.87, -0.99, 0.47, and 0.63, respectively, all being very close to 0, and they are not significant, with P-values of 0.96, 0.94, 0.96, and 0.93, respectively, all greater than 0.05.
2. The coefficients for the change $E2 \rightarrow E3$ for PTR, PTN, PAR, and PAN are 1.05, -1.73, 0.56, and -0.75, respectively, all being close to 0, and they are not significant, with P-values of 0.92, 0.84, 0.92, and 0.88, respectively, all greater than 0.05.
3. The coefficients for the change $E1 \rightarrow E3$ for PTR, PTN, PAR, and PAN are 1.93, -2.73, 1.03, and -0.11, respectively, all being close to 0, and they are not significant, with P-values of 0.88, 0.80, 0.88, and 0.98, respectively, all greater than 0.05.

So now, let us examine the effect of group size changes on the rescaled dependent variables. The effects of group size changes after scaling are summarized in the parenthesized triples in the POEPMcreate section of Table 11. Again, some, but not all group size changes have significant effects on some but not all dependent variables. Perhaps surprisingly, in fact, all and only those results that were significant with unscaled values are significant with scaled values, albeit, in a few cases, less strongly so. Moreover, the direction and approximate magnitude of each coefficient is unchanged after scaling. So, the conclusions drawn from the unscaled results still hold, and the significant effect of the experiment number on the dependent variables does not affect the conclusions about the subhypotheses.

That these results are basically unchanged as a result of scaling

- gives us confidence that scaling was the correct thing to do,
- combined with that after scaling, the effect of changes in experiment number is not significant, gives us confidence that the scaling factors used are correct, and

- gives us confidence that analyzing the combined data of the three experiments is legitimate and gives meaningful results.

The results of this section justify rejection of the null hypotheses that the number of members of an elicitation group has *no* effect on the quantity and quality of the requirement ideas generated by the group and by a member of the group. So, we can say for sure that the number of members of an elicitation group *has* such an effect. While the data suggest, in some cases with statistical significance, specific directions for the effect, it is premature to be certain. Section 10 begins the future work to arrive at this certainty.

9 Threats to Validity and Future Work to Address Them

Since the conduct of the three experiments were identical, many of the possible threats to the validity of the conclusions to Experiment 3 and the analysis of the three experiments threatened also the validity of the conclusions to Experiments 1 and 2. The discussion of these enduring threats is repeated from the paper about Experiments 1 and 2 (Sakhnini et al. 2012).

9.1 Construct Validity

Construct validity is the extent to which the experiment and its various measures test and measure what they claim to test and measure. Certainly, the groups were trying to be creative in their idea generation. As mentioned in Section 3.3, counting raw ideas is correct, because each of EPMcreate, POEPMcreate, and the first stage of brainstorming has as a principal goal the generation of as many ideas as possible, under the long-believed principles that an idea's quality is evaluated only later and that quality follows quantity (Osborn 1953). However, there is some recent new evidence that shows that in a comparison of several CETs applied to requirements elicitation, the CET that generated the most raw ideas was different from the CET that generated both the most creative ideas and the most ideas that ended up being requirements downstream (Berntsson Svensson and Taghavianfar 2015).

The shakiest measure used in the experiment is the Williams test of individual creativity. With any psychometric test, such as the Williams test and the standard IQ tests, there is always the question of whether the test measures what its designers say it measures. The seminal paper describing the test discusses this issue (Williams and Taylor 1966), and the test seems to be accepted in the academic psychometric testing field (Dow 2016). The original test was designed for testing children, and the test seems to be used in U.S. schools to identify gifted and talented students (West Side School District 2016). We modified the test to be for adults attending a university or working (Mich et al. 2005, 2010). Each of the authors has examined the test and has determined for him- or herself that the test does examine at least something related to creativity if not individual creativity itself. Finally, the same modified-for-adults Williams test, in Italian and English versions, has been used in all of our past experiments about CETs and will be used in all of our future experiments about CETs. Therefore, even if the test does not measure individual creativity exactly or fully, the same error is made in all our experiments. Thus, the results of all of these experiments should be comparable.

One other threat to construct validity has been discussed already in earlier sections of the paper, as it arose in dealing with the data. Specifically, Section 3.3 deals with using only the newness of a requirement idea relative to the existing Web site as the measure of the idea's quality.

9.2 Internal Validity

Internal validity is that one can conclude the causal relationship that is being tested by the experiment. In this case, we are claiming that the differences in groups size caused the observed differences in the quantity and quality of the requirement ideas generated. We know from being in the room with the groups that each group was actively using the same POEPMcreate while it was generating its requirement ideas. We carefully assigned subjects to the groups so that the groups were balanced in all personal factors, especially individual creativity, that we thought might influence the subjects' abilities to generate requirement ideas. The original regressions did show that experiment number has a significant effect on the numbers of raw and new requirement ideas generated, but after rescaling the numbers of raw and new requirement ideas generated, this effect disappeared. Therefore, we believe that, after rescaling, the only factor that can account for the differences in the number of requirement ideas among groups is the sizes of the groups.

There is also the question of the impacts of the subjects' domain knowledge on the quantity and quality of the generated requirement ideas: How does the subjects' domain knowledge affect the results? A common belief among practitioners of brainstorming is that a group's creativity is boosted when the group has a mix of different competencies, backgrounds, viewpoints, and domain knowledge (Osborn 1953; von Bertalanaffy 1976). This belief has been empirically validated in the context of using brainstorming to help generate requirement ideas (Niknafs and Berry 2016). Also, we do know that differences in domain knowledge make a difference in the ideas generated: About half of the subjects in another experiment, in which two of the authors of the present paper were involved, were experts in the Web site's domain and half were not, and different results were observed as a result of the difference in domain knowledge (Mich et al. 2010). The experiments described in this paper entirely avoided this issue. All of the subjects were CS students with similar sets of competencies and backgrounds. These students are all part of the generation that grew up with Web sites and they all were quite recently in high schools, which probably had their own Web sites. Therefore, they all had similar and similarly complete domain knowledge. Thus, it is unlikely that any group gained any advantage over another on the basis of this issue.

Two other threats to internal validity have been discussed already in earlier sections of the paper, as they arose in dealing with the data.

1. Section 6.1, all but the first paragraph of Section 8.1, and Section 8.3 observe and deal with an unwanted effect of the experiment number on the dependent variables, in a way that the effect is no longer a threat.
2. Section 6.2 and the first paragraph in Section 8.1 deal with the potential, but unrealized effect of differences in group members' individual creativity on the dependent variables.

9.3 External Validity

External validity is that the results can be generalized to other cases, with different kinds of subjects, with different kinds of CBS. There are several threats to external validity:

- that a controlled experiment runs the risk of not duplicating the real-life situation it is supposed to model: This threat is present in *every* controlled experiment, as control removes realism and vice versa. However, we did conduct the EPMcreate and POEPMcreate sessions as in real life, as described in Section 2.2 and used a realistic RE problem with a real-life Web site.

- the use of students as subjects instead of requirements elicitation or software development professionals: However, our student subjects had all studied at least a few courses in computer science and software engineering. Moreover, each group had at least one subject with professional experience in computing. In addition, most students at the university educating the subjects are in its cooperative education program. A typical student works one of three terms per year in a paying industrial job, preferably in his or her major area.

  Therefore, one could argue that the subjects were equivalent to young professionals, each at an early stage in his or her CBS development career (Berander 2004; Daun et al. 2016). Indeed, in one of the early experiments (Mich et al. 2005) comparing EPMcreate with brainstorming, namely the Civilia experiment, professional analysts were used as the subjects. The results and the shape of the results of Experiments 1, 2, and 3 are the same as in the early experiment in which professional analysts were used as subjects.

  Independently of whether the subjects are students or professionals and whether computer science students accurately represent professional requirements analysts, it is fair to say that with respect to knowledge about Web sites in general, and about high school Web sites in particular, our subjects were at least equivalent to, if not better than professional requirements analysts. As mentioned, the students grew up with Web sites and they were recently in high schools, which probably had Web sites.
- the particular choice of the types of stakeholders whose viewpoints were used by POEPMcreate sessions: Would other choices, e.g., of teachers, work as well?
- the choice of a particular Web site as the CBS for which to generate requirement ideas: How representative of a CBS is a Web site and how representative of Web sites is the chosen Web site? Would a different Web site or even a different kind of CBS give different results? Our experience in using EPMcreate and POEPMcreate to generate requirement ideas for four different CBSs (Mich et al. 2005, 2006) leads us to believe that the kind of CBS and among Web sites, the kind of Web site have no effect on the effectiveness of any variant of EPMcreate and on the shape of the results.

  In any case, *every* controlled experiment in RE — and in SE for that matter — must pick some small set of specific CBSs, usually only one, to provide the CBS artifacts that are treated in the experiment. Therefore, the threat of non-representativeness is universal in empirical RE and SE, and we live with it.

Not considered in these experiments is what happens downstream in the stage after requirement idea generation, when the idea generators meet with other stakeholders to discuss, prune, enhance, combine, etc. the generated ideas to transform them into actual requirements. Addressing this lack would require a much lengthier experimental procedure to do the additional stage. This lengthening would reduce the number of people willing to participate in the experiment, thus reducing the statistical validity of the results, if there were even results. This lack is typical of experimental studies of a single task that is embedded in a larger process.

9.4 Conclusion Validity

Conclusion validity is the extent to which the statistical inferences drawn from the data lead to a correct conclusion.

Even though the collected data yield statistically significant results, the medium number of groups of each size increases the probability that any positive observations were random false positives. Thus, there is the threat of a so-called Type I error (Wohlin et al. 2000), that of accepting a non-null hypothesis, making a positive claim, when it should be rejected. The only remedy for this threat is to do more experiments in the future with more groups of the same sizes.

Closely aligned with and building on the threat of a Type I error is the *multiple comparisons problem* (Jeff et al. 2016; Gelman et al. 2012). The multiple comparisons problem occurs any time that

- a researcher conducts a number M of statistical tests regarding a collection of hypotheses,
- the number $N < M$ of the tests whose results are statistically significant is less than or equal to[7] the number, $\frac{M}{\alpha}$, of tests that are expected to be statistically significant by pure chance, given the significance level, α, used for the statistical tests, and
- the researcher then claims on the basis of these N statistically significant results, support for some, usually N of the hypotheses in the collection of hypotheses.[8]

Note that when the multiple comparisons problem occurs, there is a high likelihood that a Type I error has occurred for the supported hypotheses. The special case that the collection of hypotheses is one hypothesis being repeatedly tested until a test is found to have a statistically significant result, amounts to unethically ignoring all the negative results and reporting only the positive results.

While there are multiple comparisons in the analysis of Section 8, for each group size change, and for each of the raw and new ideas, the number of ideas generated either by the group or by a group member exhibits a significant difference, many more than would be the case if the observed difference were due to chance. In addition, we are claiming support for only rejecting the null hypothesis that there is no difference and are not claiming any support for any specific hypotheses about the direction of the observed difference.

For the same reason, the error rate threat (Wohlin et al. 2012) that can occur when there are multiple comparisons does not arise.

Two other threats to conclusion validity have been discussed already in earlier sections of the paper, as they arose in dealing with the data.

1. Section 6.3 deals with the possibility that the relationship between group sizes and the dependent variables is not linear by opting to treat the individual group sizes as nominal values.
2. The last paragraph in Section 7.1 deals with the potential, but unrealized threat of fishing for hypotheses.

[7] This “less than or equal to” should be taken with the same grain of salt as is the test of whether $p < \alpha$. If the number N is only a small percentage more than exactly equal to $\frac{M}{\alpha}$, the researcher is still on shaky grounds claiming support for the hypotheses.

[8] The second and third conjuncts in this definition of the multiple comparisons problem came from private communication with William Berry, one of the authors of (Berry and Sanders 2000).

9.5 Dealing with Threats in Future Work

To address these threats to validity, we plan future experiments to get more data points and to do other experiments with different kinds of subjects, different sized groups, different stakeholder viewpoints, and different CBSs.

10 Postanalysis Speculation About Lack of Synergy

Groups are traditionally thought to have synergy, by which the effect of a group is greater than the sum of the effects of its members (Osborn 1953). The results of Section 8 suggest that synergy, if indeed it is present, is not very helpful. In particular, according to Table 11, when POEPMcreate is used to help generate requirement ideas, according to both the original and the rescaled data,

- a four-person group generates on average, fewer raw and new ideas than a two-person group, and significantly so for raw ideas,
- a four-person group member generates on average, significantly fewer raw and new ideas than a two-person group member, and
- a four-person group member generates on average, significantly fewer raw and new ideas than a one-person group member.

Perhaps, synergy is getting drowned out in the larger group, because it has more group-management overhead than the smaller group. Remember that the number of lines of intermember communication in a group is increasing quadratically with increasing group size.

However, not always is the smaller of two groups more effective. According to the same table, when POEPMcreate is used to help generate requirement ideas, according to both the original and the rescaled data,

- a two-person group generates on average, significantly more raw and new ideas than a one-person group, and
- a two-person group member generates on average, more raw and new ideas than a one-person group member.

It appears that among the three tested group sizes, one, two, and four, the group size two is optimal, both in total ideas generated and per group member. Evidently, two people bring out enough synergy to more than counteract the moderate amount of group overhead they incur.

The natural question to ask is “If you have more than two people available for generating requirement ideas with POEPMcreate, which is better:

1. have only one two-person group generating as many requirement ideas as it can, and send the rest of the people to do something else, or
2. form as many two-person groups as possible and let each two-person group and the one remaining one-two-person group independently generate as many requirement ideas as possible?”

The second choice will surely yield more ideas. For example, the original data in Table 5 say that for POEPMcreate, from two two-person groups, you will get on average 110.76 ideas.

Table 12 Shared Ideas Among Pairs of Two-Person Groups in Experiment 2

	F P2 40	G P2 42	H P2 63
E P2 45	32.5 % 13 28.89%	16.67% 7 15.56%	22.22% 14 31.11%
F P2 40		26.19% 11 27.5 %	22.22% 14 35 %
G P2 42			25.4 % 16 38.1 %

For the rescaled data, the effect is even more pronounced. From two two-person groups, you will get on average 126.43 ideas. However, it is possible that there is so much sharing among the ideas of different groups, that there is no advantage to having more than one group.

We searched for shared ideas among the ideas generated by all possible pairs of the four two-person POEPMcreate groups in Experiment 2. Table 12 shows the results of these searches. In the head of a column or row, which is about one group, the first element of the triple is a unique label for the group; the second element is of the form Tn, where (1) T is the CET used by the group, which in this case is always “P” meaning “POEPMcreate”, and (2) n is the number of members in the group, which in this case, is always “2”; and the third element is the number of raw requirement ideas generated by the group.

The triangle (in rows E, F, and G by columns F, G, and H) shows the idea sharing among the pairs of POEPMcreate groups. The reading of the cell in the row for Group E, which generated 45 raw requirement ideas using POEPMcreate, and the column for Group F, which generated 40 raw requirement ideas using POEPMcreate, is that there were 13 ideas in common among the generated ideas of the two groups; 32.5 % of Group F’s ideas were in common with Group E’s ideas; and 28.89 % of Group E’s ideas were in common with Group F’s ideas.

This search shows that among pairs of POEPMcreate two-person groups in Experiment 3, the average and maximum percentage overlap of raw requirement ideas generated were 26.78 and 38.1, respectively.

We searched for shared ideas also among the ideas generated by all possible pairs of the four two-person POEPMcreate groups, all possible pairs of the five one-person POEPMcreate groups, and all possible pairs of a one-person POEPMcreate group with a two-person POEPMcreate group in Experiment 3. Table 13 shows the results of these searches in the same format used in Table 12.

The upper left-hand triangle (in rows A, B, and C by columns B, C, and D) shows the idea sharing among the pairs of two-person groups, the lower right-hand triangle (in rows E, F, G, and H by columns F, G, H, and I) shows the idea sharing among the paris of one-person groups, and the upper right-hand rectangle between the triangles (in rows A, B, C, D by columns E, F, G, H, and I) shows the idea sharing among the pairs of a one-person group with a two-person group.

This search shows that for POEPMcreate in Experiment 3, among pairs of two-person groups, the average and maximum percentage overlap of raw requirement ideas generated were 17.18 and 30, respectively; among pairs of one-person groups, the average and maximum percentage overlap of raw requirement ideas generated were 11.93 and 27.78, respectively; and among pairs of one- and two-person groups together, the average and

Table 13 Shared Ideas Among Pairs of Groups in Experiment 3

	B P2 67	C P2 66	D P2 30	E P1 30	F P1 27	G P1 27	H P1 18	I P1 18
A P2 90	11.94% 8 8.89%	12.12% 8 8.89%	30 % 9 10 %	33.3 % 10 11.11%	25.93% 7 7.78%	7.4 % 2 2.22%	27.78% 5 5.56%	11.11% 2 2.22%
B P2 67		21.21% 14 20.9 %	26.67% 8 11.94%	26.67% 8 11.94%	29.63% 8 11.94%	14.81% 4 5.97%	33.33% 6 8.96%	38.89% 7 10.45%
C P2 66			30 % 9 13.64%	26.67% 8 12.12%	22.22% 6 9.1 %	14.81% 4 6.06%	44.44% 8 12.12%	33.33% 6 9.1 %
D P2 30				20 % 6 20 %	25.93% 7 23.33%	7.4 % 2 6.67%	38.89% 7 23.33%	27.78% 5 16.67%
E P1 30					22.22% 6 20 %	7.4 % 2 6.67%	16.67% 3 10 %	22.22% 4 10.33%
F P1 27						0.0 % 0 0.0 %	16.67% 3 11.11%	27.78% 5 8.23%
G P1 27							16.67% 3 11.11%	5.56% 1 3.7 %
H P1 18								11.11% 2 11.11%

maximum percentage overlap of raw requirement ideas generated were 16.27 and 44.44, respectively.

Combining the two sets of results about idea sharing among two-person groups using POEPMcreate shows that among pairs of POEPMcreate two-person groups, the average and maximum percentage overlap of raw requirement ideas generated were 21.98 and 38.1, respectively.

These results mean that

- from each additional two-person group applying POEPMcreate to generate requirement ideas, only an average of 21.98 % or, at worst, 38.1 % of the ideas generated by the additional group will duplicate an idea generated before, and
- from the additional, last one-person group applying POEPMcreate to generate requirement ideas, only an average of 16.27 % or, at worst, 44.44 % of the ideas generated by the additional group will duplicate an idea generated before.

In no case, will even close to 100 % of the ideas generated by the additional group will duplicate an idea generated before. Therefore, if you have more than two people available for generating requirement ideas with POEPMcreate, and the only possible group sizes are one, two, and four, then form as many two-person groups as possible and let each two-person group and the one remaining one-two-person group independently generate as many requirement ideas as possible.

Future work is needed with experiments designed specifically to test the speculative conclusions of this section. Of course, it is not known how three-person groups fit into this synergy–overhead balance. Only additional experiments can provide the data to answer this question.

As mentioned in Section 1.1, several researchers (Dornburg et al. 2008; Dennis and Valacich 1993; Furnham and Yazdanpanahi 1958; Taylor et al. 1958; Isaksen and Gaulin 2005; Aurum and Martin 1998; Kohn and Smith 2011; Ocker 2010) had noticed that for brainstorming, smaller groups are more effective per person than larger groups, and that individuals are the most effective. Thus, it appears that in brainstorming, group overhead overpowers synergy already with only two people. However, our observation is that for POEPMcreate, an individual working a two-person group is more effective than an individual working alone, although not significantly. Perhaps there is something about POEPMcreate's procedure that mitigates the drag that group overhead places on synergy in brainstorming. Future work is needed to understand the break-even point in the synergy–overhead balance and how the CET affects at what group size is the break-even point.

Section 6.3 explains how examination of plots like that for the PTR data in Fig. 5 (The plot for the PTN data is almost the same) shows that the relations between the independent variable of group size and each dependent variable, the number of raw ideas and the number of new ideas generated by whole groups, is clearly not linear (and hence we needed to consider the group sizes to be nominal values rather than numerical values). The actual slopes of the two regressions on $s1 \rightarrow s2$ and on $s2 \rightarrow s4$, e.g., the two upper dashed lines in Fig. 5, suggest that a quadratic regression on a numerically valued group size variable would work. (For the reason explained below, it is premature to actually do this regression).

A possible explanation of these plots is the following model playing group overhead against group synergy. Since each of overhead and synergy is a group phenomenon, we expect that each will grow quadratically with group size, just as does the number of arcs between nodes at the corners of polygons grows with the number of corners. If we measure synergy as the number S of ideas arising from it, we measure overhead as the number O

of ideas lost as a result of it, and let n be the number of persons in a group, then we expect that

$$S = an^2 + b$$
$$O = An^2 + B$$

for some constants, a, b, A, and B. Then, the total number I of ideas generated by a group of size n is

$$I = S - O$$
$$I = (an^2 + b) - (An^2 + B).$$

For each CET, a, b, A, and B are set to cause the peak at a different n. For example, for brainstorming, the peak is at $n = 1$, while for POEPMcreate, the peak is some where between 2 and 3, inclusive. Lacking data for $n = 3$, we cannot say where the peak is for POEPMcreate. For example if a group with three people generates the same number of ideas as a group with two people, the peak would be at $n = 2.5$.

So, for each CET c, the constants are a_c, b_c, A_c, and B_c, and

$$I_c = (a_c n^2 + b_c) - (A_c n^2 + B_c).$$

We propose this equation for I_c as a theory to be tested empirically for a variety of CETs. For each CET, an experiment similar to those described in this paper will be conducted with all group sizes ranging from 1 through at least 4, or more if necessary, to establish the values of the constants for the CET. We invite the promoters of the various CETs used in requirements elicitation to conduct these experiments with their CETs.

11 Qualitative Triangulation

That we had confirmed that group size makes a difference in applications of POEPMcreate and that others have reported the same about applications of brainstorming led us to wonder what practicing business or requirements analysts (BoRAs) in industry had observed about group sizes in applications of the CETs, including POEMPcreate, that they were using to generate requirement ideas. We decided to use a questionnaire to find out. We hoped that this questionnaire would amount to a qualitative triangulation of the results of our experiments.

In designing the questionnaire, we accepted that very few, if any, industrial BoRAs were using POEPMcreate or even EPMcreate. In fact, the responses to the questionnaire showed that no respondent was using either. We even suspected that many were not even using any described CET. Therefore, to ensure that all BoRAs would be able to answer the questions, we decided to let the questionnaire be about group and individual requirements elicitation, the inherently creative activity of which requirements idea generation is a part. Among the questions would be one that asked which technique was used in group requirements elicitation at the respondent's company. Among the possible answers were "brainstorming" and "other creativity technique". Despite not asking specifically about group and individual use of POEPMcreate, we hoped that the results of the questionnaire would tell us something about optimal or preferred group sizes for requirements elicitation and for applications of CETs for requirement idea generation.

We deployed in late August 2012 an online questionnaire (Mich et al. 2012) titled "Requirements elicitation (ReqElic) in my company". The most important of the questions were about:

- Group vs. individual activity in ReqElic — Requirements are identified as an individual activity, by a single BoRA, working alone.

- Group vs. individual activity in ReqElic — Requirements are identified as an individual activity, by more than one BoRA, each working separately.
- Group vs. individual activity in ReqElic — Requirements are identified as a group activity.

Answering these questions involved choosing between "all", "most", "some", or "none" as an indication of the fraction of projects in which the statement of the question is true. Answering some other questions, e.g.,

- Size of the groups — Groups usually consist of:

involved choosing between some numbers or ranges of numbers.

We sent an advertisement describing the questionnaire to requirements analysts or software development managers that we knew and asked that they send the advertisement on to other people in similar roles. We posted the advertisement and the propagation request on the Facebook, Google+, LinkedIn, ResearchGate, Slideshare, and Twitter accounts of one of the authors. We posted the advertisement and the request also on several e-mail lists, e.g., IIBA,[9] INCOSE,[10] Requirements Engineering Network's forum,[11] RE-online,[12] and Yahoo's Requirements-Engineering Group[13] as well as several LinkedIn groups, including AICA, Community of Practice Systems Engineering (CoP SE); America's Requirements Engineering Association; Business Analysts — Banaglore; ICT Africa; ICT Australia; IEEE Computer Society Italy Chapter; INCOSE; IREB Certified Professional for Requirements Engineering (CPRE); ModernAnalyst.com — Business Analyst Community; Requirements Engineering Specialist Group (RESG); Systems Engineers; and Requirements Engineering. Sometimes we were assisted by the help of a friend who was in the organization and could post advertisements. Thus, we have a convenience-assisted-by-a-snowball sampling.

In the end, we got 53 responses. We cannot know the response rate because we have no idea how many people saw the advertisement inviting people to fill the questionnaire. Normally, the small number of responses would be a concern. However, our goal was an exploratory corroboration of the speculation of the previous section for the purpose of deciding about future work. This questionnaire could end up being a pilot for a future study. The rest of this section gives an analysis of the data from these 53 responses. See https://cs.uwaterloo.ca/~dberry/FTP_SITE/tech.reports/53Xresponses.pdf for an automatically generated summary of the responses.

The answers to the demographics question about the roles the respondent plays in his or her organization shows that many respondents are involved in more than one role, each of 45 % of the respondents is a business or requirements analysts (BoRAs) in all or most of his or her organization's projects, each of 17 % is a software engineer (SWE), and each of 34 % is a project manager (PM).

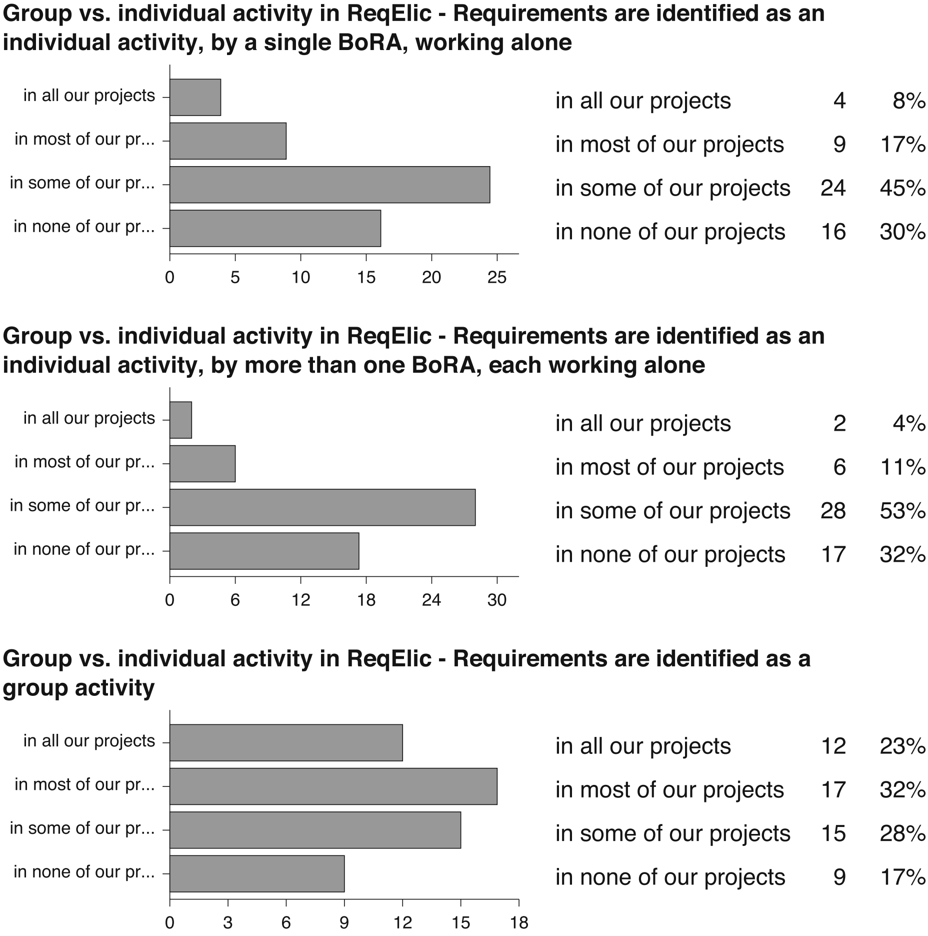
[9] http://www.iiba.org/

[10] http://www.incose.org/

[11] http://www.requirementsnetwork.com/

[12] http://discuss.it.uts.edu.au/mailman/listinfo/re-online

[13] http://tech.groups.yahoo.com/group/Requirements-Engineering/



Group vs. individual activity in ReqElic - Requirements are identified as an individual activity, by a single BoRA, working alone

in all our projects	4	8%
in most of our projects	9	17%
in some of our projects	24	45%
in none of our projects	16	30%

Group vs. individual activity in ReqElic - Requirements are identified as an individual activity, by more than one BoRA, each working alone

in all our projects	2	4%
in most of our projects	6	11%
in some of our projects	28	53%
in none of our projects	17	32%

Group vs. individual activity in ReqElic - Requirements are identified as a group activity

in all our projects	12	23%
in most of our projects	17	32%
in some of our projects	15	28%
in none of our projects	9	17%

Fig. 6 Requirements Elicitation Done by Individuals and Groups

Figure 6 shows that requirements elicitation is described as an individual activity, by a single BoRA, working alone in all or most projects by 25 % of the respondents, as an individual activity, by more than one BoRA, each working separately in all or most projects by 15 % of the respondents, and as a group activity in all or most projects by 55 % of the respondents.

The same figure shows additionally, requirements elicitation is described as an individual activity, by a single BoRA, working alone in some through all projects by 70 % of the respondents, as an individual activity, by more than one BoRA, each working separately in some through all projects by 68 % of the respondents, and as a group activity in some through all projects by 83 % of the respondents.

This figure shows also that conversely, requirements elicitation is described as an individual activity, by a single BoRA, working alone in no project by 30 % of the respondents,

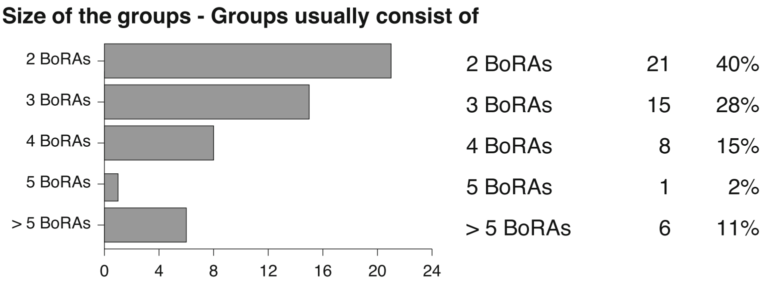


Fig. 7 Sizes of Groups Doing Requirements Elicitation

as an individual activity, by more than one BoRA, each working separately in no project by 32 % of the respondents, and as a group activity in no project by 17 % of the respondents. While both individuals and groups *are* used for requirements elicitation, it appears that groups are *not* used more often than are individuals.

Figure 7 shows that the usual number of BoRAs in a requirements elicitation group for the all or most projects that use groups is given as 2 by 40 %, as 3 by 28 %, as 4 by 8 %, as 5 by 2 %, and as more than 5 by 11 % of the respondents. Thus, groups of sizes 2 and 3 comprise 68 %, a majority, of the groups.

It seems that BoRAs in industry have noticed that smaller is better in forming groups for requirements elicitation, even without the benefit of controlled experiments. Moreover, they are even forming small groups consciously, according to what they have noticed. This observation suggests that the synergy–overhead balance observed for POEPMcreate may be applicable to other CETs and to requirements elicitation in general.

12 Related Work

This section describes older related work of several kinds, about creativity in general and definitions of creativity; creativity in software engineering, business and product planning, and information systems; creativity in requirements engineering; and CETs, most of which has been reported in our earlier publications with other authors (Mich et al. 2005, 2006, 2010; Sakhnini et al. 2012). It then reports work on collaboration in RE and more recent, mostly empirical studies that test the effectiveness of CETs in the context of requirements elicitation.

Other related work concerning

- how to measure the effectiveness of a CET is cited in Section 3.3 and
- how smaller groups have been shown to be more effective than larger groups for other CETs, including brainstorming, is cited in Sections 1.1, 8.2, and 10.

12.1 Creativity in General and Definitions of Creativity

Of course, creativity was not invented for the purpose of being used to develop CBSs; it was the subject of numerous studies in the context of general problem solving (e.g., Poincaré 2001; Fromm 1959; Simon and Newell 1972; Rickards 1974; Leigh 1983; Binnig 1989; de Bono 1985, 1993) before the advent of computing and even now. Nevertheless, creativity is difficult to define, because it plays a role in each of technical innovation, teaching, business, the arts and sciences, and many other fields. Consequently, there are many definitions of creativity, with each field having its own, and with some fields having several competing definitions (Runco 2007). However, fortunately there appears to be convergence among researchers on a concept of creativity related to problem solving (e.g., Rickards 1974; Leigh 1983), encompassing also problem finding and solution thinking. Creativity, in general, is the ability of an individual or a group to think of new and useful ideas (Simonton 1988; Amabile 1988; Feist 1993; Runco 2007). In particular, creativity is understood as the generation of innovative, unexpected solutions to complex, non-trivial problems, or to ill-formed, wicked problems (Rittel and Webber 1973), whose very definition is part of the problem itself.

12.2 Creativity in Software Engineering, Business and Product Planning, and Information Systems

While creativity plays an important role throughout the development of a CBS, it plays an essential role in the lead-up to the development of the CBS, during the identification and definition of the business problem to be solved by building a CBS and during the elicitation, analysis, and specification of the requirements for the CBS to solve the business problem. Some authors describe creativity's multiple roles during the whole CBS development process (Glass 1995; McBreen 2001; Glass and DeMarco 2006). Some authors have investigated creativity's importance in developing business and product plans that solve problems in highly competitive contexts and that address critical business challenges (Geschka 1983; Rickards 1999; LeadershipReview 2016; de Bono and Heller 2010). Some authors have investigated creativity in the context of the development of information systems (Evans 1991; Couger et al. 1993; Couger 1995; Nagasundaram and Bostrom 1995; Couger 1996; Sweeney 2003); the study by Nagasundaram and Bostrom included an empirical validation (Nagasundaram and Bostrom 1995).

12.3 Creativity in Requirements Engineering

Other authors have investigated creativity's importance to RE, particularly for discovering and inventing requirements during elicitation of requirements for CBSs (Gause and Weinberg 1989; Goguen 1993; 1994; Nguyen et al. 2000; Maiden and Gizikis 2001; Browne and Rogich 2001; Robertson 2001; 2002; Robertson and Maiden 2002; Cybulski et al. 2003; Mavin and Maiden 2003; Maiden et al. 2004; Hoffmann et al. 2005; Maiden et al. 2006; Schlosser et al. 2008; Nguyen and Shanks 2009; Ocker 2010; Ficalora and Cohen 2010; Mahaux et al. 2013). In fact, about 90 % of these investigations of creativity in RE focus on requirements elicitation, the step in which requirement ideas are generated. Of these, the studies by Browne and Rogich and by Mavin and Maiden are empirical (Browne and Rogich 2001; Mavin and Maiden 2003).

To understand the role of creativity in RE, it is useful to note that RE itself has all the characteristics of a wicked problem that calls for creativity, in particular when many stakeholders, each with a different view, are involved (Gause and Weinberg 1989; Mullery 1996). Some see creativity as an enhancing factor in RE for CBSs, particularly in helping to identify, indeed invent, new requirements (Glass 1995; Nagasundaram and Bostrom 1995; Couger 1996; McBreen 2001; Robertson 2002; Robertson and Maiden 2002; Maiden et al. 2006; Glass and DeMarco 2006; Zachos and Maiden 2008), that excite the users

(Kano et al. 1984; Ficalora and Cohen 2010). On the other hand, some see creativity in RE as a threat, as something to monitor and control carefully in order to prevent it from compromising their projects (Nguyen et al. 2000). That is, new requirement ideas discovered after implementation has started can be very expensive to accommodate and are often not appreciated by the implementers. Ironically, this danger is a problem only if the creative invention happens not during RE, but after a requirements specification has been delivered and implementation has started.

Creativity gives the hope of attacking what some consider to be the most difficult problem in RE, that of discovering missing requirements (The Standish Group 1994; Al-Ani et al. 1998; Wiegers 2001). The source of many a disaster is a real-world situation that the CBS involved was not prepared to handle because no one thought of the situation, what Don Gause has called “Nature’s Last Laugh” (Gause 2000). Creativity gives the hope that more of these otherwise overlooked situations will be identified before the CBS is built.

More recently, some have applied role playing with scenarios in an attempt to bring more creativity to requirements elicitation (Mavin and Maiden 2003) and to Joint Application Development (JAD) (Wood and Silver 1999). Still others have set up workshops that integrate creativity provocation with use-case and system-context modeling for eliciting requirements, e.g., for air-traffic control (Maiden et al. 2004, 2004). Finally, there is a recognition that in spite of all the tools and technology available to help requirements elicitation, the requirement engineer’s problem-solving skills are key to defining good system requirements and that creativity and imagination are essential components in successful problem solving (Gause and Weinberg 1989; Schenk et al. 1998; Aurum et al. 2001; Gallagher et al. 2004).

Fuller discussions of creativity and of applying these techniques to requirements elicitation, in the specific, and to RE, in general, can be found in overviews elsewhere (etourism Website 2011; Mich et al. 2005; Nguyen and Shanks 2009; Sakhnini et al. 2012; Ocker 2010; Saha et al. 2012; Mahaux et al. 2013).

12.4 Creativity Enhancement Techniques

The most popular CET used for requirements elicitation is brainstorming (Gause and Weinberg 1989), a classical technique which dates back to 1935[14] (Osborn 1953). Many techniques, e.g., brainstorming (Osborn 1953); variations of brainstorming, both manual (Gundy 1984) and computer assisted (Gallupe and Cooper 1993; Ocker 2010); Six Thinking Hats (de Bono 1985); P.a.p.s.a. (Jaoui 1991); Creative Problem Solving (Parnes 1992); and the Creative Pause Technique (de Bono 1993), have been developed to help people be more creative. A characteristic of some of these techniques is that each of them tries to address the problem of identifying and enhancing the viewpoints of all the stakeholders, albeit in a different way. Some of these and other techniques have been applied to RE (Aurum and Martin 1998; Maiden et al. 2004; Ocker 2010), and some, including EPMcreate and POEPMcreate, have also been subjected to empirical validation of their effectiveness in helping requirements elicitation (Aurum and Martin 1998; Jones et al. 2008; Sakhnini et al. 2012). Ocker specifically discusses supporting *group* creativity in the upstream activities, e.g., requirements elicitation, of RE (Ocker 2010). The support includes both

[14] Even though the original work was done in 1935, the work was not published formally until 1953.

software tools and techniques for avoiding the social problems that impede effective group functioning.[15]

12.5 Collaboration in RE

Many requirements elicitation methods assume and promote collaboration of the stakeholders and possibly the analysts. Nearly every CET assumes that a group will collaborate in carrying out the CET's steps. Nevertheless, as mentioned, we are beginning to see evidence that maybe individuals outperform groups in some CETs. Several have noted that RE works with input from a variety of stakeholders and is therefore inherently collaborative (Price and Cybulski 2004, 2006; Damian et al. 2007; Fricker 2010). Others have provided methods, tools, and advice for collaborating, often creatively, in RE (Ang et al. 1998; Boehm et al. 2001; Maiden et al. 2004; Ocker 2010; Mahaux et al. 2013). There have been several empirical studies, all experience reports, of collaborative RE efforts (Mavin and Maiden 2003; Maiden et al. 2007; Jones et al. 2008; Schlosser et al. 2008).

12.6 Empirical Studies of Effectiveness of CETs for RE

While empirical studies tend to be more recent, the older related work cited in the previous subsections includes some empirical studies (Nagasundaram and Bostrom 1995; Aurum and Martin 1998; Browne and Rogich 2001; Mavin and Maiden 2003; Jones et al. 2008; Sakhnini et al. 2012).

The most recent related work is mostly empirical, including experiments, case studies, and systematic reviews. For example, Kauppinen et al. (2007) observed the RE activities of six different commercial software development organizations in Finland. They found three situations in which innovation, and thus creativity, is beneficial for exposing hidden customer and user requirements, inventing new features to satisfy these requirements, and finding innovative solutions to technical problems. Therefore, more research is needed into the application of creativity in RE.

Zachos and Maiden (2008) studied using creativity to address the difficult problem of ensuring completeness of a requirements specification. They describe a parser-based tool, called AnTiQue, that algorithmically retrieves Web services in domains that are analogous to the system whose requirements are being elicited. The paper describes two empirical evaluations of the effectiveness of the tool and its algorithm. The first evaluation compares the tool's recall and precision to those of humans doing the same task on medium-sized problem. The tool's recall was 100 %, i.e., it found all the analogies that the humans did. The second evaluation was to assess the novelty of the requirements human analysts generated after doing walkthroughs of analogies found for a subject domain by AnTiQue and other tools. Here, "novelty" was equated to "dissimilarity" to existing requirements for the domain.

Lemos et al. (2012) conducted a systematic mapping study of creativity in RE in order to find all studies about CETs in RE, to determine what these studies offer to RE research and practice, and to determine the benefits and limitations of these studies. Among the CETs they describe are EPMcreate and POEPMcreate. Their conclusions are that research is needed to provide

[15] Beyond mentioning the overwhelming data showing that individuals are more effective than face-to-face groups at generating ideas with brainstorming, and that one negative social influence is the power of a majority in a group to inhibit a minority in the group, Ocker makes no mention of the effect of group sizes.

- more *empirical* evidence about the effectiveness of all of these CETs,
- tools for enhancing creativity that are integrated into RE tool sets,
- a taxonomy of CETs for RE, and
- guidelines for selecting CETs for each RE phase.

Finally, they suggest that creative thinking needs to be applied more than just during RE, in order that creativity permeate the entire software development lifecycle. Vieira et al. (2012) built on this systematic mapping study to offer a Creativity Pattern Guide that helps a requirements engineer to choose the right CET to apply in any RE situation.

Berntsson Svensson et al. (2015) conducted both (1) a systematic literature review of the use of CETs in RE and (2) an online survey (with a questionnaire) of practitioners about their use of the same. They conclude from these two studies that

- there is insufficient empirical evidence to be able to evaluate whether the CETs actually help generate more creative requirements,
- there is actually only a limited use of CETs in real-life RE.

Our online survey, described in Section 11 found that one CET, brainstorming, *is* used by practicing business or requirements analysts. Of course, since brainstorming is so pervasive, this use of brainstorming could be considered a limited use with respect to more powerful CETs. Berntsson Svensson and Taghavianfar (2015) chose four of the CETs, Hall of Fame, Constraint Removal, brainstorming, and Ideal Box for empirical comparison of their use in creativity workshops conducted for industrial requirements elicitation. They conclude that while brainstorming can generate by far the most raw requirement ideas, Hall of Fame generates the most creative ideas, as measured by the customers, and the largest number of ideas that ended up being requirements for future releases of the product for which the ideas were generated.

13 Conclusions

The research question that this paper attempts to answer is

> In POEPMcreate, how does the number of members of an elicitation group affect the number and quality of requirement ideas generated by the group and by each member?

This question was refined, for the purposes of this paper, into two null hypotheses:

H1 In POEPMcreate, the number of members of an elicitation group has no effect on the quantity and quality of the requirement ideas generated by the group.

H2 In POEPMcreate, the number of members of an elicitation group has no effect on the quantity and quality of the requirement ideas generated on average by each member of the group.

As explained in Section 3.3, we decided to measure the quantity of ideas generated by a group by counting the number of raw ideas generated by the group. We decided also to measure the quality of ideas generated by a group by counting the number of ideas generated by the group that were new with respect to the existing system for which the requirements ideas were generated.

The data from three experiments with identical design and conduct are combined to conclude that when one-, two-, and four-person groups are using POEPMcreate to generate requirement ideas, the null hypotheses can definitely be rejected. Moreover, the data say that a two-person group generates more raw and new requirement ideas, overall and per group member, than a one-person group and than a four-person group.

Because of shortness in the supply of potential subjects, we chose not to test three-person groups. So it is not known how three-person groups fit in these conclusions. Nevertheless, subject to the caveats arising from the threats discussed in the paper, these conclusions can be used to guide staffing of POEPMcreate groups for generation of requirement ideas.

The closer examination of the data in Section 10 to estimate the number of ideas shared by pairs of independently operating groups shows that if there are more than two people available to staff POEPMcreate groups, distributing the people over as many two-person groups as possible is probably the best use of their requirement idea generation power. We noted empirical evidence that in brainstorming, another CET, small group sizes are better and that individuals working alone are better than groups. It is natural to ask for any CET, “What is the optimal group size?”, with the size possibly being one.

The corroborating questionnaire data indicate that industry seems to have come to similar conclusions about CETs and requirements elicitation in general on its own, as a result of good old fashioned observation, (1) that small group sizes are better and (2) that a group size of two is the most popular. More work is needed to study the effect of group size in other CETs and in requirements elicitation in general. The important question to resolve for all of these processes is “What is the tradeoff between group overhead and group synergy?”

Acknowledgments Each of Victoria Sakhnini’s and Luisa Mich’s work was supported in part by a Cheriton School of Computer Science addendum to the same Canadian NSERC–Scotia Bank Industrial Research Chair that was supporting Daniel Berry. Daniel Berry’s work was supported in parts by a Canadian NSERC grant NSERC-RGPIN227055-00 and by a Canadian NSERC–Scotia Bank Industrial Research Chair NSERC-IRCPJ365473-05. The authors thank William Berry for his graciously offered and personal advice on multivariate regressions. All blame for any misapplication of this advice falls on the authors.

Compliance with Ethical Standards This paper is an enhancement of a similarly titled, shorter paper (Sakhnini et al. 2013), by the same authors, published in the *Proceedings of the Workshop on Creativity in Requirements Engineering (CreaRE) at the 18th Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ’2013)*. The workshop paper has been extended by a more detailed description of the techniques, reporting of more data gathered since submitting the workshop paper, a more detailed statistical analysis, and a strengthening of the conclusions.

This paper uses in Sections 1, 2.1 through 2.3, 9, and 12 material copied verbatim from the authors’ and others’ previous papers (Mich et al. 2005; Sakhnini et al. 2012), describing EPMcreate, POEPMcreate, the conduct of the experiment, threats, and related work.

The research conducted with human subjects described in this paper was approved in advance by the University of Waterloo’s Office of Research Ethics. Each subject was given, during his or her Step 1, the approved description of the project and was asked to sign an informed-consent form. The only subjects actually used were those that signed this form.

Conflict of interests There is no known potential or actual conflict of interest.

References

Administrator: Sir John A MacDonald High School Web Site (Viewed 16–20 November 2009 and 7–12 March 2010), http://sja.ednet.ns.ca/index.html

Al-Ani B, Lowe D, Leany J (1998) Incomplete requirements: when requirements go missing. In: Proceedings of the 3rd Australian conference on requirements engineering. Deakin University, Australia

Amabile TM (1988) A model of creativity and innovation in organizations. Res Organ Behav 10:123–167

Ang D, Lim LH, Chan HC (1998) Collaborative requirements engineering: an overview and a proposed integrated model. In: Proceedings of the Thirty-First Hawaii international conference on system sciences, vol 5, pp 355–364

Aurum A, Martin E (1998) Requirements elicitation using solo brainstorming. In: Proceedings of the 3rd Australian conference on requirements engineering. Deakin University, Australia, pp 29–37

Aurum A, Handzic M, Cross J, Toorn CV (2001) Software support for creative problem solving. In: IEEE International conference on advanced learning technologies (ICALT'01), pp 160–162. Madison

Berander P (2004) Using students as subjects in requirements prioritization. In: Proceedings of the international symposium on empirical software engineering (ISESE'04), pp 167–176. IEEE Computer Society

Berntsson Svensson R, Taghavianfar M (2015) Selecting creativity techniques for creative requirements: an evaluation of four techniques using creativity workshops. In: Proceedings of the 23rd IEEE international requirements engineering conference (RE), pp 66–75

Berntsson Svensson R, Taghavianfar M, Gren L (2015) Creativity techniques for more creative requirements: theory vs. practice. In: Proceedings of the forty-first euromicro conference on software engineering and advanced applications (SEAA), pp 104–111

Berry W, Sanders M (2000) Understanding multivariate research: a primer for beginning social scientists. Westview Press, New York

Binnig G (1989) Aus dem Nichts. Über die Kreativität von Natur und Mensch. Piper, München. in German

Briggs RO, Reinig BA, Shepherd MM, Yen J, Nunameker JF Jr (1997) Quality as a function of quantity in electronic brainstorming. In: Hawaii International conference on system sciences, pp 94–103

Boehm B, Grünbacher P, Briggs RO (2001) Developing groupware for requirements negotiation: lessons learned. IEEE Softw 18:46–55

Browne GJ, Rogich MB (2001) An empirical investigation of user requirements elicitation: comparing the effectiveness of prompting techniques. J Manag Inf Sys 17:223–249

Brooks FP (1995) The mythical man-month: essays on software engineering, 2nd edn. Addison-Wesley, Reading

Conboy K, Wang X, Fitzgerald B (2009) Creativity in agile systems development: a literature review. In: Information systems — creativity and innovation in small and medium-sized enterprises, proceedings of the IFIP WG8.2 international conference, CreativeSME 2009. Volume IFIP AICT 301, pp 122–134

Couger JD (1995) Creative problem solving and opportunity finding. Boyd & Fraser, San Francisco

Couger JD (1996) Creativity and innovation in information systems organizations. Boyd & Fraser, San Francisco

Couger JD, Higgens LF, McIntyre SC (1993) (Un)structured creativity in information systems organizations. MIS Q 17:375–398

Cybulski JL, Nguyen L, Thanasankit T, Lichtenstein S (2003) Understanding problem solving. In: Proceedings of the 7th Pacific Asia conference on information systems, pp 465–482. Adelaide

de Bono E (1985) Six thinking hats. Viking, UK

de Bono E (1993) Serious creativity: using the power of lateral thinking to create new ideas. Harper Collins, UK

de Bono E, Heller R (2010) Can creative management techniques help you survive the recession (Viewed 10 August 2010) http://www.thinkingmanagers.com/management/creative-management-techniques

Damian D, Marczak S, Kwan I (2007) Collaboration patterns and the impact of distance on awareness in requirements-centred social networks. In: Proceedings of the 15th IEEE international requirements engineering conference (RE), pp 59–68

Daun M, Salmon A, Bandyszak T, Weyer T (2016) Common threats and mitigation strategies in requirements engineering experiments with student participants. In: Proceedings of the 21st international working conference on requirements engineering: foundation for software quality (REFSQ'2016), pp 269–295

Dean DL, Hender JM, Rodgers TL, Santanen EL (2006) Identifying quality, novel, and creative ideas: constructs and scales for idea evaluation. J Assoc Inf Syst 7

Dennis AR, Valacich JS (1993) Computer brainstorms: more heads are better than one. J Appl Psychol 78:531–537

Dornburg CC, Stevens SM, Hendrickson SML, Davidson GS (2008) LDRD final report for improving human effectiveness for extreme-scale problem solving: assessing the effectiveness of electronic brainstorming in an industrial setting. Technical Report SAND2008-5971, Sandia National Laboratories. http://prod.sandia.gov/techlib/access-control.cgi/2008/085971.pdf

Dow G (2016) Creativity test: creativity assessment packet (Williams, 1980), R546 instructional strategies for thinking, collaboration, and motivation, AKA: Best of Bonk on the Web (BOBWEB). Technical report, Indiana University (Viewed 11 April 2016) http://www.indiana.edu/bobweb/Handout/d16.cap.html

etourism Website (2011) Online bibliographies, click on (1) creativity, (2) business creativity, (3) creativity techniques or (4) brainstorming as a technique for software requirements elicitation (viewed April 2011) http://etourism.economia.unitn.it/bibliographies/?locale=en

Evans JR (1991) Creative thinking in the decision and management sciences. South Western, Cincinnati

Feist GJ (1993) A structural model of scientific eminence. Psychol Sci 4:366–371

Ficalora JP, Cohen L (eds) (2010) Quality function deployment and six sigma, 2nd edn: a QFD handbook. Pearson Education, Boston

Fricker S (2010) Requirements value chains: Stakeholder management and requirements engineering in software ecosystems. In: Proceedings of the working conference on requirements engineering: foundation for software quality (REFSQ), pp 60–66

Fromm E (1959) The creative attitude. In: Anderson H (ed) Creativity and its cultivation. Harper & Row, New York, pp 44–54

Furnham A, Yazdanpanahi T (1958) Personality differences and group versus individual brainstorming. Person Individ Diff 19:73–80

Gallagher K, Mason RM, Vandenbosch B (2004) Managing the tension in IS projects: balancing alignment, engagement, perspective and imagination. In: Proceedings of the 37th Hawaii international conference on system sciences. Honolulu

Gallupe RB, Cooper WH (1993) Brainstorming electronically. Sloan Manag Rev 35:27–36

Gause DC (2000) User DRIVEN design—the luxury that has become a necessity, a works hop in full lifecycle requirements management. ICRE 2000 Tutorial T7, Schaumberg

Gelman A, Hill J, Yajima M (2012) Why we (usually) don't have to worry about multiple comparisons. J Res Educ Effect 5:189–211

Geschka H (1983) Creativity techniques in product planning and development: a view from West Germany. R&D Manag 13:169–183

Givant S, Halmos P (2009) Introduction to Boolean algebras. Springer Science+Business Media, New York

Glass R (1995) Software creativity. Prentice Hall, Englewood Cliffs

Glass RL, DeMarco T (2006) Software Creativity 2.0. developer.* Books, Atlanta

Gause D, Weinberg G (1989) Exploring requirements: quality before design. Dorset House, New York

Gause D, Weinberg G (1990) Are your lights on? How to figure out what the problem REALLY is. Dorset House, New York

Goguen JA (1993) Requirements engineering as the reconciliation of technical and social issues. Technical report, Centre for Requirements and Foundations, Programming Research Group, Oxford University Computing Lab. Modified version later published as (Goguen 1994)

Goguen JA (1994) Requirements engineering as the reconciliation of technical and social issues. In: Requirements engineering, social and technical issues, pp 165–199. Academic Press

Gundy ABV (1984) Managing group creativity. American Management Association, New York

Hoffmann O, Cropley D, Cropley A, Nguyen L, Swatman P (2005) Creativity, requirements and perspectives. Aust J Inf Syst 13:159–174

Isaksen SG, Gaulin JP (2005) A reexamination of brainstorming research: implications for research and practice. Gifted Child Q 40(Fall):315–329

Jaoui H (1991) La Créativité Mode d'Emploi. E.S.F. Editeur – Entreprise Moderne d'Edition – Librairies Techniques. Paris, in French

Jeff, Berg, Mike (2016) 882: Significant (Viewed 11 August 2016) https://www.explainxkcd.com/wiki/index.php/882:_Significant

Jones S, Lynch P, Maiden N, Lindstaedt S (2008) Use and influence of creative ideas and requirements for a work-integrated learning system. In: Proceedings of the 16th IEEE international requirements engineering conference (RE), pp 289–294

Kano N, Seraku N, Takahashi F, Tsuji S (1984) Attractive quality and must-be quality (in japanese). J Japan Soc Qual Control 14:39–48

Kauppinen M, Savolainen J, Männisto T (2007) Requirements engineering as a driver for innovations. In: Proceedings of the 15th IEEE international requirements engineering conference (RE), pp 15–20

Kaufman JC, Sternberg RJ (eds) (2006) The international handbook of creativity. Cambridge University Press, Cambridge

Kohn NW, Smith SM (2011) Collaborative fixation: effects of others' ideas on brainstorming. Appl Cogn Psychol 25:359–371

LeadershipReview (2016) Using creativity to help your business stand out from the crowd (Viewed 11 April 2016) http://www.leadershipreview.net/using-creativity-help-your-business-stand-out-crowd

Leigh A (1983) Decisions, decisions!: a practical management guide to problem solving and decision making. Gower Aldershot, Hampshire

Lemos J, Alves C, Duboc L, Rodrigues GN (2012) A systematic mapping study on creativity in requirements engineering. In: Proceedings of the 27th Annual ACM symposium on applied computing (SAC), pp 1083–1088

Maiden N, Gizikis A (2001) Where do requirements come from? IEEE Softw 18:10–12

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

Maiden N, Manning S, Robertson S, Greenwood J (2004) Integrating creativity workshops into structured requirements processes. In: Proceedings of the conference on designing interactive systems (DIS'2004), pp 113–122. Cambridge

Maiden N, Robertson S, Robertson J (2006) Creative requirements: Invention and its role in requirements engineering. In: Proceedings of the 28th international conference on software engineering (ICSE), pp 1073–1074

Maiden N, Ncube C, Robertson S (2007) Can requirements be creative? experiences with an enhanced air space management system. In: Proceedings of the 29th international conference on software engineering, pp 632–641

Mahaux M, Nguyen L, Gotel O, Mich L, Mavin A, Schmid K (2013) Collaborative creativity in requirements engineering: Analysis and practical advice. In: Proceedings of the 7th IEEE international conference on research challenges in information science (RCIS), pp 1–10

Mavin A, Maiden N (2003) Determining socio-technical systems requirements: experiences with generating and walking through scenarios. In: Proceedings of the 11th IEEE International requirements engineering conference, pp 213–222. Monterey Bay

McBreen P (2001) Tutorial 38: creativity in software development. In: OOPSLA 2001. Tampa Bay

Mich L, Anesi C, Berry DM (2005) Applying a pragmatics-based creativity-fostering technique to requirements elicitation. Require Eng J 10:262–274

Mich L, Berry DM, Franch M (2006) Classifying web-application requirement ideas generated using creativity fostering techniques according to a quality model for web applications. In: Proceedings of the 12th international workshop requirements engineering: foundation for software quality, REFSQ' 06

Mich L, Berry DM, Alzetta A (2010) Individual and end-user application of the EPMcreate creativity enhancement technique to website requirements elicitation. In: Proceedings of the workshop on creativity in requirements engineering (CreaRE) at REFSQ'2010

Mich L, Sakhnini V, Berry DM (2012) Requirements elicitation (ReqElic) in my company. Technical report, University of Trento (Deployed 31 August 2012) https://docs.google.com/spreadsheet/viewform?formkey=dFI2UWx0MWJuRUdvQ1JNZnh1NFN0SGc6MQ

Mullery G (1996) The perfect requirements myth. Requir Eng J 1:132–134. also at http://link.springer.com/content/pdf/10.1007

Nagasundaram M, Bostrom RP (1995) Structuring creativity with GSS: an experiment. In: Proceedings of the Americas conference on information systems. Paper 145, http://aisel.aisnet.org/amcis1995/145/

Niknafs A, Berry DM (2016) The impact of domain knowledge on the effectiveness of requirements engineering activities. Empirical Software Engineering Online First. http://link.springer.com/article/10.1007/s10664-015-9416-2

Nguyen L, Shanks G (2009) A framework for understanding creativity in requirements engineering. J Inf Softw Technol 51:655–662

Nguyen L, Carroll J, Swatman PA (2000) Supporting and monitoring the creativity of IS personnel during the requirements engineering process. In: Proceedings of the 33rd Hawaii international conference on system sciences. HICSS-33, Maui, http://csdl2.computer.org/comp/proceedings/hicss/2000/0493/07/04937008.pdf

Ocker RJ (2010) Promoting group creativity in upstream requirements engineering. Human Technol Interdiscip J Humans ICT Environ 6:55–70

Osborn A (1953) Applied imagination. Charles Scribner's, New York

Parnes S (1992) Source book for creative problem solving. Creative Foundation, USA

Poincaré H (2001) Science and method. Key Texts, South Bend. Originally published in 1914

Preparata FP, Yeh RTY (1973) Introduction to discrete structures for computer science and engineering. Addison-Wesley Longman, Boston

Price J, Cybulski JL (2004) Influence of stakeholder communication on consensus making in requirements negotiation. In: Proceedings of the 11th Austrlian workshop on requirements engineering (AWRE)

Price J, Cybulski JL (2006) The importance of IS stakeholder perspectives and perceptions to requirements negotiation. In: Proceedings of the 11th Austrlian workshop on requirements engineering (AWRE)

Rickards T (1974) Problem solving through creative analysis. Gower, New York

Rickards T (1999) Creativity and the management of change. Blackwell, Oxford

Rittel H, Webber M (1973) Dilemmas in a general theory of planning. Polic Sci 4:155–169

Robertson S (2001) Requirements trawling: techniques for discovering requirements. Int J Human-Comput Stud 55:405–421

Robertson J (2002) Eureka! Why analysts should invent requirements. IEEE Softw 19:20–22

Robertson S, Maiden N (2002) Tutorial notes T08: creativity, the path to innovative requirements. In: IEEE Joint international requirements engineering conference. Essen

Runco MA (2007) Creativity: theories and themes: research, development, and practice. Elsevier Academic Press, Burlington

Saha SK, Selvi M, Büyükcan G, Mohymen M (2012) A systematic review on creativity techniques for requirements engineering. In: Proceedings of the international conference on informatics, electronics vision (ICIEV), pp 34–39

Sakhnini V, Mich L, Berry DM (2012) The effectiveness of an optimized EPMcreate as a creativity enhancement technique for Website requirements elicitation. Require Eng J 17:171–186

Sakhnini V, Mich L, Berry DM (2013) On the sizes of groups using the full and optimized EPMcreate creativity enhancement technique for Web site requirements elicitation. In: Proceedings of the workshop on creativity in requirements engineering (CreaRE) at REFSQ'2013, pp 23–38. http://www.icb.uni-due.de/fileadmin/ICB/research/research_reports/ICB-Report-No56.pdf

Sakhnini V, Mich L, Berry DM (2016) Group versus individual use of an optimized and the full EPMcreate as creativity enhancement techniques for web site requirements elicitation. Technical report, School of Computer Science, University of Waterloo (Viewed 11 April 2016) http://se.uwaterloo.ca/dberry/FTP_SITE/tech.reports/SakhniniMichBerryTR.pdf

Sakhnini V, Berry DM, Mich L (2016) Materials for comparing POEPMcreate, EPMcreate, and brainstorming. Technical report, School of Computer Science, University of Waterloo (Viewed 11 April 2016). http://se.uwaterloo.ca/dberry/FTP_SITE/software.distribution/EPMcreateExperimentMaterials/

Salzer H, Levin I (2004) Atomic requirements in teaching logic control implementation. Int J Eng Educ 20:46–51

Schenk KD, Vitalari NP, Davis KS (1998) Differences between novice and expert systems analysts: what do we know and what do we do? J Manag Inf Syst 15:9–50

Schlosser C, Jones S, Maiden N (2008) Using a creativity workshop to generate requirements for an event database application. In: Proceedings of the international workshop requirements engineering: foundation for software quality REFSQ'08. LNCS, vol 5025. Springer, Berlin, pp 109–122

Simon H, Newell A (1972) Human problem solving. Prentice Hall, Englewood Cliffs

Simonton DK (1988) Scientific genius: a psychology of science. Cambridge University Press, Cambridge

Sweeney RB (2003) Creativity in the information technology curriculum proposal. In: Proceedings of the 4th conference on information technology curriculum, CITC4'03, pp 139–141. Lafayette

Taylor DW, Berry PC, Block CH (1958) Does group participation when using brainstorming facilitate or inhibit creative thinking? Admin Sci Q 3:23–47

The Standish Group (1994) The CHAOS report. Technical report, The Standish Group

Vieira ER, Alves C, Duboc L (2012) Creativity patterns guide: Support for the application of creativity techniques in requirements engineering. In: Proceedings of the 4th international conference on human-centered software engineering (HCSE), pp 283–290. Springer-Verlag

von Bertalanaffy L (1976) General systems theory: foundations, development, applications. revised edn. George Braziller, New York

West Side School District (2016) Gifted and talented program. Technical report, West Side Public Schools, Higden, AR, U.S.A. (Viewed 11 April 2016) http://wseagles.k12.ar.us/GT.pdf

Wiegers KE (2001) Inspecting requirements. Technical report, StickyMinds.com Original Column. http://www.stickyminds.com/se/S2697.asp

Williams F, Taylor CW (1966) Instructional media and creativity. In: Proceedings of the 6th Utah creativity research conference. New York, Wiley

Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2000) Experimentation in software engineering: an introduction. Kluwer Academic Publishers, Norwell

Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in software engineering. Springer, Heidelberg

Wood J, Silver D (1999) Joint application development. Wiley, New York

Zachos K, Maiden N (2008) Inventing requirements from software: An empirical investigation with web services. In: Proceedings of the 16th IEEE international requirements engineering conference (RE), pp 145–154

Zhou J (ed) (2016) The Oxford handbook of creativity, innovation, and entrepreneurship. Oxford Library of Psychology Series. Oxford University Press, Oxford

Victoria Sakhnini is a lecturer at the School of Computer Science at the University Of Waterloo. Her career spans 22 years as an educator. She has over 18 years of experience teaching high school computer science AP courses. She also served as a senior assessor for the Israeli Minister of Education in the CS AP Exams Assessment.

For 10 years, Victoria worked with the Israeli National Center for CS Teachers and conducted training classes for CS teachers to guide them through the best practices for teaching the new curriculum, developing new teaching methods and tools for certain topics, etc.

In addition to her BSc in Computer Science (1990), Victoria earned a PhD in Computer Science Education (2006) from the Technion, where she worked on: systems approach in teaching, integrating models in education, evaluation of educational projects, curriculum planning, innovations in environmental education and assessments, and seminars in learning, research, and assessment.

Since 2009, Victoria's research interests are software engineering in general and innovative creativity fostering techniques in particular. More information about Victoria's work can be found at https://cs.uwaterloo.ca/~vsakhnin/.

Luisa Mich is an Associate Professor of of Computer Science at the University of Trento, Italy. Her research interests include requirements engineering, creativity and web strategies. She is an author of more than 150 papers that have appeared in journals, conferences, and workshops. See http://www4.unitn.it/Ugcvp/en/Web/ProdottiAutore/PER0001016 for a complete list. She serves and has served on the organizing and program committees of several conferences and workshops, including NLDB, ENTER, REFSQ, and RE.

She has lectured at and collaborated with several Italian and foreign universities.

Luisa Mich is a member of the IEEE Computer Society, of the Association for Computing Machinery (ACM); of the International Federation for Information Technology and Tourism (IFITT), for which she is a member of the Italian Chapter's Committee; and of the AICA (Associazione Italiana per l'Informatica ed il Calcolo Automatico). She has been a board member of AICA-CINI-CRUI, the National Observatory for Computer Science Certifications. She has been an initiator of many didactic initiatives for introducing computer science to different degrees, both technical and non technical, at the University of Trento.

Daniel M. Berry got his B.S. in Mathematics from Rensselaer Polytechnic Institute, Troy, New York, USA in 1969 and his Ph.D. in Computer Science from Brown University, Providence, Rhode Island, USA in 1974. He was on the faculty of the Computer Science Department at the University of California, Los Angeles, California, USA from 1972 until 1987. He was in the Computer Science Faculty at the Technion, Haifa, Israel from 1987 until 1999. From 1990 until 1994, he worked for half of each year at the Software Engineering Institute at Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, where he was part of a group that built CMU's Master of Software Engineering program. During the 1998–1999 academic year, he visited the Computer Systems Group at the University of Waterloo in Waterloo, Ontario, Canada. In 1999, Berry moved to what is now the the Cheriton School of Computer Science at the University of Waterloo. Between 2008 and 2013, Berry held an Industrial Research Chair in Requirements Engineering sponsored by Scotia Bank and the National Science and Engineering Research Council of Canada (NSERC). Prof. Berry's current research interests are software engineering in general, and requirements engineering and electronic publishing in the specific. For more details see https://cs.uwaterloo.ca/~dberry/resume.db.pdf.