Group Versus Individual Use of an Optimized and the Full EPMcreate as Creativity Enhancement Techniques for Web Site Requirements Elicitation *

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

[1] Cheriton School of Computer Science, University of Waterloo
Waterloo, ON, N2L 3G1 Canada
vsakhnini@gmail.com, dberry@uwaterloo.ca

[2] Department of Industrial Engineering, University of Trento
I-38122 Trento, Italy
luisa.mich@unitn.it

Abstract. [Context] Creativity is often needed in requirements elicitation, i.e., generating ideas for requirements, and therefore, techniques to enhance creativity are believed to be useful. **[Objective]** How does the size of a group using the EPMcreate creativity enhancement technique or an optimization of it, POEPMcreate, affect the group's and each member of the group's effectiveness in generating requirement ideas? **[Method]** 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. **[Results]** The data of this experiment combined with the data of two previous experiments involving two-person and four-person groups using EPMcreate and POEPMcreate indicate that the size of a group using EPMcreate or POEPMcreate does affect the number of raw and new requirement ideas generated by the group and by the average member of the group. **[Conclusions]** A conclusion from the data is that generally, the larger a group is, up to a particular group size that depends on the creativity enhancement technique used, the more raw and new requirement ideas it generates. After that particular group size, the larger a group is, the fewer raw and new requirement ideas it generates. Another conclusion from the data is that generally, the larger a group is, up to a particular group size that depends on the creativity enhancement technique used, the more raw and new requirement ideas the average group member generates. After that particular group size, the larger a group is, the fewer raw and new requirement ideas he or she generates. These conclusion are partially corroborated by qualitative data gathered from a survey of professional business or requirements analysts about group sizes and creativity enhancement techniques.

* 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.

1 Introduction

Many have observed the importance of creativity in requirements engineering, particularly for discovering and inventing requirements during elicitation of requirements for computer-based systems (CBSs) [1–11], for those solving wicked problems, for those in highly competitive contexts, for those addressing critical business challenges, and for Web sites with requirements for high quality [12–16].

Creativity is difficult to define, because it plays a role in technical innovation, teaching, business, the arts and sciences, and many other fields, and each field has its own definition [17]. Creativity, in general, is the ability of an individual or a group to think of new and useful ideas [18–20]. Many techniques, e.g., brainstorming [21], Six Thinking Hats [22], and the Creative Pause Technique [23], have been developed to help people be more creative. Some of these techniques have been applied to requirements engineering [24, 7], and some have also been subjected to experimental validation of their effectiveness [24–26]. A fuller discussion of creativity and of applying these techniques to requirements elicitation can be found elsewhere [27, 28, 26].

This paper investigates the use of an optimized and the full *EPMcreate (Elementary Pragmatic Model Creative Requirements Engineering [A] TEchnique)* [28, 15] creativity enhancement techniques (CET) to help in generating ideas for requirements for Web sites. The optimization is called the Power-Only EPMcreate (POEPMcreate).

The feasibility of applying POEPMcreate and the full EPMcreate to help idea generation in requirements elicitation was established by earlier experiments [28, 15, 26]. 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 do 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.

These experiments exposed a number of issues to be explored in the future. These include the question that is taken as the research question of this paper:

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

The purpose of the research leading to this paper is to answer this question by conducting experiments in the context of eliciting requirements for a high school's Web site. 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 legitmate 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. The need for transforming some of the data is discovered. That transformation is done, and the support for the hypotheses is reconsidered. 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. 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 Full and Optimized EPMcreate Techniques

The explanation of EPMcreate given here is abbreviated to what is necessary to understand this paper. A fuller description of EPMcreate can be found elsewhere [28, 26].

2.1 Basic, Full 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

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

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$\wedge$SH2, represents the analyst's looking for ideas that agree with SH1's viewpoint and with SH2's viewpoint.
- $f2 =$ SH1 $\wedge$ $\neg$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.
$f5 =$ 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.

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. See Section 2.4 for an alternative, direct, way to deal with three types of 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. EPMcreate is by no means the only technique for identifying requirements; it is but one of many that can be used. 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

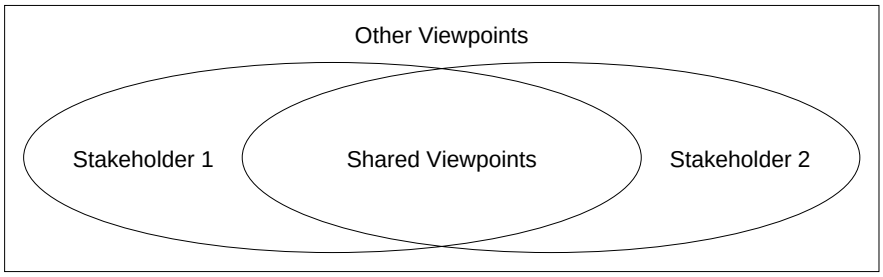


Fig. 1: Venn Diagram of Two Stakeholders' Viewpoints

diagram that the leader will show 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.

While showing the diagram of Figure 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 e-learning platform, the clear stakeholders are the student and the teacher. 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. See Section 2.4 for an alternative, direct, way to deal with three types of stakeholders.

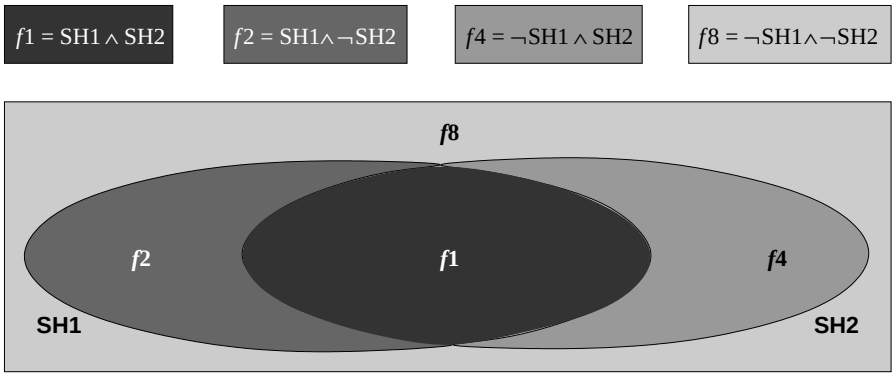


Fig. 2: The Four Steps of the Optimization and the Four Regions of the Venn Diagram

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 include 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 [26], to be more effective than the full EPMcreate, because as illustrated by Figure 2, the Boolean function of each of the power-of-two steps corresponds to exactly one of the four regions of Figure 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. In other words, these four regions are the four atoms of the 16-element free Boolean algebra that is generated from the two stakeholder viewpoints [29, 30].

2.4 POEPMcreate for Three Stakeholders

An EPMcreate for three types of stakeholders would require 256 steps[3], which are more than the 48 steps that are required for doing EPMcreate three[4] times to cover the logical combinations of viewpoints of all possible pairings of two out of three stakeholders. However, as observed by Ingrid Giel in personal communication, a POEPMcreate for three stakeholders would require only eight[5] steps, namely those corresponding to the powers of two that are less than 256, just as the two-stakeholder POEPMcreate requires only four steps, namely those corresponding to the powers of two that are less than 16. Moreover, these eight steps are fewer than the 12 steps required to do the 4-step

[3] $2^{2^3} = 256$ is the number of Boolean functions of three variables [30].

[4] $\binom{3}{2} = 3$

[5] The Venn diagram of three fully overlapping ellipses has eight regions.

POEPMcreate three times for all three pairings of three stakeholders' viewpoints. In addition, probably considering three viewpoints at one time will expose ideas that considering all possible pairings of three viewpoints will not find. The effectiveness of this method will have to be tested empirically.

3 Experimental Design

As mentioned, the feasibility of applying the full EPMcreate as a CET was established by experiments conducted by us with Anesi [28, 26]. The effectiveness of POEPMcreate as a CET and as an *improvement* over EPMcreate was established by two identically designed experiments, Experiments 1 and 2, conducted by us [26].

This paper describes

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

to answer the research question mentioned in Section 1, which is to determine whether the number of members of an elicitation group, using EPMcreate or POEPMcreate as a CET, affects the number of requirement ideas generated by the group and by each member.

This section describes the experimental design that was used in all three experiments, independent of any particular experiment. The next section 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 previous experiments [28]. 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

The research question to be answered by this paper is:

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

We could not predict with confidence any answer to the question. It did seem reasonable that the number of requirement ideas per group would be smaller with fewer members, but we really had no idea how the number of ideas per member would be affected by the number of members in a group. On one hand, it might be that with fewer members in a group, each member would have more time to generate more ideas; it might even be in this case, that the number of additional ideas per member is enough that the number of ideas per group is as large or even larger than with more members. On the other hand, it might be that with more members in a group, the group's management overhead goes up and reduces the number of ideas generated by each member of the group. Back on the first hand, it might be that fewer members in a group reduces

the synergistic effect that members are supposed to have on others' generation of ideas [21, 31]. Unpredictable is the effect of a group's avoidance of duplicate ideas on both the number of ideas per group and the number of ideas per member. Mich, Alzetta, and Berry [32] 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 each of EPMcreate and 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 each of EPMcreate and 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 [33]. 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,

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 wrote 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 [34].

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 [7, 35, 25, 11, 36]. Beyond newness, there is no universal agreement. Many definitions stress also usefulness [25, 11]. Other characteristics that have been mentioned include applicability, effectiveness, implementability, non-obviousness, originality, relevance, realizability, specificity, thoroughness, usefulness, workability, [28, 35].

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 [37]. Moreover, when we were doing 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 [26]. Combining the results of multiple experiments requires following this decision in all experiments.

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 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.

3.4 Refining Hypotheses into Subhypotheses

The two hypotheses H1 and H2 may be refined into eight different subhypotheses, each one about the CET applied by a group, taking the numbers of raw and new requirement ideas produced either by the whole group or on average by a member of the group[6]. The eight subhypotheses are, therefore:

H1:

[6] 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.

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

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

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.

H2:

HEAR: The number of members of an elicitation group using **EPMcreate** has no effect on the **average number of requirement ideas per group member** of **raw** requirement ideas generated.

HEAN: The number of members of an elicitation group using **EPMcreate** has no effect on the **average number of requirement ideas per group member** of **new** requirement ideas generated.

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 rest of the paper, an individual working alone to generate requirement ideas using a CET 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,

Step 2: 30 minutes for each subject to take an adult version of Frank Williams's Creativity Assessment Packet [38], hereinafter called the *Williams test* to measure the subject's individual[7] creativity.

[7] 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 [39].

Step 3: 10 minutes for us to deliver to all groups an explanation about the experiment and the CET, EPMcreate or POEPMcreate, that they were to use. The explanation of EPMcreate was basically the second paragraph of Section 2.2 of this paper, and the explanation of POEPMcreate was basically the same paragraph, but using only Function Steps 1, 2, 4, and 8. Note that all the groups in any session used the same CET so that one explanation sufficed.

Step 4: 120 minutes for each group to carry out its requirements elicitation session using the assigned CET, EPMcreate or 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 [40].

3.6 Recruiting and Assigning Subjects into Balanced Groups

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

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 the superiority of POEPMcreate over 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.

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 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 (reproduced from the paper about Experiments 1 and 2 [26]). The average Williams test scores for the six groups in Experiment 1 were in the range from 70.25 to 71.6 out of a possible 100. The average Williams test scores for the 8 groups in Experiment 2 were in the range from 59.5 to 82; in particular, the average of the average Williams test score for the four groups that used EPMcreate technique was 74 while the average of the average Williams test score for the four groups that used POEPMcreate technique was 72.875. There was no way to form groups with closer average Williams test scores without being unbalanced 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[8]. In the first round of Experiment 3, only 15 students replied, 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, 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.

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.

[8] 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 2: Characteristics of Groups of Experiment 1 and Their Subjects [26]

Group	Technique	# 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
1	POEPMcreate	3	1	1	3	2	2	2	2	25.5	70.66
2	POEPMcreate	2	2	2	2	2	2	3	1	23.8	71.00
3	EPMcreate	2	2	2	2	1	3	3	1	21.5	70.75
4	EPMcreate	3	1	1	3	2	2	1	3	23.4	70.60
5	Brainstorming	4	0	3	1	1	3	1	3	20.2	71.60
6	Brainstorming	2	2	2	2	3	1	3	1	25	70.25

Table 3: Characteristics of Groups of Experiment 2 and Their Subjects [26]

Group	Technique	# 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
A	EPMcreate	1	1	0	2	2	0	2	0	26	82
B	EPMcreate	0	2	1	1	0	2	0	2	25	72.5
C	EPMcreate	1	1	0	2	0	2	0	2	24	59.5
D	EPMcreate	1	1	0	2	1	1	2	0	24	82
E	POEPMcreate	1	1	0	2	0	2	2	0	30.5	75.5
F	POEPMcreate	2	0	0	2	0	2	0	2	25	80.5
G	POEPMcreate	0	2	0	2	1	1	1	1	24	72.5
H	POEPMcreate	2	0	0	2	2	0	2	0	26	63

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 professionally	# not worked professionally	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

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 every group that participated in one of the three experiments. The structure of the table is explained columnwise and then rowwise.

The first three columns give a group's characteristics:

1. its experiment,
2. its assigned CET, and
3. 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 "Scaled", are data whose need is explained in Section 8.2 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 two letters of its name, i.e., "TR", "AR", "TN", and "AN", matches the header of the column. Under the two-letter header of a column is a description of the data displayed in the column.

Rowwise, between each pair of triple rules is a section of the table consisting of rows about one CET, EPMcreate or POEPMcreate. In each section, separated by double rules are subsections, each about one size of group doing the CET of the containing section. At the head of each data column in each CET's subsection is the two-letter name of the entire data column prepended with an "E" or "P", for EPMcreate or POEPMcreate

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

			Original				Scaled			
			TR	AR	TN	AN	TR	AR	TN	AN
Exp #	Assigned CET	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
			ETR	EAR	ETN	EAN	ETR	EAR	ETN	EAN
1	EPMcreate	4	63	15.75	62	15.5	34.02	8.505	33.48	8.37
1	EPMcreate	4	60	15	56	14	32.4	8.1	30.24	7.56
Avg	EPMcreate	4	61.5	15.38	59	14.75	33.21	8.3025	31.86	7.965
2	EPMcreate	2	30	15	24	12	39.9	19.95	31.92	15.96
2	EPMcreate	2	35	17.5	26.5	13.25	46.55	23.28	35.25	17.62
2	EPMcreate	2	36	18	30	15	47.88	23.94	39.9	19.95
2	EPMcreate	2	40	20	21	10.5	53.2	26.6	27.93	13.97
Avg	EPMcreate	2	35.25	17.63	25.38	12.69	46.8825	23.4425	33.75	16.875
			PTR	PAR	PTN	PAN	PTR	PAR	PTN	PAN
1	POEPMcreate	4	74	18.5	70.5	17.625	39.96	9.99	38.07	9.5175
1	POEPMcreate	4	76	19	70.5	17.625	41.04	10.26	38.07	9.5175
3	POEPMcreate	4	40	10	36.5	9.125	40	10	36.5	9.125
3	POEPMcreate	4	40	10	35.5	8.875	40	10	35.5	8.875
3	POEPMcreate	4	44	11	36	9	44	11	36	9
3	POEPMcreate	4	38	9.5	28	7	38	9.5	28	7
Avg	POEPMcreate	4	52	13	46.17	11.54	40.5	10.125	35.3567	8.8392
2	POEPMcreate	2	40	20	32.5	16.25	53.2	26.6	43.23	21.613
2	POEPMcreate	2	42	21	32	16	55.86	27.93	42.56	21.28
2	POEPMcreate	2	45	22.5	36	18	59.85	29.93	47.88	23.94
2	POEPMcreate	2	63	31.5	51.5	25.75	83.79	41.9	68.5	34.248
3	POEPMcreate	2	66	33	46	23	66	33	46	23
3	POEPMcreate	2	30	15	20.5	10.25	30	15	20.5	10.25
3	POEPMcreate	2	90	45	68.5	34.25	90	45	68.5	34.25
3	POEPMcreate	2	67	33.5	57.5	28.75	67	33.5	57.5	28.75
Avg	POEPMcreate	2	55.38	27.69	43.06	21.53	63.2125	31.6075	49.3338	20.7914
3	POEPMcreate	1	27	27	19.5	19.5	27	27	19.5	19.5
3	POEPMcreate	1	30	30	29	29	30	30	29	29
3	POEPMcreate	1	18	18	18	18	18	18	18	18
3	POEPMcreate	1	18	18	17	17	18	18	17	17
3	POEPMcreate	1	27	27	15.5	15.5	27	27	15.5	15.5
Avg	POEPMcreate	1	24	24	19.8	19.8	24	24	19.8	19.8

CET, respectively, to serve as the header for the data column in the CET's subsection. 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.

At the end of a subsection, about one size of group doing the CET of the containing section, comes a row whose first column says "Avg", that gives data-column-by-data-column, the average of the subsection's data for the column.

Figures 3 and 4 plot the "Original" data of Table 5, but not correspondingly. Specifically, Figure 3 shows a graph plotting the numbers of raw and new requirement ideas generated for both CETs by entire groups in all three experiments while Figure 4 shows a graph plotting the numbers of raw and new requirement ideas generated on average for both CETs by each member of groups, again from all three experiments. In each of these graphs, there is a thick vertical line between two pairs of bars. The bars to the left of the vertical line are about POEPMcreate, and the bars to the right are about EPM-create. From these graphs, it is already apparent that in some ways, a two-person group outperforms a four-person group.

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 essentially identical to those of Experiments 1 and 2 [26]. 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 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

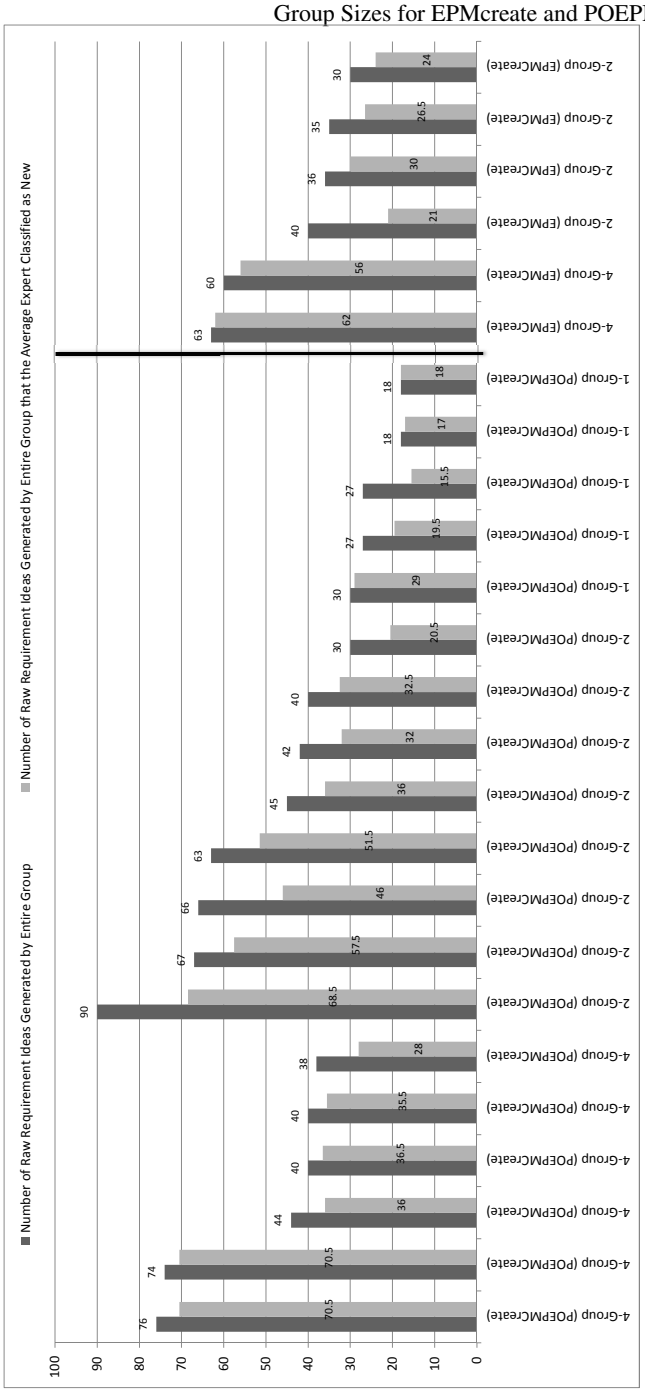


Fig. 3: Numbers of Raw and New Requirements Ideas Generated by Entire Groups

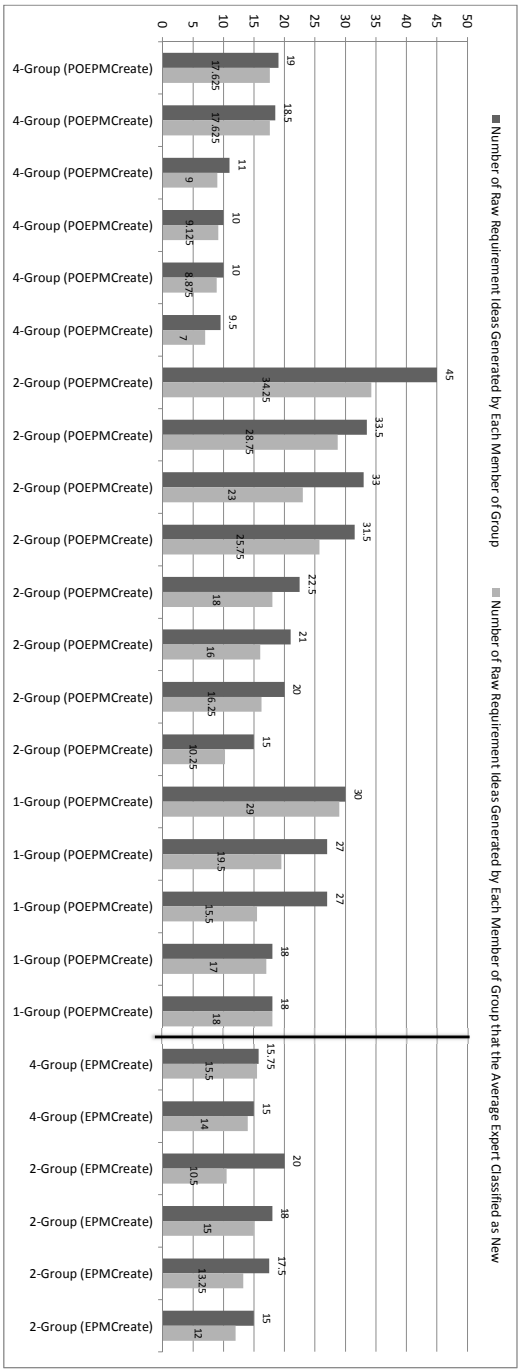


Fig. 4: Numbers of Raw and New Requirements Ideas Generated by Each Member of Groups

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 dispell 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 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, Figure 5 shows a linear regression for the number of raw requirement ideas generated per group as a function of group size. Figure 6 shows that the overall linear regression, depicted as a solid line, is very different from the regression, depicted as a dashed line, for each of the three possible 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 [41] to compute the coefficients of the effect on dependent variables of changes in the independent variables and to compute their statistical significance.

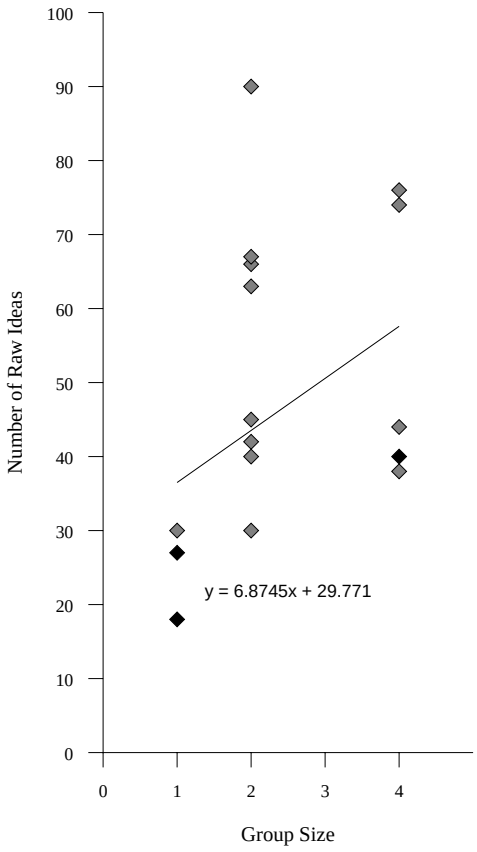


Fig. 5: Linear Regression for PTR Data

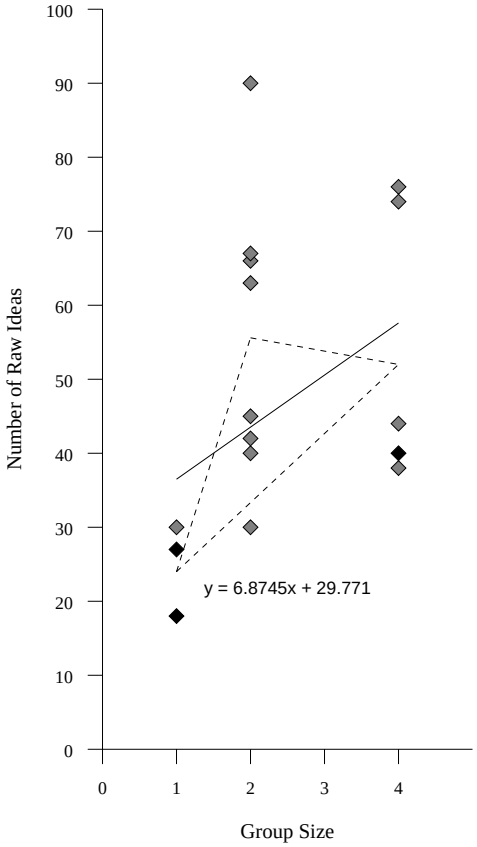


Fig. 6: Point-by-Point Linear Regression for PTR Data

7.1 Regressions for POEPMcreate

For the POEPMcreate data, we decided to run for each group's values of each dependent variable, PTR, PTN, PAR, and PAN, a regression of it on the group's values for each of the following independent numerical variables: the group's average Williams test score, crt in the range [0 – 100], and the group's size as a numerical variable, sz in the range [1 – 4][9], and on the group's values of the following changes in the values of independent nominal variables: $s1 \rightarrow s2$, $s2 \rightarrow s4$, $s1 \rightarrow s4$, $E1 \rightarrow E2$, $E2 \rightarrow E3$, and $E1 \rightarrow E3$.

7.2 Regressions for EPMcreate

For the EPMcreate data, the numerical variables are the same as for POEPMcreate. So the regressions to do involving numerical values are the same. For the nominal data, the story is different. Since there are only two sizes, $s1$ and $s2$, and only two experiments, $E1$, and $E2$, involved with EPMcreate data, there are only two changes for which to calculate regressions: $s2 \rightarrow s4$ and, without loss of generality, $E2 \rightarrow E1$. It turns out that all of the groups of size 4 are from Experiment 2 and vice versa, and all of the groups of size 2 are from Experiment 1 and vice versa. Thus, the $s2 \rightarrow s4$ change is identical to and is indistinguishable from the $E2 \rightarrow E1$ change. Therefore, for each dependent variable, we had to run only *one* regression of it on these two changes *together*. Moreover, it is impossible, at least from the data alone, to determine the contribution of each independent variable change to the cause of any dependent variable's change.

7.3 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 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 $\alpha = 0.05$. A table generated for a regression shows at least a coefficient and its P-value.

To do the regressions, we used MS Excel 2010 and the Analysis ToolPak.

[9] This linear regression is the discounted linear regression on group size as a numerical value. It's included so that the reader can see how poorly the linear regressions fit the data.

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

Independent Variable or Change Thereof	Original		Rescaled	
	Coefficient	P-value	Coefficient	P-value
crt	-0.291580199	0.567169434	-0.391747566	0.463615259
sz	4.004075962	0.372219723	4.016406215	0.379442735
$s1 \rightarrow s2$	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
$E1 \rightarrow E2$	-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

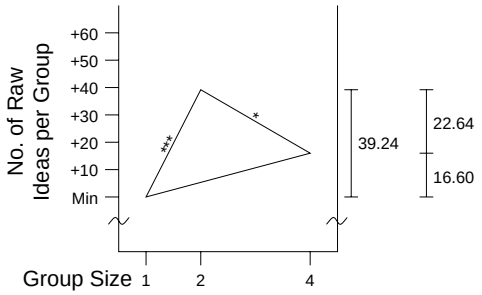


Fig. 7: Plot of PTR, Number of Raw Ideas per Group, against Group Size Changes for POEPMcreate

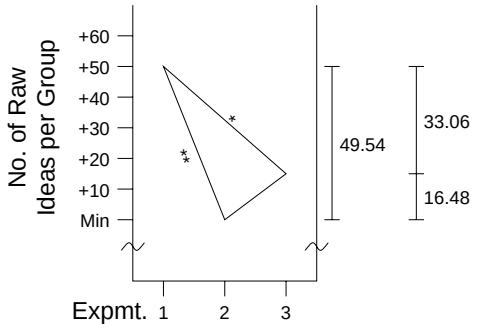


Fig. 8: Plot of PTR, Number of Raw Ideas per Group, against Experiment Number Changes for POEPMcreate

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

Independent Variable or Change Thereof	Original		Rescaled	
	Coefficient	P-value	Coefficient	P-value
crt	-0.181949813	0.676068493	-0.272127704	0.552595134
sz	3.684972045	0.289007366	3.696072629	0.299263136
$s1 \rightarrow s2$	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
$E1 \rightarrow E2$	-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

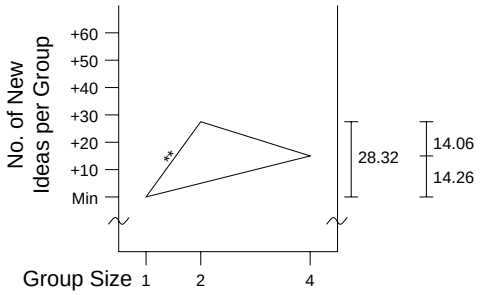


Fig. 9: Plot of PTN, Number of New Ideas per Group, against Group Size Changes for POEPMcreate

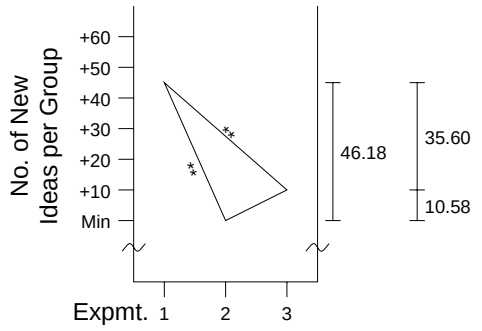


Fig. 10: Plot of PTN, Number of New Ideas per Group, against Experiment Number Changes for POEPMcreate

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

Independent Variable or Change Thereof	Original		Rescaled	
	Coefficient	P-value	Coefficient	P-value
crt	-0.159017017	0.561493124	-0.209117916	0.463874233
sz	-5.161418231	0.018261852	-5.155250985	0.020711643
$s1 \rightarrow s2$	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
$E1 \rightarrow E2$	-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

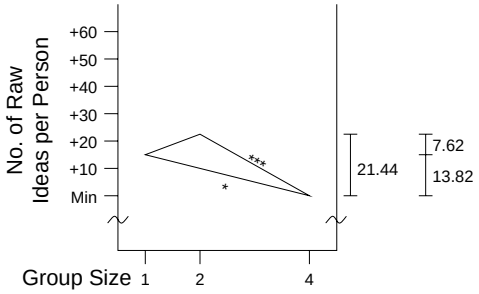


Fig. 11: Plot of PAR, Number of Raw Ideas per Group Member, against Group Size Changes for POEPMcreate

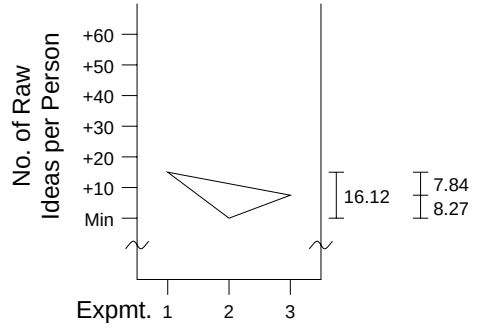


Fig. 12: Plot of PAR, Number of Raw Ideas per Group, Member against Experiment Number Changes for POEPMcreate

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

Independent Variable or Change Thereof	Original		Rescaled	
	Coefficient	P-value	Coefficient	P-value
crt	-0.138556740	0.556947638	-0.183625140	0.455529535
sz	-4.114031862	0.017238353	-4.108484099	0.020082549
$s1 \rightarrow s2$	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
$E1 \rightarrow E2$	-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

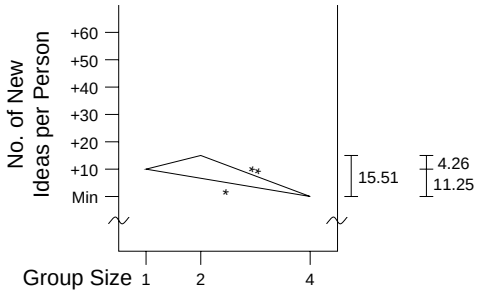


Fig. 13: Plot of PAN, Number of New Ideas per Group Member, against Group Size Changes for POEPMcreate

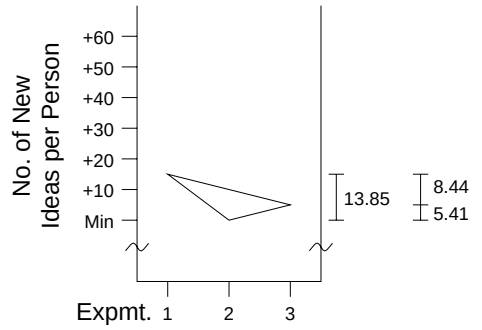


Fig. 14: Plot of PAN, Number of New Ideas per Group Member, against Experiment Number Changes for POEPMcreate

Table 10: Regression for ETR — EPMcreate, Total per group, Raw ideas

Independent Variable or Change Thereof	Original		Rescaled	
	Coefficient	P-value	Coefficient	P-value
crt	-0.041922257	0.867727865	-0.056278610	0.861726476
sz	13.05530425	0.006197000	-6.929813190	0.064549549
$s2 \rightarrow s4$ & $E2 \rightarrow E1$	26.11060850	0.006197000	-13.85962638	0.064549549

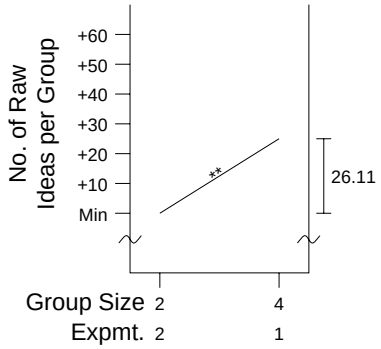


Fig. 15: Plot of ETR, Number of Raw Ideas per Group, against Group Size and Experiment Number Changes for EPMcreate

Table 11: Regression for ETN — EPMcreate, Total per group, New ideas

Independent Variable or Change Thereof	Original		Rescaled	
	Coefficient	P-value	Coefficient	P-value
crt	-0.335671729	0.112870142	-0.447509444	0.032538447
sz	16.25444575	0.000943312	-1.688984450	0.179140138
$s2 \rightarrow s4$ & $E2 \rightarrow E1$	32.50889150	0.000943312	-3.377968900	0.179140138

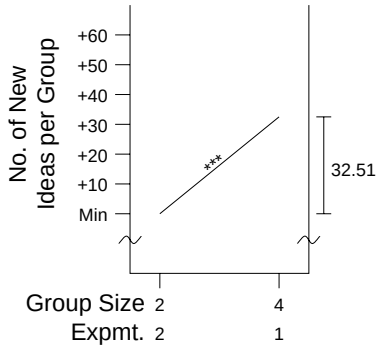


Fig. 16: Plot of ETN, Number of New Ideas per Group, against Group Size and Experiment Number Changes for EPMcreate

Table 12: Regression for EAR — EPMcreate, Average per group member, Raw ideas

Independent Variable or Change Thereof	Original		Rescaled	
	Coefficient	P-value	Coefficient	P-value
crt	-0.021126321	0.862437485	-0.028250535	0.860423976
sz	-1.160122509	0.293943982	-7.616966515	0.007994883
$s2 \rightarrow s4$ & $E2 \rightarrow E1$	-2.320245018	0.293943982	-15.23393303	0.007994883

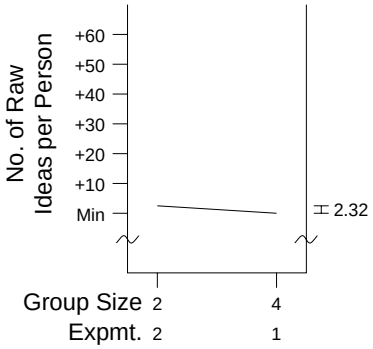


Fig. 17: Plot of EAR, Number of Raw Ideas per Group Member, against Group Size and Experiment Number Changes for EPMcreate

Table 13: Regression for EAN — EPMcreate, Average per group member, New ideas

Independent Variable or Change Thereof	Original		Rescaled	
	Coefficient	P-value	Coefficient	P-value
crt	-0.168166250	0.040727752	-0.223793634	0.020865985
sz	0.751673609	0.154645977	-4.827056916	0.001304228
$s2 \rightarrow s4$ & $E2 \rightarrow E1$	1.503347219	0.154645977	-9.654113833	0.001304228

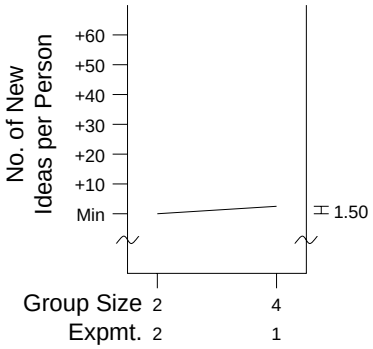


Fig. 18: Plot of EAN, Number of New Ideas per Group Member, against Group Size and Experiment Number Changes for EPMcreate

8 Regression Results

Tables 6 through 13 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 described in Section 8.2. This section concerns the coefficients and P-values of the original data that are in columns under the header "Original".

Interleaved among these tables are Figures 7 through 18 that show the plots derived from the nominal data parts of these tables, i.e., the last 6 rows for a POEPMcreate table and the last row for an EPMcreate table. In any graph, the y-axis gives coefficient values for expected changes in the values of the dependent variable described by the legend of the axis. The axis's 0 tick is labeled "Min" to remind us that we are looking at *changes* in the dependent variable's value and not its actual values. The x-axis's legend indicates the nominal independent variable whose changes are being considered. The axis's ticks indicate values of this nominal variable. The range lines to the right of a plot give the exact values of the y-axis changes. Each line has from zero to three asterisks indicating the order of magnitude of the 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. For example, in Figure 8, the line from (1,49.54) to (2,Min) means that the expected change in the dependent variable PTR (**P**OEPMcreate **T**otal group **R**aw Ideas) as a result of the change $E1 \rightarrow E2$ (from Experiment 1 to Experiment 2) is -49.54. Because of the three asterisks on this line, this change is significant at the $0.00n$ level for some n. The corresponding table row, the row for $E1 \rightarrow E2$ in Table 6 shows -49.54437592 as the coefficient and 0.005243091 as the P-value.

8.1 POEPMcreate Results

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 14 summarizes the results of the regressions relative to their support for the subhypotheses. The table's structure is described first columnwise and then rowwise.

The first three columns give:

1. the CET under examination, establishing the value of "c"s that appear in column headers as "P" for POEPMcreate and "E" for EPMcreate;
2. the compared group sizes, $s1 \rightarrow s2$, $s2 \rightarrow s4$, $s1 \rightarrow s4$, and $sz : [1 - 4]$; and
3. the compared experiments, $E1 \rightarrow E2$, $E2 \rightarrow E3$, and $E1 \rightarrow E3$.

The last four columns concern subhypotheses, as indicated by the headers of the four columns and superheaders of pairs and the quadruple of columns.

Rowwise, the table is divided into three sections separated by double rules:

1. the header section,
2. the POEPMcreate section, and

Table 14: Summary of Subhypothesis Conclusions

CET	Compared Group Sizes (s)	Compared Experiments (E)	Hypotheses: H1 — # Raw Requirement Ideas Generated by Whole Group: HcTR	Hypotheses: H1 — # New Requirement Ideas Generated by Whole Group: HcTN	Hypotheses: H2 — # Raw Requirement Ideas Generated by Group Member: HcAR	Hypotheses: H2 — # New Requirement Ideas Generated by Group Member: HcAN
POEPMcreate c = P		$E1 \rightarrow E2$	HPTR $E1 \rightarrow E2$ ** $\downarrow$ 49.54	HPTN $E1 \rightarrow E2$ ** $\downarrow$ 46.18	HPAR $E1 \rightarrow E2$ $\downarrow$ 16.12	HPAN $E1 \rightarrow E2$ $\downarrow$ 13.85
		$E2 \rightarrow E3$	HPTR $E2 \rightarrow E3$ $\uparrow$ 16.48	HPTN $E2 \rightarrow E3$ $\uparrow$ 10.58	HPAR $E2 \rightarrow E3$ $\uparrow$ 8.27	HPAN $E2 \rightarrow E3$ $\uparrow$ 5.41
		$E1 \rightarrow E3$	HPTR $E1 \rightarrow E3$ * $\downarrow$ 33.06	HPTN $E1 \rightarrow E3$ ** $\downarrow$ 35.60	HPAR $E1 \rightarrow E3$ $\downarrow$ 7.84	HPAN $E1 \rightarrow E3$ $\downarrow$ 8.44
	$s1 \rightarrow s2$		HPTR $s1 \rightarrow s2$ *** $\uparrow$ 39.24 (*** $\uparrow$ 39.24)	HPTN $s1 \rightarrow s2$ ** $\uparrow$ 28.32 (** $\uparrow$ 28.32)	HPAR $s1 \rightarrow s2$ $\uparrow$ 7.62 ($\uparrow$ 7.62)	HPAN $s1 \rightarrow s2$ $\uparrow$ 4.26 ($\uparrow$ 4.32)
	$s2 \rightarrow s4$		HPTR $s2 \rightarrow s4$ * $\downarrow$ 22.64 (* $\downarrow$ 22.60)	HPTN $s2 \rightarrow s4$ $\downarrow$ 14.06 ($\downarrow$ 14.02)	HPAR $s2 \rightarrow s4$ *** $\downarrow$ 21.44 (** $\downarrow$ 21.44)	HPAN $s2 \rightarrow s4$ ** $\downarrow$ 15.51 (** $\downarrow$ 15.49)
	$s1 \rightarrow s4$		HPTR $s1 \rightarrow s4$ $\uparrow$ 16.60 ($\uparrow$ 16.64)	HPTN $s1 \rightarrow s4$ $\uparrow$ 14.26 ($\uparrow$ 14.30)	HPAR $s1 \rightarrow s4$ * $\downarrow$ 13.82 (* $\downarrow$ 13.80)	HPAN $s1 \rightarrow s4$ * $\downarrow$ 11.25 (* $\downarrow$ 11.24)
	$sz : [1-4]$		HPTR $sz : [1-4]$ $\uparrow$ 4.00 ($\uparrow$ 4.02)	HPTN $sz : [1-4]$ $\uparrow$ 3.68 ($\uparrow$ 3.70)	HPAR $sz : [1-4]$ * $\downarrow$ 5.16 (* $\downarrow$ 5.16)	HPAN $sz : [1-4]$ * $\downarrow$ 4.11 (* $\downarrow$ 4.11)
EPMcreate c = E	$s2 \rightarrow s4$	$E2 \rightarrow E1$	HETR $s2 \rightarrow s4$ ** $\uparrow$ 26.11 ($\downarrow$ 13.85)	HETN $s2 \rightarrow s4$ *** $\uparrow$ 32.51 ($\downarrow$ 3.38)	HEAR $s2 \rightarrow s4$ $\downarrow$ 2.32 (** $\downarrow$ 15.23)	HEAN $s2 \rightarrow s4$ $\uparrow$ 1.50 (** $\downarrow$ 9.65)

3. the EPMcreate section.

In the header section, the headers for the first three columns are simple descriptions. 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 hypotheses. The first two of these four 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 groups of subhypotheses:

- about numbers of raw requirement ideas generated by whole groups, the HcTR hypotheses, for "c" = "P" or "E"; and
- about numbers of new requirement ideas generated by whole groups, the HcTN hypotheses, for "c" = "P" or "E".

H2 has two groups of subhypotheses:

- about average numbers of raw requirement ideas generated by group members, the HcAR hypotheses, for "c" = "P" or "E"; and
- about average numbers of new requirement ideas generated by group members, the HcAN hypotheses, for "c" = "P" or "E".

The POEPMcreate section is divided into seven subsections, one for each of the six regressions on changes in the nominal experiment and group size independent variables, $E1 \rightarrow E2$, $E2 \rightarrow E3$, $E1 \rightarrow E3$, $s1 \rightarrow s2$, $s2 \rightarrow s4$, and $s1 \rightarrow s4$ and one for the regression on group size treated as a numerical variable, $sz : [1 - 4]$.

A cell that is at the intersection of

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

is about a regression of the dependent variable cXY on $CIVoIV$, e.g., PTR on $E1 \rightarrow E2$, and has three rows:

1. The first row of the cell gives the subhypothesis, HcXY, of the cell paired with $CIVoIV$, e.g., "HPTR $E1 \rightarrow E2$".
2. The second row of the cell gives a triple (described just below) not enclosed in parentheses, reporting the result of the regression of the cell.
3. The third row of the cell gives a triple enclosed in parentheses, reporting the result of a regression on the scaled version of the values used for the regression of the cell, whose result is reported in the second row. (See Section 8.2.)

In the second and third rows, a triple is used to report the result of a regression, and it consists of three parts:

1. zero to three asterisks, reporting the strength of statistical significance of the result of the regression, as is done in the plots in Figures 7 through 18;
2. an arrow to report the direction of the coefficient of regression, with $\uparrow$ for positive and $\downarrow$ for negative; and
3. a numeral to indicate the magnitude of the coefficient of the regression.

The EPMcreate section deals with only one group size change and one experiment change, and it deals with them together in one regression. The notation used in this section is the same as that for a subsection of the POEPMcreate section.

Returning to consideration of the POEPMcreate results, examination of the non-parenthesized triples in the cells of the POEPMcreate section of Table 14 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.
- Overall, as indicated by the regression on $sz : [1 - 4]$, the larger of two groups generates about 4 and 3.6 more more raw and new requirement ideas, respectively, for each additional member it has over the smaller of the two groups, and neither difference is significant. Thus, a group of size 4 is expected to generate about 12 and 11 more raw and new requirement ideas, respectively, than does a group of size 1.

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. This phenomenon has been observed in brainstorming [42–46, 24, 47]. 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 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 14, to include the relevant group sizes in the statement of a subhypothesis.

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 22 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.
- Overall, as indicated by the regression on $sz : [1-4]$, per member, the larger of two groups generates about 5 and 4 *fewer* raw and new requirement ideas, respectively, for each additional member it has over the smaller of the two groups, and each difference is significant. Thus, per member, a group of size 4 is expected to generate about 15 and 12 more raw and new requirement ideas, respectively, than does a group of size 1.

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 requirements 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 grows quadratically with an increasing number of group members [48]. 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.

Table 14 shows that some, but not all experiment changes have significant effects on some but not all dependent variables. 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 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, it is necessary to try to factor them out of the results.

8.2 Scaling POEPMcreate Data and New POEPMcreate Results

We decided to try scaling the dependent variables by amounts that eliminate the effect of the experiment nominal variable. That is, we wanted the regression on changes in the experiment 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 POEPMcreate rows of 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" that give the coefficients and their P-values.

Figures 19 through 22 show the plots derived from the nominal data parts of these tables.

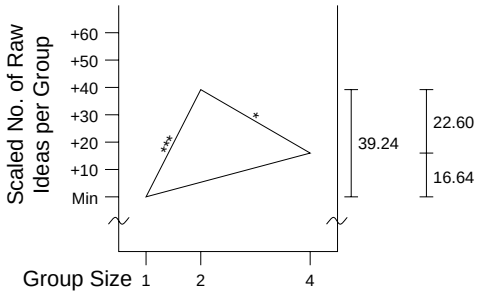


Fig. 19: Plot of Scaled PTR, Number of Raw Ideas per Group, against Group Size Changes for POEPMcreate

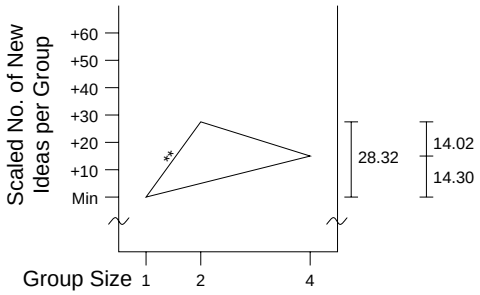


Fig. 20: Plot of Scaled PTN, Number of New Ideas per Group, against Group Size Changes for POEPMcreate

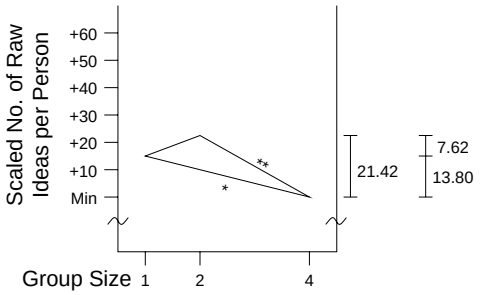


Fig. 21: Plot of Scaled PAR, Number of Raw Ideas per Group Member, against Group Size Changes for POEPMcreate

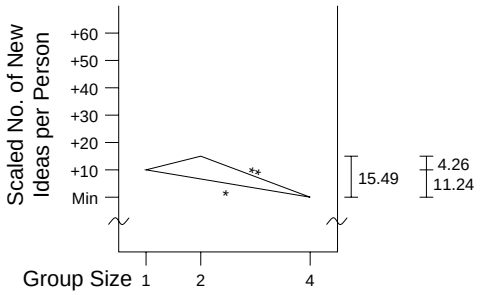


Fig. 22: Plot of Scaled PAN, Number of New Ideas per Group Member, against Group Size Changes for POEPMcreate

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 the parentesized triples in the POEPMcreate section of Table 14. 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.

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, changes in experiments are not significant, gives us confidence that the scaling factors used are correct.

This confidence in the scaling ended up being useful for dealing with the EPMcreate results.

8.3 EPMcreate Results

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 ETR, ETN, EAR, and EAN are -0.04, -0.36, -0.02, and -0.17, respectively, all being very close to 0, and they are not significant, with P-values of 0.87, 0.11, 0.86, and 0.04, respectively, all but the last being greater than 0.05.

Without Scaling Recall that for the EPMcreate dependent variables, ETR, ETN, EAR, and EAN, it is impossible to distinguish the effects of a change in experiment from Experiment 2 to Experiment 1 from the effects of a change in group size from two to four. The non-parenthesized triples in the cells of the EPMcreate section of Table 14 summarize the effects that these changes jointly have on the dependent variables.

For H1 about total requirement ideas generated by whole groups:

- A group of size 4 and in Experiment 1 generates about 26 and 32 more raw and new requirement ideas, respectively, than does a group of size 2, and each difference is very significant.

Thus, rejection of HETR is supported significantly for $s2 \rightarrow s4$ combined with $E2 \rightarrow E1$, and rejection of HETN is supported significantly for $s2 \rightarrow s4$ combined with $E2 \rightarrow E1$. For EPMcreate, group size and experiment make a difference in the numbers of raw and new requirement ideas generated by groups.

The directions of these rejections were as expected. For HETR and $s2 \rightarrow s4$ combined with $E2 \rightarrow E1$, and HETN and $s2 \rightarrow s4$ combined with $E2 \rightarrow E1$, the more group members, the more raw and new ideas are generated.

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

- Per member, a group of size 4 generates about 2.3 *fewer* raw requirement ideas than does a group of size 2, and this difference is not significant.
- Per member, a group of size 4 generates about 1.5 more new requirement ideas than does a group of size 2, and this difference is not significant.

Thus, rejection of HEAR is not supported for $s2 \rightarrow s4$ combined with $E2 \rightarrow E1$, and rejection of HEAN is not supported for $s2 \rightarrow s4$ combined with $E2 \rightarrow E1$. For EPMcreate, group size and experiment make no difference in the average numbers of raw and new requirement ideas generated by group members.

With Scaling To try to separate the effects of the group size change from the effects of the experiment change, we assumed that the effects of the experiment change on the EPMcreate variables were the same as those on the POEPMcreate variables. This assumption is reasonable because part of the purposes of Experiment 1 and Experiment 2 was to compare the effectiveness of EPMcreate and POEPMcreate. As a result, EPMcreate and POEPMcreate sessions were run in close time proximity. Moreover, in each experiment, the ideas generated by the EPMcreate groups and the POEPMcreate groups were lumped together into one file, sorted, and then subjected to evaluation by the same pair of evaluators. Therefore, whatever caused the change in experiment to give rise to a significant difference in the POEPMcreate dependent variables would likely cause the same change in experiment to give rise to the same significant difference in the EPMcreate dependent variables. It would then be likely that the same scaling factors would wash out the effects of the experiment changes.

Therefore, it made sense to scale values from Experiment 1 by 0.54 and to scale values from Experiment 2 by 1.33, in effect to pretend that each EPMcreate group participated in Experiment 3. The resulting scaled values are shown in the EPMcreate rows of last four columns of Table 5. Then, all of the regressions from Section 7.2 were

run again with the scaled values. The results of these regressions are actually in the two rightmost numerical columns of Tables 10 through 13, the columns under the header "Rescaled" that give the coefficients and their P-values.

Figures 23 through 26 show the plots derived from the nominal data parts of these tables.

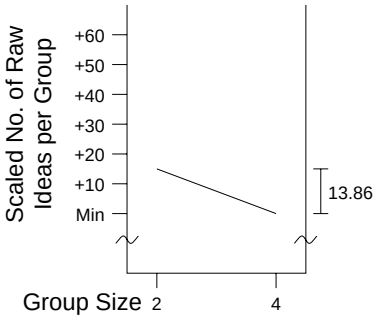


Fig. 23: Plot of Scaled ETR, Number of Raw Ideas per Group, against Group Size and Experiment Number Changes for EPMcreate

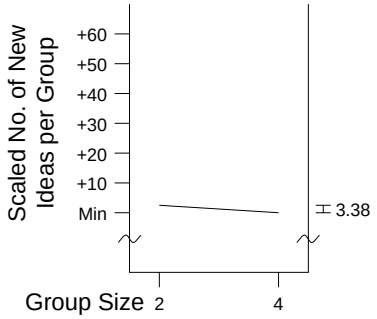


Fig. 24: Plot of Scaled ETN, Number of New Ideas per Group, against Group Size and Experiment Number Changes for EPMcreate

After this scaling, a group's average Williams test score, crt, still has no effect on any dependent variable. The coefficients for the numerical variable crt for ETR, ETN, EAR, and EAN are -0.06, -0.45, -0.03, and -0.22, respectively, all being very close to 0, and they are not significant, with P-values of 0.86, 0.03, 0.86, and 0.02, respectively, all but the second and the last being greater than 0.05.

When we run the regressions to determine the effects on the scaled variables of the joint changes in experiment from Experiment 2 to Experiment 1 and in group size from

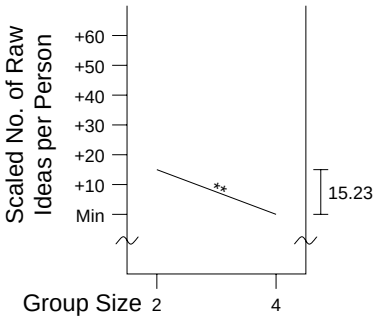


Fig. 25: Plot of Scaled EAR, Number of Raw Ideas per Group Member, against Group Size and Experiment Number Changes for EPMcreate

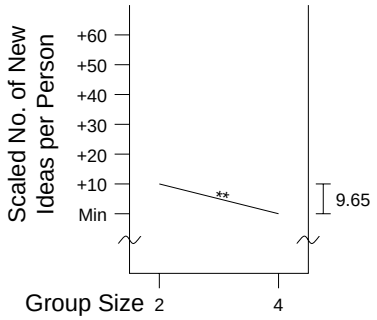


Fig. 26: Plot of Scaled EAN, Number of New Ideas per Group Member, against Group Size and Experiment Number Changes for EPMcreate

s2 to s4, we expect that the fiction that all groups participated in Experiment 3 to wash out the effect of the experiment change.

An examination of the non-parenthesized and parenthesized triples in the cells of the EPMcreate section of Table 14 shows that the results have changed dramatically as a result of scaling and the attempt to factor out the contribution of the experiment variable to the effects. For H1 about total requirement ideas generated by whole groups:

- A group of size 4 generates about 14 and 3 *fewer* raw and new requirement ideas, respectively, than does a group of size 2, and neither difference is significant.

Thus, rejection of HETR is not supported for $s2 \rightarrow s4$, and rejection of HETN is not supported for $s2 \rightarrow s4$. For EPMcreate, group size and experiment make no difference in the numbers of raw and new requirement ideas generated by groups.

The directions of these effects were contrary to expectation: For HETR and $s2 \rightarrow s4$ and HETN and $s2 \rightarrow s4$, the more group members, the *fewer* raw and new ideas are generated. Here, again, it appears that the larger group is suffering from larger group-management overhead.

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

- Per member, a group of size 4 generates about 15 and 10 *fewer* raw and new requirement ideas, respectively, than does a group of size 2, and each difference is very significant.

Thus, rejection of HEAR is supported significantly for $s2 \rightarrow s4$, and rejection of HEAN is supported significantly for $s2 \rightarrow s4$. For EPMcreate, group size and experiment make a significant difference in the average numbers of raw and new requirement ideas generated by group members.

The directions of these rejections were contrary to expectations: For HEAR and $s2 \rightarrow s4$ and HEAN and $s2 \rightarrow s4$, the more group members, the *fewer* raw and new ideas are generated per group member. Also here, it appears that the larger group is suffering from larger group-management overhead.

Which Results to Accept Whereas the scaling did not change the conclusions about support for the POEPMcreate subhypotheses, HPTR, HPTN, HPAR, and HPAN, the scaling completely changes the conclusions about support for the EPMcreate subhypotheses, HETR, HETN, HEAR, and HEAN. Each rejection that was significant without scaling is not significant with scaling and vice versa. Furthermore, the scaling flipped the direction of the rejection in three of the four subhypotheses. The question of which set of conclusions to believe naturally arises.

While the conclusions do appear to change in fundamental ways, there is a way that they can be viewed as saying approximately the same thing. In either case, the average number of raw requirement ideas generated per group member is smaller in the larger group. The difference in the conclusions is at which group size does the lower effectiveness of a group member in the larger group cause the whole group to generate fewer ideas. Without scaling, the reduction in the larger group's effectiveness has not happened yet, and with scaling, the reduction in the larger group's effectiveness happens in the change from group size 2 to group size 4. If one accepts the thinking in the last two

paragraphs at the end of Section 8.1, the overall better effectiveness of the participants of Experiment 1 allowed the groups of size 4, which were all from Experiment 1, to be more effective than they would be otherwise, thus counteracting the dominance of group-management overhead.

The conclusion of Section 8.2 is that, at least for the POEPMcreate results, both the scaling and the factors used in the scaling made sense. The calculation of each scale factor used data from both CETs to get a scale factor that can be legitimately applied to the data about both CETs. Thus, there is good reason to believe that scaling *should be* applied to the EPMcreate data, and that the results from the scaled data are more likely correct than are those of the unscaled data. Therefore, we conclude it is likely that for EPMcreate, the reduction in the larger group's effectiveness happens in the change from group size 2 to group size 4, and that the scaled results should be accepted as definitive.

9 Threats to Validity and Future Work to Address Them

Many of the possible threats to the validity of the conclusions to Experiment 3 and the analysis of the three experiments threatened also Experiments 1 and 2 [26]. The discussion of these enduring threats is repeated from the paper about Experiments 1 and 2.

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. Counting of raw ideas is correct, because as mentioned, at least one famous CET, i.e., brainstorming, has as a principal goal for its first stage the generation of as many ideas as possible, under the principles that an idea's quality is evaluated only after the first stage is finished and that quality follows quantity [21].

The method to evaluate the quality of an idea, determining its newness and its realizability, is based on (1) an accepted definition of creativity, that it generates new and useful ideas and (2) taking newness relative to the existing implementation as the measure of newness and realizability in the current context as the measure of usefulness. Admittedly, this measure of quality is subject to debate. For example, Briggs *et al.* [49] 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".

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 [38], and the test seems

to be accepted in the academic psychometric testing field [50]. 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 [51]. We modified the test to be for adults attending a university or working [28, 32]. 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.

Section 6 discusses three other threats to construct validity, namely (1) whether combining data from multiple experiments is valid, (2) whether balancing the groups' average creativity scores eliminated differences in group members' individual creativity as a cause for observed differences in the groups' requirement idea generation, and (3) whether group size should be a numerical variable. Section 6 discusses in detail how these threats were dealt with.

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 its assigned CET 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 factors that can account for the differences in the number of requirement ideas among groups are the CET being used by the groups and the sizes of the groups.

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:

- 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 from which the subjects are are in the 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 [52]. Indeed, in one of the early experiments [28] 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.

- the particular choice of the types of stakeholders whose viewpoints were used by EPMcreate and POEPMcreate sessions: Would other choices, e.g., of teachers, work as well?
- the single Web site as the CBS for which to generate requirement ideas: Would a different Web site or even a different kind of CBS inhibit the effectiveness of any CET? Our experience in using the full EPMcreate and a variant of it to generate requirement ideas for four different CBSs [28, 15] leads us to believe that the kind of CBS has no effect on the effectiveness of any variant of EPMcreate.
- 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 [21, 31]. The experiments described in this paper entirely avoided this issue. All of the subjects were CS students with very similar sets of competencies, backgrounds, and domain knowledge. Thus it is unlikely that any group gained any advantage over another on the basis of this issue. 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 [32].
- Finally, 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 [53], 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.

9.4 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

It is useful to step back from the definitive results of an experiment and to take a closer look at the data to begin to understand why the results are what they are and to possibly speculate about additional results that may require additional work in the future to confirm.

The definitive results described in Section 8 can be summarized in a structured way that showcases some surprises.

1. When EPMcreate is used to help generate ideas for requirements elicitation,
   A. according to the original data,
      – per whole group,
        a four-person group generates on average significantly more raw and new ideas than a two-person group;
      – however, *per group member*,
        there is no significant difference in the numbers of raw and new ideas generated on average by a two-person group member and a four-person group member:
        - a four-person group member generates on average slightly fewer raw ideas than a two-person group member, while
        - a four-person group member generates on average slightly more new ideas than a two-person group member,
   B. but, according to the rescaled data,
      – per whole group,
        there is no significant difference in the numbers of raw and new ideas generated on average by a two-person group and a four-person group:
        - a four-person group generates on average fewer raw ideas than a two-person group, while
        - a four-person group generates on average slightly fewer new ideas than a two-person group;
      – however, *per group member*,
        a four-person group member generates on average significantly fewer raw and new ideas than a two-person group member.

2. When POEPMcreate is used to help generate ideas for requirements, elicitation, *according to both the original and the rescaled data*,
   – per whole group,
     - a two-person group generates on average significantly more raw and new ideas than a one-person group,
     - a four-person group generates on average fewer raw and new ideas than a two-person group, and significantly so for raw ideas,
     - there is no significant difference in the numbers of raw and new ideas generated on average by a one-person group and a four-person group:
       * a four-person group generates on average more raw and new ideas than a one-person group,
   – however, *per group member*,
     - there is no significant difference in the numbers of raw and new ideas generated on average by a one-person group member and a two-person group member:
       * a two-person group member generates on average more raw ideas than a one-person group member, and
       * a two-person group member generates on average slightly more new ideas than a one-person group member,

    - 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.

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 [21]. These data suggest that synergy, if indeed it is present, is not uniformly helpful. In particular,

- when EPMcreate is used to help generate ideas for requirements elicitation,
    - according to the original data, a four-person group member generates on average slightly fewer raw ideas than a two-person group member, and
    - according to the rescaled data,
        - a four-person group generates on average fewer raw ideas and slightly fewer new ideas than a two-person group, and
        - a four-person group member generates on average significantly fewer raw and new ideas than a two-person group member,

  and
- when POEPMcreate is used to help generate ideas for requirements, elicitation, 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 lines of intermember communication in a group is increasing quadratically with increasing group size.

The natural question to ask is "What is more important for requirements engineering uses of EPMcreate or POEPMcreate to optimize,

1. the total number of ideas in a group, or
2. the number of ideas per member in a group?"

For instance, the original data in Table 5 say that for EPMcreate,

1. a four-person group generates on average 61.5 raw requirement ideas, 15.38 per member, but
2. a two-person group generates on average 35.25 raw requirement ideas, 17.63 per member,

and that for POEPMcreate,

1. a four-person group generates on average 52 raw requirement ideas, 13 per member, but
2. a two-person group generates on average 55.38 raw requirement ideas, 27.69 per member.

3. a one-person group generates on average 24 raw requirement ideas, 24 per member.

These data say that

1. the more or most effective total group is the four-person group, and
2. the more or most effective group member is that of the two-person group.

What is the better group size, four, because the total number of ideas is bigger, or two, because each member will be more productive? If we believe the data presented thus far, the answer would be "two".

- You want to get as many people as you can afford working to make ideas. The more people you have, the more ideas you get.
- However, to maximize the power of the individual and thus ultimately of the group, you should split the people you have into groups of two.

For example, if you are using EPMcreate and can afford four people, then make two groups of two. In one four-person group, you will get on average 61.5 ideas. In two two-person groups, you will get on average 70.5 ideas. The same is even more true with POEPMcreate. In one four-person group, you will get on average 52 ideas. In two two-person groups, you will get on average 110.76 ideas. For the rescaled data, the effect is even more pronounced. With EPMcreate, in one four-person group, you will get on average 33.21 ideas. In two two-person groups, you will get on average 93.77 ideas. With POEPMcreate, in one four-person group, you will get on average 40.5 ideas. In two two-person groups, you will get on average 126.43 ideas. Of course, the questions are "How much overlap is there between two groups? Is the overlap high enough that the benefit of having smaller but more groups is lost?"

We searched for shared ideas among the ideas generated by all possible pairs of the four two-person EPMcreate groups and all possible pairs of the four two-person POEPMcreate groups in Experiment 2. We did this search in only the original data because they are more conservative with respect to what we are trying to determine. Table 15 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, with "E" meaning "EPMcreate" and "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 upper left-hand triangle (in rows A, B, and C by columns B, C, and D) shows the idea sharing among the pairs of EPMcreate groups, and the lower right-hand 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 A, which generated 30 raw requirement ideas using EPMcreate, and the column for Group B, which generated 35 raw requirement ideas using EPMcreate, is that there were 6 ideas in common among the generated ideas of the two groups; 17.14% of Group B's ideas were in common with Group A's ideas; and 20% of Group A's ideas were in common with Group B's ideas.

This search shows that among pairs of EPMcreate two-person groups, the average and maximum percentage overlap of raw requirement ideas generated were 15.55 and

Table 15: Shared Ideas Among Pairs of Two-Person Groups in Experiment 2

			B		C		D		F		G		H	
			E2	35	E2	40	E2	36	P2	40	P2	42	P2	63
A	E2	30	6 20 %	17.14%	3 10 %	7.5 %	3 10 %	8.33%						
B	E2	35			8 22.86%	20 %	6 17.14%	16.67%						
C	E2	40					7 17.5 %	19.44%						
E	P2	45							13 28.89%	32.5 %	7 15.56%	16.67%	14 31.11%	22.22%
F	P2	40									11 27.5 %	26.19%	14 35 %	22.22%
G	P2	42											16 38.1 %	25.4 %

22.86, respectively, and among pairs of POEPMcreate two-person groups, the average and maximum percentage overlap of raw requirement ideas generated were 26.78 and 38.1, respectively. It is not unexpected that the overlap is higher in the CET that was demonstrated to be more effective.

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 16 shows the results of these searches in the same format used in Table 15.

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, 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 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[10].

Let us continue the example of what to do if you have four people to use POEPMcreate to generate requirement ideas. With one four-person group, you will get on average 52 ideas. With two two-person groups, you will get on average about 111 ideas. At worst, 38 of them will be in common, leaving 73 usable ideas, about 40% more than the four-person group. However, on average, only 22 of them will be in common, leaving about 89 usable ideas, 71.15% more than the 52 ideas from the four-person group. With four one-person groups, i.e., four individuals, you will get on average 96 ideas. At worst 38 will be common, leaving 58 usable ideas, fewer than with two two-person groups, but still more than one four-person group. However, on average only 22 of the ideas will

[10] A little thought shows that this search for shared ideas is very expensive. It requires a comparison of every pair of the dozens of ideas of every pair of the tested groups. The comparison of two ideas is not syntactic, because a judgment must be made whether the compaired *mean* the same thing. With a group generating on average about 45 raw ideas, there are about 1000 comparisons for each pair of groups. Then if there are n groups to pair, there are $n^2/2 + n$ of these 1000 comparisons to do. Because of this quadratic growth of these 1000 comparisons, it is considerably cheaper to do a search among the pairings of four groups and a search among the pairings of another four groups than to do a search among the pairings of eight groups. Moreover, there is no reason to expect that the amount of sharing among two distinct sets of four groups will be significantly different from the amount of sharing among one combined set of eight groups made from the two sets of four groups.

Table 16: 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
					11.94%			12.12%			30 %			33.3 %			25.93%			7.4 %			27.78%			11.11%
A	P2	90		8			8			9			10			7			2			5			2	
			8.89%			8.89%			10 %			11.11%			7.78%			2.22%			5.56%			2.22%	
B	P2	67					21.21%			26.67%			26.67%			29.63%			14.81%			33.33%			38.89%
							14			8			8			8			4			6			7	
						20.9 %			11.94%			11.94%			11.94%			5.97%			8.96%			10.45%	
C	P2	66								30 %			26.67%			22.22%			14.81%			44.44%			33.33%
										9			8			6			4			8			6	
									13.64%			12.12%			9.1 %			6.06%			12.12%			9.1 %	
D	P2	30										20 %		26.67%			25.93%			7.4 %			38.89%			27.78%
												6		8			7			2			7			5
												20 %		23.33%			6.67%			6.67%			23.33%			16.67%
E	P1	30														22.22%			7.4 %			16.67%			22.22%
																6			2			3			4
															20 %			6.67%			10 %			10.33%	
F	P1	27																	0.0 %			16.67%			27.78%
																			0			3			5
																		0.0 %			11.11%			8.23%	
G	P1	27																				16.67%			5.56%
																						3			1
																						11.11%			3.7 %
H	P1	18																							11.11%
																									2
																									11.11%

be in common, leaving 74 usable ideas, again fewer than with two two-person groups, but still more than one four-person group. Thus, it appears that it is better to split a large group into smaller groups when generating requirement ideas with EPMcreate or POEPMcreate, and at least for POEPMcreate, one should not split two-person groups into individuals.

On the assumption that this conclusion is generalizable, for EPMcreate or POEPMcreate, if one has more than two people available, she should split them into as many two-person groups as possible, and then a one-person group if the number of people is odd. These groups should work independently, and then their ideas should be combined while eliminating repeated ideas.

Several researchers [42–46, 24, 47] had noticed a similar phenomenon for brainstorming, that smaller groups are more effective per person than larger groups, and that individuals are the most effective. Future work is needed with experiments designed specifically to test the speculative conclusions of this section.

A mystery is that data of this paper show that for POEPMcreate, an individual is generating more raw and new ideas when working a two-person group than when working alone, although the difference is not statistically significant. The empirical brainstorming literature says that the average individual is better than any group per person [47]. Perhaps there is something about POEPMcreate's procedure that mitigates the drag that a group places on individuals in brainstorming. Future work is needed also to solve this mystery.

11 Qualitative Triangulation

To interpret the combined results of the three experiments, we designed and deployed in late August 2012 an online questionnaire that can be found at:
`https://docs.google.com/spreadsheet/viewform?formkey=dFI2UWx0MWJuRUdvQ1JNZnh1NFN0SGc6MQ`
The questionnaire's main goal was to learn what industrial practitioners knew about individual versus group requirements elicitation. Another goal was to learn the extent of the use of CETs in industrial requirements elicitation.

The most important of the questions were:

- 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
- Size of ideal groups

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[11], INCOSE[12], Requirements Engineering Network's forum[13], RE-online[14], and Yahoo's Requirements-Engineering Group[15] 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. See
`https://cs.uwaterloo.ca/~dberry/FTP_SITE/tech.reports/53Xresponses.pdf` [16]
for an automatically generated summary of the responses. 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.

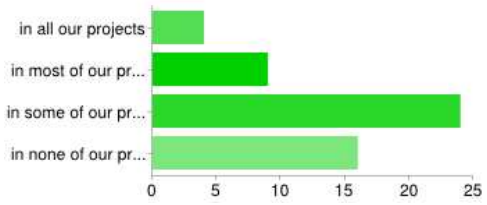
The answers to the demographics question about the roles the respondent plays in his or her organization shows that many respondents are involved in more than one role, each of 45% of the respondents is a business or requirements analysts (BoRAs) in all or most of his or her organization's projects,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

Figure 27 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 sep-

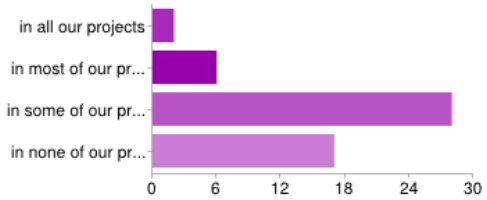
[11] `http://www.iiba.org/`
[12] `http://www.incose.org/`
[13] `http://www.requirementsnetwork.com/`
[14] `http://discuss.it.uts.edu.au/mailman/listinfo/re-online`
[15] `http://tech.groups.yahoo.com/group/Requirements-Engineering/`
[16] If you copy and paste this URL into a browser, before hitting "Enter", please change the character before "dberry" that only looks like a tilde to a true ASCII tilde.

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



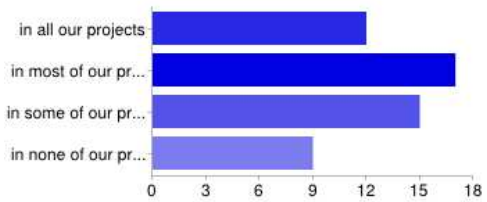
in all our projects	**4**	8%
in most of our projects	**9**	17%
in some of our projects	**24**	45%
in none of our projects	**16**	30%

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



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. 27: Requirements Elicitation Done by Individuals and Groups

arately 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, 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 28 shows that the usual number of BoRAs in a requirements elicitation group

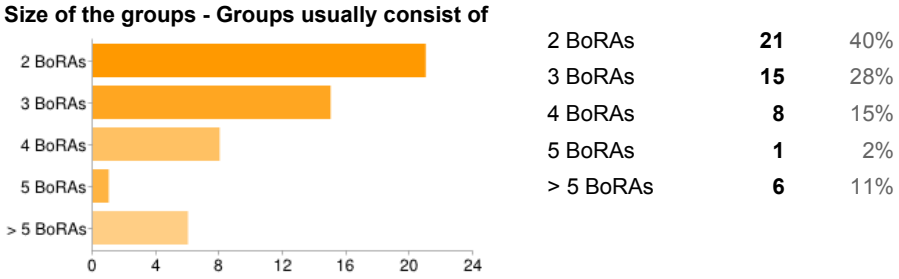


Fig. 28: Sizes of Groups Doing Requirements Elicitation

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.

Figure 29 shows that when respondents were asked specifically how they would

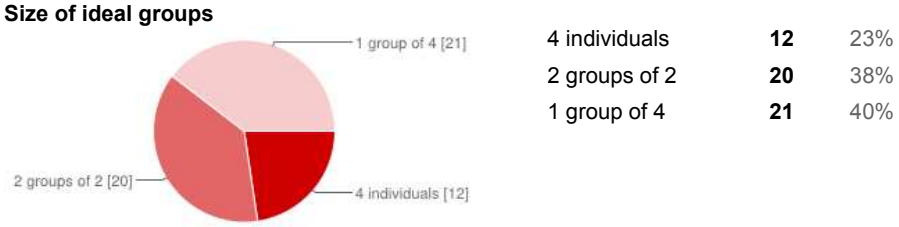


Fig. 29: Ideal Group Size for Doing Requirements Elicitation

distribute an available 4 BoRAs to do a requirements elicitation task, 23% said that they would have the BoRAs work individually, 38% said that they would have the BoRAs work in 2 groups of 2, and 40% said that they would have the BoRAs work in 1 group of 4. That is, 61% of the respondents said that they would have the 4 BoRAs work in groups of sizes 1 and 2. Therefore, smaller groups seem to be chosen even when more

BoRAs are available for a task, and if more BoRAs are available, the extra ones would be used to make other groups rather than to beef up any one group.

It seems that BoRAs, PMs, and SWEs 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 conclusions about EPMcreate and POEPMcreate that the experimental results weakly support may be correct and that more work needs to be done to strengthen the results.

12 Related Work

The introductions to our earlier papers introducing EPMcreate and POEPMcreate and describing empirical validations of their effectiveness [28, 26] discuss related work about

- definitions of creativity,
- the role of creativity in SE in general,
- the role of creativity in RE in specific,
- particularly in requirements elicitation to invent requirements, to discover missing requirements, and to deal with wicked problems,
- the role of communication and interaction in creativity for RE, and
- CETs, including not only the granddaddy of them all, brainstorming, against which most others are compared, but also more focused techniques.

Some of this work describes empirical validations of the effectiveness of the techniques they desccribe. The introduction to the current paper summarizes this related work.

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 Section 8.1.

The most recent related work is mostly empirical, including experiments, case studies, and systematic reviews. For example, Kauppinen, Savolainen, and Männisto [54] 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 [37] study 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 exiting requirements for the domain.

Lemos, Alves, Duboc, and Rodrigues [55] conduct 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

- more *empirical* evidence about the effectiveness 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 develpment lifecycle.

Svensson, Taghavianfar, and Gren [56] conduct 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, and
- 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.

13 Conclusions

The data from three experiments with identical design and conduct are combined to draw conclusions that

1. among pairs of different sized EPMcreate or POEPMcreate elicitation groups, the larger of the two groups is more effective overall;
2. among POEPMcreate elicitation groups, per group member, a two-person group is more effective on average than a four-person group, while among EPMcreate elicitation groups, per group member, there is no significant difference between a two-person and a four-person group; and
3. among POEPMcreate elicitation groups, per group member, a two-person group is slightly but not significantly more effective than a one-person group, and in the last analysis, there is no significant difference between a one-person and a two-person group.

The slight but insignificant difference in Conclusion 3 together with the surprising Conclusion 2 leads us to speculate about optimal group size and the possibility that dividing the available EPMcreate or POEPMcreate practitioners into groups of two, but no

further, may be the best strategy. The corroborating survey data indicate that industry seems to have come to the same conclusions about CETs 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 and is considered ideal at least as often as any other size. More work is needed to resolve this speculation.

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 is 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 paper [57], by the same authors, published in *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, and a strengthening of the conclusions.

Note to Editor: This paper uses in Sections 1, 2.1 through 2.3, and 9 material copied verbatim from the authors' previous paper [26], describing EPMcreate, POEPMcreate, the conduct of the experiment, and threats.

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.

There is no known potential or actual conflict of interest.

References

1. Gause, D., Weinberg, G.: Exploring Requirements: Quality Before Design. Dorset House, New York, NY, USA (1989)
2. Goguen, J.A.: Requirements engineering as the reconciliation of technical and social issues. Technical report, Centre for Requirements and Foundations, Programming Research Group, Oxford University Computing Lab (1993) modified version later published as [3].
3. Goguen, J.A.: Requirements engineering as the reconciliation of technical and social issues. In: Requirements Engineering: Social and Technical Issues, Academic Press (1994) 165–199

4. Nguyen, L., Carroll, J., Swatman, P.A.: Supporting and monitoring the creativity of IS personnel during the requirements engineering process. In: Proc. 33rd Hawaii Int. Conf. on System Sciences. HICSS-33, Maui, HI, USA (2000) http://csdl2.computer.org/comp/proceedings/hicss/2000/0493/07/04937008.pdf.
5. Maiden, N., Gizikis, A.: Where do requirements come from? IEEE Software **18** (2001) 10–12
6. Robertson, J.: Eureka! Why analysts should invent requirements. IEEE Software **19** (2002) 20–22
7. Maiden, N., Robertson, S., Gizikis, A.: Provoking creativity: Imagine what your requirements could be like. IEEE Software **21** (2004) 68–75
8. Hoffmann, O., Cropley, D., Cropley, A., Nguyen, L., Swatman, P.: Creativity, requirements and perspectives. Australasian J. Information Systems **13** (2005) 159–174
9. Maiden, N., Robertson, S., Robertson, J.: Creative requirements: Invention and its role in requirements engineering. In: Proceedings of the 28th International Conference on Software Engineering (ICSE). (2006) 1073–1074
10. Schlosser, C., Jones, S., Maiden, N.: Using a creativity workshop to generate requirements for an event database application. In: Proc. Int. Workshop Requirements Engineering: Foundation for Software Quality, REFSQ'08. LNCS 5025, Berlin, Germany, Springer (2008) 109–122
11. Nguyen, L., Shanks, G.: A framework for understanding creativity in requirements engineering. J. Information & Software Technology **51** (2009) 655–662
12. Rittel, H., Webber, M.: Dilemmas in a general theory of planning. Policy Sciences **4** (1973) 155–169
13. Geschka, H.: Creativity techniques in product planning and development: A view from West Germany. R&D Management **13** (1983) 169–183
14. Rickards, T.: Creativity and the Management of Change. Blackwell, Oxford, UK (1999)
15. Mich, L., Berry, D.M., Franch, M.: Classifying web-application requirement ideas generated using creativity fostering techniques according to a quality model for web applications. In: Proc. 12th Int. Workshop Requirements Engineering: Foundation for Software Quality, REFSQ'06. (2006)
16. de Bono, E., Heller, R.: Can creative management techniques help you survive the recession (Viewed 10 August 2010) http://www.thinkingmanagers.com/management/creative-management-techniques.
17. Runco, M.A.: Creativity: Theories and Themes: Research, Development, and Practice. Elsevier Academic Press, Burlington, MA, USA (2007)
18. Simonton, D.K.: Scientific Genius: A Psychology of Science. Cambridge University Press, Cambridge, UK (1988)
19. Amabile, T.M.: A model of creativity and innovation in organizations. Research in Organizational Behaviour **10** (1988) 123–167
20. Feist, G.J.: A structural model of scientific eminence. Psychological Science **4** (1993) 366–371
21. Osborn, A.: Applied Imagination. Charles Scribner's, New York, NY, USA (1953)
22. de Bono, E.: Six Thinking Hats. Viking, London, UK (1985)
23. de Bono, E.: Serious Creativity: Using the Power of Lateral Thinking to Create New Ideas. Harper Collins, London, UK (1993)
24. Aurum, A., Martin, E.: Requirements elicitation using solo brainstorming. In: Proc. 3rd Australian Conf. on Requirements Engineering, Deakin University, Australia (1998) 29–37
25. Jones, S., Lynch, P., Maiden, N., Lindstaedt, S.: 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). (2008) 289–294

26. Sakhnini, V., Mich, L., Berry, D.M.: The effectiveness of an optimized EPMcreate as a creativity enhancement technique for Website requirements elicitation. Requirements Engineering Journal **17** (2012) 171–186
27. etourism Website: 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.
28. Mich, L., Anesi, C., Berry, D.M.: Applying a pragmatics-based creativity-fostering technique to requirements elicitation. Requirements Engineering Journal **10** (2005) 262–274
29. Wikipedia: Free Boolean algebra (Viewed 10 August 2010) http://en.wikipedia.org/wiki/Free_Boolean_algebra.
30. Preparata, F.P., Yeh, R.T.Y.: Introduction to Discrete Structures for Computer Science and Engineering. Addison-Wesley Longman, Boston, MA, USA (1973)
31. von Bertalanaffy, L.: General Systems Theory: Foundations, Development, Applications. revised edn. George Braziller, New York, NY, USA (1976)
32. Mich, L., Berry, D.M., Alzetta, A.: 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. (2010)
33. 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.
34. Salzer, H., Levin, I.: Atomic requirements in teaching logic control implementation. International Journal of Engineering Education **20** (2004) 46–51
35. Dean, D.L., Hender, J.M., Rodgers, T.L., Santanen, E.L.: Identifying quality, novel, and creative ideas: Constructs and scales for idea evaluation. Journal of the Association for Information Systems **7** (2006)
36. Conboy, K., Wang, X., Fitzgerald, B.: Creativity in agile systems development: A literature review. In: Information Systems — Creativity and Innovation in Small and Medium-Sized Enterprises, Proceedings IFIP WG8.2 International Conference, CreativeSME 2009. Volume IFIP AICT 301. (2009) 122–134
37. Zachos, K., Maiden, N.: Inventing requirements from software: An empirical investigation with web services. In: Proceedings of the 16th IEEE International Requirements Engineering Conference (RE). (2008) 145–154
38. Williams, F., Taylor, C.W.: Instructional media and creativity. In: Proc. 6th Utah Creativity Research Conf., New York, NY, USA, Wiley (1966)
39. Kaufman, J.C., Sternberg, R.J., eds.: The International Handbook of Creativity. Cambridge University Press, Cambridge, UK (2006)
40. Sakhnini, V., Berry, D.M., Mich, L.: Materials for Comparing POEPMcreate, EPMcreate, and Brainstorming. Technical report, School of Computer Science, University of Waterloo (Viewed 7 March 2010) http://se.uwaterloo.ca/~dberry/FTP_SITE/software.distribution/EPMcreateExperimentMaterials/.
41. Berry, W., Sanders, M.: Understanding Multivariate Research: A Primer For Beginning Social Scientists. Westview Press, New York, NY, USA (2000)
42. Dornburg, C.C., Stevens, S.M., Hendrickson, S.M.L., Davidson, G.S.: 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 (2008) http://prod.sandia.gov/techlib/access-control.cgi/2008/085971.pdf.
43. Dennis, A.R., Valacich, J.S.: Computer brainstorms: More heads are better than one. Journal of Applied Psychology **78** (1993) 531–537
44. Furnham, A., Yazdanpanahi, T.: Personality differences and group versus individual brainstorming. Personality and Individual Differences **19** (1958) 73–80

45. Taylor, D.W., Berry, P.C., Block, C.H.: Does group participation when using brainstorming facilitate or inhibit creative thinking? Administrative Science Quarterly **3** (1958) 23–47
46. Isaksen, S.G., Gaulin, J.P.: A reexamination of brainstorming research: Implications for research and practice. Gifted Child Quarterly **40** (Fall 2005) 315–329
47. Kohn, N.W., Smith, S.M.: Collaborative fixation: Effects of others' ideas on brainstorming. Applied Cognitive Psychology **25** (2011) 359–371
48. Brooks, Jr., F.P.: The Mythical Man-Month: Essays on Software Engineering. Second edn. Addison-Wesley, Reading, MA (1995)
49. Briggs, R.O., Reinig, B.A., Shepherd, M.M., Yen, J., Nunameker, Jr., J.F.: Quality as a function of quantity in electronic brainstorming. In: Hawaii International Conference on System Sciences. (1997) 94–103
50. Dow, G.: 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 7 March 2010)
51. West Side School District: Gifted and Talented Program. Technical report, West Side Public Schools, Higden, AR, U.S.A. (Viewed 7 March 2010)
52. Berander, P.: Using students as subjects in requirements prioritization. In: Proceedings of the International Symposium on Empirical Software Engineering (ISESE'04), IEEE Computer Society (2004) 167–176
53. Wikipedia: Type I and type II errors (Viewed 20 February 2012) `http://en.wikipedia.org/wiki/Type_I_and_type_II_errors`.
54. Kauppinen, M., Savolainen, J., Männisto, T.: Requirements engineering as a driver for innovations. In: Proceedings of the 15th IEEE International Requirements Engineering Conference (RE). (2007) 15–20
55. Lemos, J.a., Alves, C., Duboc, L., Rodrigues, G.N.: A systematic mapping study on creativity in requirements engineering. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC). (2012) 1083–1088
56. Svensson, R.B., Taghavianfar, M., Gren, L.: Creativity techniques for more creative requirements: Theory vs. practice. In: Proceedings of the Forty-First Euromicro Conference on Software Engineering and Advanced Applications (SEAA). (2015) 104–111
57. Sakhnini, V., Mich, L., Berry, D.M.: 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. (2013) `http://www.icb.uni-due.de/fileadmin/ICB/research/research_reports/ICB-Report-No56.pdf`.