Empir Software Eng (2017) 22:80–133
DOI 10.1007/s10664-015-9416-2

The impact of domain knowledge on the effectiveness of requirements engineering activities

Ali Niknafs[1] · **Daniel Berry**[1]

Published online: 30 April 2016

Abstract One factor that seems to influence an individual's effectiveness in requirements engineering activities is her knowledge of the problem being solved, i.e., domain knowledge. While in-depth domain knowledge enables a requirements analyst to understand the problem easier, she can fall for tacit assumptions and fail to consider issues that she believes to be obvious. This paper investigates the impact of domain knowledge on requirements engineering activities. Its main research question is "How does one form the most effective team, consisting of some mix of domain ignorants and domain awares, for a requirements engineering activity involving knowledge about the domain of the computer-based system whose requirements are being determined by the team?" For completeness, a number of other factors, such as educational background, are considered for their effect on teams' effectiveness. Two controlled experiments test a number of hypotheses derived from the question, including mainly that for a computer-based system in a particular domain, a team consisting of a mix of requirements analysts that are both ignorant and aware of the domain, is more effective at requirement idea generation than a team consisting of only analysts that are aware of the domain. The results show no significant effect of the mix by itself on effectiveness in requirement idea generation. However, the results do show surprising significant effects of the educational background and of the mix combined with the educational background. Combining the results, the main conclusion is that the presence of a domain ignorant

Communicated by: Daniel Amyot

Electronic supplementary material The online version of this article (doi:10.1007/s10664-015-9416-2) contains supplementary material, which is available to authorized users.

✉ Daniel Berry
dberry@uwaterloo.ca

Ali Niknafs
aniknafs@uwaterloo.ca

[1] David R. Cheriton School of Computer Science, University of Waterloo,
Waterloo, Ontario, Canada

with a computer science or software engineering background improves the effectiveness of a requirement idea generation team.

Keywords Requirements engineering · Requirements idea generation · Domain knowledge · Empirical software engineering

1 Introduction

A key step of any software development is deciding precisely what to build (Brooks 1995). The process of arriving at a set of features that need to be developed is referred to as *requirements engineering* (RE). The quality of the final product of a software development project depends on the extent to which the product satisfies its stakeholders' needs (Finkelstein 1994). Therefore, the more emphasis that is given to RE, the better the chances are of obtaining high quality software.

One of the challenges in RE is the huge gap between what the customer wants and what the analysts think the customer wants. To overcome this gap, it has long been believed that requirements analysts need to be experienced in the customer's problem domain to be productive when performing an RE activity (Al-Rawas and Easterbrook 1996; Jarke et al. 1993; Rose et al. 2009).

However, deep knowledge of the problem domain seems to lead to falling into the tacit assumption tarpit, including failing to investigate what is considered to be obvious (Berry 1995). Lack of domain knowledge might, in fact, have some benefits in RE activities. One such benefit has been observed by Berry (1995), namely the abilities of a domain ignorant to state his[1] ideas independently of any domain assumptions and to ask revealing questions that can lead to exposing issues that domain experts have overlooked. Domain ignorance is a good tool to surface the tacit assumptions of domain experts (Fischer 1999). This surfacing can lead to the necessary shared understanding of the topics of the tacit assumptions.

This paper describes two controlled experiments to test a number of hypotheses derived from the research question, "How does one form the most effective team, consisting of some mix of domain ignorants and domain awares, for a requirements engineering activity involving knowledge about the domain of the computer-based system whose requirements are being determined by the team?" The main hypothesis that is tested is that for a computer-based system in a particular domain, a team consisting of a mix of requirements analysts that are both ignorant and aware of the domain, is more effective at requirement idea generation than a team consisting of only analysts that are aware of the domain.

The results show no significant effect of the mix by itself on effectiveness in requirement idea generation by itself. However, the results do show a significant effect of the educational background and of the mix combined with the educational background.

It is hoped that the results of this study will help RE managers in forming more effective teams for doing requirement idea generation and other domain-knowledge-intensive RE activities and in making more effective use of the personnel available to them, by

- providing advice on the best mix for requirement idea generation of personnel who know the domain and who are ignorant of the domain,
- which domain ignorance is at least helpful, and

[1] Although a person could be a man or woman, we have assumed any nonspecific person is a man throughout the body of this paper.

- providing a useful role for new hires that allows them to be productive from the start while learning about the domain slowly without being a time drain on their mentors.

Section 2 of this paper describes related work. Sections 3 through 6 describe the experiment's design and method. Sections 7 through 11 describe and discuss the results of the experiment, including the threats. Sections 12 and 14 conclude the paper with a wrap up of the results, a comparison with results of previous work, and a description of future work.

The previous work described in Section 13 include our RE'12 paper (Niknafs and Berry 2012) and the first author's PhD thesis (Niknafs 2014). The subject of Section 13 is the differences between the results of this paper and those of the RE'12 paper. The PhD thesis has details omitted in this paper to keep this paper down to a reasonable size.

2 Background and Related Work

There is background and related work both in software engineering (SE) and RE and in the Humanities.

2.1 From Software and Requirements Engineering

Very few studies have investigated the impact of domain knowledge on SE activities. This section describes the relevant existing studies conducted in either academic or industrial settings.

Most SE research studies presume that domain knowledge is fundamental to an effective software development, and these studies do not assess whether this assumption holds. There is even no clear distinction between "knowledge" and "experience", as they are commonly used. The two are usually taken to mean the same thing. However, this study clearly distinguishes knowledge from experience.

Berry (1995) made one of the early observations of the benefits of domain ignorance as a result of his better-than-expected performances helping to write requirements specifications for software in two domains of which he was quite ignorant. As he noted later (Berry 2002), an even earlier observation of the impact of ignorance is from Burkinshaw's statement during the second NATO conference on SE in 1969 (Naur and Randell 1969):

> Get some intelligent ignoramus to read through your documentation and try the system; he will find many "holes" where essential information has been omitted. Unfortunately intelligent people don't stay ignorant too long, so ignorance becomes a rather precious resource. Suitable late entrants to the project are sometimes useful here.

In an experiment conducted on software design, Sharp (1991) defines three knowledge facets to design experience: 1) a designer's knowledge of the solutions to similar problems, 2) a designer's general knowledge of software design, and 3) a designer's knowledge of the application domain. Sharp's experiment was focused on the third facet. She found that the quality of the produced design is not affected by the designers' domain knowledge.

Kristensson et al. (2004) studied idea generation for a problem in the mobile technology domain using three types of participants: 1) advanced users who were CS students, 2) ordinary users who were non-CS students, and 3) professional product developers. The results obtained from this study showed that the ideas generated by ordinary users were considered more valuable by the authors than those generated by advanced users and professionals.

McAllister (2006) gathered studies about the role of user advocates in requirements determination. A user advocate is introduced during requirements determination to bridge the gap between the users and the developers, and to mediate the flow of events, which include the passing of information, between the two parties. McAllister observes that a user advocate understands the users' and developers' perspectives, but is not a domain expert in either the users' or developers' area. Therefore, a user advocate could help uncover tacit assumptions in the users' domain and could also avoid being influenced by a prior understanding of the developers' technology. Thus, in the vocabulary to be established for this paper, a user advocate is a domain ignorant in each of the users' and developers' domains.

Ferrari and Madhavji (2007) studied the impact of requirements knowledge and experience on software architecture tasks without considering domain knowledge. Their study suggests that architects with requirements knowledge and experience perform better than those without.

From a survey on requirements elicitation techniques, Dieste et al. (2008) concluded that a requirements analyst's experience with interviewing as an elicitation method and his experience with the problem domain does not affect the quantity of the ideas generated during an interview.

Carver et al. (2008) conducted a controlled experiment having two types of participants, those who have studied computer science (CS) as their university major and those who have studied something else. They observed that the general knowledge of CS did not improve the quality of the inspection, and the individuals in non-computing majors did even better than those in computing majors in detecting defects.

Mehrotra (2011) conducted a survey that showed that several activities are thought by experienced software development managers to be at least helped by domain ignorance. Based on the results obtained from the survey, Mehrotra categorized software development activities into three categories: 1) activities helped by domain ignorance, 2) activities not affected by domain ignorance, and 3) activities hindered by domain ignorance. Later, he showed, by mining histories reported by Dagenais et al. (2010) of immigrations of newbies to software development projects, a small positive correlation between a successful immigration for a newbie and the newbie's assignment to tasks that are thought to be at least helped by domain ignorance. Here, the term "newbie" comprises new hires and existing employees assigned to new projects.

One of the results of Mehrotra's work is that for requirements documents inspection, domain awareness is considered to be necessary, but domain ignorance is considered also to be helpful. For other inspection activities, e.g., of test plans and user manuals, both domain ignorance and domain awareness were considered to be helpful. These results seem to imply that a team with a mix of domain ignorance and awareness might be more effective at inspection than a team with no mix.

Hadar et al. (2014) studied the effect of domain knowledge on conducting interviews and on the preferences for different elicitation techniques throughout the elicitation process. They determined that those without domain knowledge can be effective in interviews. They did not explore the specific effect of an analyst's prior domain knowledge.

So, the SE and RE fields are slowly coming to the recognition that domain knowledge is an important factor in software development and that domain ignorance may even be beneficial.

2.2 From Humanities

There is a lot more work on the effect of domain ignorance in the Humanities. Here is a sampling of it:

Luchins (1942) demonstrated experimentally the *Einstellung effect* as the tendency for a person to apply to a new problem, solutions that worked on related problems in the past and to continue to do so even in the face of the failure of these past solutions to solve the new problem. Luchins and Luchins (1950) later showed that giving the simple instruction, “Don’t be blind” helped to extinguish the Einstellung effect.

Wiley (1998) observed conditions that may inhibit creative problem solving, by conducting three separate experiments. Her hypothesis was that domain knowledge may act as a mental set and as a result, domain knowledge promotes fixation in problem-solving attempts. She observed that high-knowledge subjects were biased toward their first solution attempts, and therefore, their knowledge stopped them from investigating a broad range of solutions. She observed also that the low-knowledge subjects were more successful in arriving at appropriate solutions. She concluded that two conditions gave rise to this effect for a given task:

1. the task involves creative problem solving, and
2. the task is not so complex that it inhibits novices from participating.

Dunbar (1999) studied how scientists study things in practice. He found that over half of the data that scientists find are unexpected. What do they do with the unexpected data? They find an excuse and ignore it altogether. Lehrer (2009) puts it in another way; we interpret the results of an experiment the way that we want to see it and disregard what we do not want to see. Based on Dunbar’s findings, Lehrer suggests four ways of dealing with the unexpected data:

1. *Check your assumptions:* Maybe the experiment is correct, the hypothesis is not.
2. *Seek out the ignorant:* Explain your work to people ignorant about your work. It might make clear some aspects that you were not looking at before.
3. *Encourage diversity:* Nowadays, in any scientific study, groups of scientists do the reasoning about the results instead of individual scientists (Thagard 1997; Dunbar 1999). This situation is called also *distributed reasoning* (Dunbar 1999). The reason is that people with the same knowledge about a domain have the same assumptions and, therefore, expect the same sort of results and do the same sort of reasoning about the results.
4. *Beware of failure-blindness:* There is always the risk of the bias toward rejecting unexpected results in order to reject failure.

Firestein (2013) teaches a course called *Ignorance* at Columbia University. He invites scientists from different disciplines, including biology and biomedical sciences, psychology, chemistry, physics, mathematics and statistics, computer science, and earth sciences, to give lectures in the class. Each lecture is a case study in which the invited scientist discusses the recent problems he is working on. Then, the speaker and students discuss the role of ignorance in driving the scientist’s research. Firestein promotes the idea that ignorance is not something that will be transformed into knowledge, it is knowledge that transforms ignorance into higher quality ignorance. This is what Pascal refers to as natural ignorance and learned ignorance (Pascal and Krailsheimer 1968):

> The world is a good judge of things, for it is in natural ignorance, which is man's true state. The sciences have two extremes which meet. The first is the pure natural ignorance in which all men find themselves at birth. The other extreme is that reached by great intellects, who, having run through all that men can know, find they know nothing, and come back again to that same ignorance from which they set out; but this is a learned ignorance which is conscious of itself.

Apfelbaum et al. (2014) compared the effects of homogeneity and diversity in groups. They found that homogeneity in a team led to more subjectivity in an individual's judgements. On the other hand, diversity in a group led to an increase in the individual's objectivity. Therefore, the authors suggest to further study the potential effects of diversity in a team.

3 Context

The context of the research described in this paper is requirement idea generation for some *computer-based system (CBS)* for some *client*. The CBS is situated in some *domain*, and generally, at least one member of the client's organization is *aware of* and is often expert in this domain.

It is assumed that each member of the software development organization doing the requirement idea generation is at least competent in his development roles. However, each such member has a different amount of *knowledge about the domain*. In some cases, the member is *ignorant of the domain*, i.e., is a *domain ignorant (DI)*. In other cases, the member is *aware of the domain*, i.e., is a *domain aware (DA)*. Each of domain ignorance and domain awareness is a kind of *domain familiarity*.

While in real life, the boundary line between domain ignorance and domain awareness is fuzzy, conducting experiments depending on the distinction requires making sure that no participant is both and that it is possible to easily classify each participant as one or the other. Therefore, the study described herein strived to find a way to make the distinction between domain ignorance and domain awareness sharp.

4 Research Questions

Following the Goal-Question-Metric template (Basili et al. 1994), the goal of this research is to improve the effectiveness of the RE process from the viewpoint of project managers, in the context of both laboratory projects and real-world projects. Given this goal, the main research question (RQ) to answer is:

> *How does one form the most effective team, consisting of some mix of DIs and DAs, for an RE activity involving knowledge about the domain of the CBS whose requirements are being determined by the team?*

Answering this RQ properly requires particularizing the question to one activity in RE. One of these activities is requirement idea generation during requirements elicitation.

The major RQ can be decomposed into two specific RQs:

RQ_1 *Does a team consisting of a mix of DIs and DAs perform requirement idea generation more effectively than a team consisting of only DAs?*

RQ_2 *Do factors other than a team's mix of DIs and DAs impact the effectiveness of the team's performing requirement idea generation?*

Effectiveness in requirement idea generation is measured by both the quantity and quality of the ideas generated, as is described in Section 6.3.

The effect of domain knowledge cannot be assessed in isolation, since there are confounding factors that need to be considered. These factors include educational background, industrial experience, and experience in RE. Creativity is another factor to be considered, since it plays an important role in idea generation activities, such as brainstorming.

5 Main Hypotheses

The main hypothesis coming from the RQs is:

> *A team consisting of a mix of DIs and DAs is more effective in requirement idea generation than is a team consisting of only DAs.*

The corresponding null hypothesis is:

> *The mix of DIs and DAs in a team has no effect on the team's effectiveness in requirement idea generation.*

The corresponding non-directed alternative hypothesis is:

> *The mix of DIs and DAs in a team has an effect on the team's effectiveness in requirement idea generation.*

6 Experiment Design

This section explains the design of controlled experiments (Wohlin et al. 2000) that aimed to answer the RQs and test the hypotheses.

The experiment design described in this section has been applied in two separate experiments, E1 and E2. The results of E1 were reported in a conference paper written by the same authors (Niknafs and Berry 2012). E1's results were that there was some support for accepting the main hypothesis. However, E1 suffered from (1) a small number, 19, of teams and (2) an imbalance in the numbers of teams with each mix of domain familiarity, with 9 teams with only DIs and 3 or 4 teams with each other mix. The small number and the imbalance reduced the statistical strength of the results. E2 was conducted to provide 21 more teams, for a total of 40, and to balance the number of teams, 10, with each mix of domain familiarity. This paper reports on the analysis of the combined data from E1 and E2.

6.1 Pilot Studies, Lessons Learned, and Domain Selection

While controlled experiments are probably the most effective method by which to validate a hypothesis, it is usually very difficult to foresee all the factors that need to be taken into consideration. Thus, before conducting E1, two pilot studies, whose results were destined to be ignored, were conducted as completely as possible in order to identify defects in the design of the experiment and generally to improve that design.

The main lesson learned from the pilot studies was that finding a suitable CBS with a suitable domain to use in experiments was critical. The CBS chosen for the first pilot study was a requirements tracing tool, while for the second pilot study, the CBS chosen was a university admissions system. Domains in CS or university administration were too familiar to the participant population of university students that are competent in CS. For such domains, it is hard to build teams with DIs. It was clear that we needed a domain outside CS, e.g., health informatics. In addition, in the pilots, even self-reported DIs had *some* knowledge of the tracing and admissions domain. So, it was hard to classify participants as either DI or DA. There were too many participants who would be somewhere in the middle of being a DI and being a DA. Thus, the domain has to be so far out of CS that each competent software developer would be either totally ignorant or totally aware of it. Health Informatics would not be suitable on this basis.

One day, Berry realized that he and Niknafs shared knowledge of a domain that very few computer scientists and software developers in North America knew anything about: bidirectional word processing. Each of us spoke a language that is written from right to left, Persian for Niknafs and Hebrew for Berry. A document in each of these languages about high technology uses terminology in e.g., English, that is written from left to right. Moreover, in each of Arabic, Hebrew, Persian, and Urdu, a numeral is written from left to right. So, we agreed that the application for which requirement ideas would be generated would be a bidirectional word processor (BDWP). Any computer scientist from the Middle East would likely be a DA, and any computer scientist from elsewhere would likely be a DI. The expected few exceptions were easily identified and classified correctly by asking a few questions. Moreover, the division of participants would likely be sharp; there would probably not be anyone that was neither one nor the other. In fact, it is even hard to conceive of a person who could be classified as both.

6.2 Participants and Composition of Teams

Participants in E1 were all CS and SE students. Because not many of these students spoke any right-to-left language, most teams consisted of only DIs. For E2, we decided to allow participants other than CS and SE students. We knew that this decision would introduce new variables to the study, but it was the only option left at the time. We had exhausted the pool of potential CS and SE volunteer participants and would have had to wait another year for a new batch of students to arrive. We did, however, insist that each non-CS-or-SE participant be in some high technology field, such as Electrical Engineering.

Each team consisted of three members. The team size was chosen as a compromise. The smaller the team size, the more teams we can squeeze out of a participant pool. However, a team of size one is not a team. A team of size two does not allow one kind of domain familiarity to be in a majority. So, we chose a team size of three as the smallest size in which in every configuration, one domain familiarity has a majority.

6.3 Classification of Generated Ideas

The goal of the controlled experiments is to discover the effect of a team's mix of DIs and DAs on the team's performance in requirement idea generation. Since the stated goal of the first stage of brainstorming is to generate as many ideas as possible, the number of raw ideas generated by each team serves as a good quantitative measure. However, in order to better compare the performance of the teams, we considered also the quality of their generated ideas. Based on the characteristics of a good requirement in the IEEE 830 Standard (Berenbach et al. 2009), we decided to classify each idea according to three characteristics:

Relevancy An idea is considered relevant if it has something to do with the domain.

- “the ability to embed left-to-right (LR) numerals within right-to-left (RL) text” is relevant, but
- “the ability to play MP3 files” is not relevant.

Feasibility An idea is considered feasible if it is relevant and it is correct, well presented, and implementable.

- “the ability to embed left-to-right (LR) numerals within right-to-left (RL) text” is relevant and also feasible, but
- “the ability to translate Arabic text into English” is relevant but not feasible.

Innovation An idea is considered innovative if it is feasible and it is not already implemented in an existing application for the domain known to the classifier.

- “the ability to show the logical-order and visual-order views of a document simultaneously in two different editable and continually updated windows so that a change to one is reflected in the other”[2] is relevant, feasible, and innovative, since no existing BDWP known to any classifier has this feature, but
- “the ability to embed left-to-right (LR) numerals within right-to-left (RL) text” is relevant, feasible, but not innovative, since all existing BDWPs have this feature.

We decided to use ourselves, both experts in the BDWP domain, as idea classifiers. To eliminate any bias in classifying an idea that might arise from a classifier’s knowing the domain familiarity mix of the team from which the idea came, we decided to produce a list of all ideas generated by all teams, sorted using the first letters of each idea. Each domain-expert classifier would then classify the ideas in the full list. Once both classifications are done, each classifier’s classifications of each idea would be transferred to the idea’s occurrences in the individual team lists. Then, the average of the numbers of the ideas generated by a team in each classification, as determined by the classifiers, is used as the value of the classification for the team.

Later, we hired by the hour a third classifier, an Arabic and Hebrew speaker. We had discovered almost unanimous agreement, 99 %, over which ideas were relevant and feasible, but only 89.2 % agreement over which feasible ideas were innovative. So, to save money that we needed to pay participants, while getting the most classification bang for each buck, we had the third classifier evaluate only the innovativeness of the union of our feasible ideas. We felt that having the third classifier classify also the relevance and feasibility of the raw ideas would quadruple the time she spent without showing us anything new.

[2] In a Unicode-compliant BDWP, knowing the time-ordered, logically-ordered internal representation of the currently displayed visually-ordered text helps the user predict the effect of any editing change enacted on the displayed view, particularly since the internal representation may have so-called zero-width control characters that are *invisible* in the visual-order view.

6.4 Procedure

The experiment is divided into two parts. In the first part, each participant was asked to fill out a questionnaire about his education level, RE experience, industrial experience, and familiarity with the bidirectional word processing domain. Each was asked also to take the Williams creativity test (Taylor and Williams 1965) to detect the presence of significant differences in personal creativity. The gathered creativity scores would be used to balance the teams based on their average creativity scores. The information gathered in the participants' first parts allowed forming teams. Each team had one particular needed mix of DIs and DAs, and each was invited to attend a second part.

In the second part, each team attended a one-half hour lecture on reading bidirectional text. The lecture was about the basics of reading and writing text written in right-to-left languages, particularly when it is mixed with text written in left-to-right languages. The lecture described possible ways of storing and displaying bidirectional text in existing word processors.

After the lecture, the team members were reminded about brainstorming and how the focus of the first part of brainstorming is on generating as many ideas as possible, i.e., "quantity over quality".

Finally, each team participated in its own one-half hour first part of a brainstorming for ideas for requirements for the BDWP. Each team was given a laptop or a desktop computer into which to type its ideas. Ideas, one per line, were entered in unstructured natural language.

A copy of the materials for conducting this procedure can be found at https://cs.uwaterloo.ca/~dberry/FTP_SITE/NiknafsBerryMaterials/.

6.5 Variables

Values for several independent and dependent variables were gathered during the experiments about each team performing requirement idea generation for a CBS in a domain.

6.5.1 Independent Variables About a Team

The independent variables about each three-person *team* were determined from the goals, RQs, and lessons learned from the pilot studies:

- *Mix of Domain Familiarities (MIX)*: The team's MIX value is of the form *nI*, where *n* is the number of DIs it has; thus, the value is one of *0I*, *1I*, *2I*, and *3I*. The *0I* mix is designated as the control mix, since it corresponds to the norm of practice. That is, most managers would staff a three-person team as *0I*. However, in these experiments with the RQs and hypotheses as stated, there is no control in the sense of establishing a baseline. We were trying to determine which mix gives the best performance.
- *Creativity (CR)*: The team's CR value is the average of the team members' creativity scores.
- *RE Experience*: The team's RE experience is divided into two subvariables in order to differentiate between overall RE experience and industrial RE experience:
    - *Overall RE Experience (REXP)*: the average number of both academic and industrial RE projects the members of the team have done in the past, and
    - *Industrial RE Experience (IREXP)*: the average number of industrial RE projects the members of the team have done in the past.

- *Industrial Experience (IEXP)*: The team's IEXP value is the average number of years of industrial software development experience of the members of the team.
- *Educational Background*: The team's educational background is divided into three subvariables in order to expose the strength of the team's CS or SE background:
    - *Number of CS student members (NCS)*: the number, between 0 and 3, of members in the team who are CS students.
    - *Number of SE student members (NSE)*: the number, between 0 and 3, of members in the team who are SE students.
    - *Number of graduate student members (NGRAD)*: the number, between 0 and 3, of members in the team who are graduate students.

6.5.2 Dependent Variables About a Team

The dependent variables about a team are based on the classifications of the requirement ideas described in Section 6.3:

- *Raw number of ideas (RAW)*: the raw number of ideas that the team generated for the CBS used in the experiment,
- *Average number of relevant ideas (AVG_R)*: the average of the numbers of relevant ideas the two classifiers thought that the team generated for the CBS used in the experiment,
- *Average number of feasible ideas (AVG_F)*: the average of the numbers of feasible ideas the two classifiers thought that the team generated for the CBS used in the experiment, and
- *Average number of innovative ideas (AVG_I)*: the average of the numbers of innovative ideas the three classifiers thought that the team generated for the CBS used in the experiment.

With these specific variables, the effectiveness of a team in generating requirement ideas of any type is measured by the number of that type of ideas that the team generated during its half-hour requirement idea generation session.

7 Gathered Data

For each experiment, E1 or E2, a single list of all ideas generated by all teams was created. Using the procedure described in Section 6.3, two domain experts classified the ideas in each list for relevance and feasibility. Three domain experts classified the ideas in each list for innovativeness. When the classifications for E2 were finished, the data from E2 were combined with the data from E1. A Pearson test was employed to find the correlations between the pairs of classifications. The results, shown in Table 1, demonstrate that the classifications of the first two classifiers have a strong correlation ($p < 0.05$). Also the classifications of the third classifier have strong correlations with those of each of the two other classifiers.

Since the results of E1 and E2 are combined for the purpose of analysis, the correlation between the classifiers' classifications between E1 and E2 must be computed. All that really matters are the numbers of ideas of each type, since only these numbers are used in the analysis about the various types of ideas. Therefore, we decided to compare the ratios of the numbers of relevant, feasible, and innovative ideas to the number of raw ideas for E1 and E2.

Table 1 Correlation Between the Classifiers' Classifications of Ideas

	Ideas				
	Relevant	Feasible	Innovative		
			(C1,C2)*	(C1,C3)*	(C2,C3)*
Pearson Correlation	.977	.993	.987	.905	.851
Significance	.000	.000	.000	.000	.000

*C1: Classifier 1, C2: Classifier 2, C3: Classifier 3

As shown in Table 2, the differences between the E1 and E2 ratios for the relevant and feasible ideas are clearly significant. Perhaps, the classifiers were less conservative for E2 ideas than they were for E1 ideas. Perhaps the difference in the educational background of the participants was the factor, i.e., CS and SE students are less capable of identifying relevant and feasible ideas than other high technology students. Whatever the reason, a possible threat to combining the two experiments and conducting the analysis on the combined data is the difference between the classifications for relevant and feasible ideas in the two experiments. This threat is considered in detail in Section 11. Also, the statistical tests in Section 10 that see whether the NCS and NSE independent variables have an effect on the dependent variables allow determining whether the difference in the educational background is a factor in the classification differences shown in Table 2.

8 Data Preparation for Statistical Analysis

Prior to statistical analysis, the data from E1 and E2 were combined. Information about the combined set of participating teams is shown in Table 3, and a summary of the classifications of their generated ideas is shown in Table 4. These combined data were then subjected to various tests, conversions, or transformations (Warner 2012).

8.1 Data Nominalization

The values of some of the independent variables about a team were converted into nominal values and others were left unchanged. To do any nominalization, we plotted the independent variable's data, and we looked for gaps and big deltas in values to identify natural groupings of actual values into nominal values, as described below explicitly for CR.

Table 2 Ratios of the Classified Data to the Number of Raw Ideas between E1 and E2

Classifier	Experiment	Ideas		
		Relevant	Feasible	Innovative
C1	E1	.27	.20	.04
	E2	.59	.26	.03
C2	E1	.28	.20	.03
	E2	.57	.27	.03

Table 3 Combined Data about the Teams

		Creativity Score		RE Experience		Industrial RE Experience		Industrial Experience		No. CS Participants		No. CE Participants		No. Graduate Participants	
Mix of Teams	No. of Teams	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
0I	10	1.70	.48	2.00	.94	.70	.82	.90	.57	1.00	1.25	.10	.32	2.70	.48
1I	10	1.80	.63	2.40	.97	1.30	1.16	1.90	1.10	1.90	.88	1.10	.88	2.40	.70
2I	10	2.10	.57	1.50	.97	1.10	1.10	1.60	.70	2.00	1.25	1.40	1.17	2.00	1.33
3I	10	2.00	.00	1.30	.82	1.00	.87	1.80	.79	3.00	.00	2.90	.32	.10	.32
Total	40	1.90	.50	1.80	.99	1.03	.97	1.55	.88	1.98	1.19	1.38	1.25	1.80	1.28

Table 4 Combined Data of the Generated Ideas

	Number of Raw Ideas			Number of Relevant Ideas			Number of Feasible Ideas			Number of Innovative Ideas		
Mix of Teams	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
0I	23.50	12.00	27.86	10.30	7.25	11.85	5.05	1.75	9.75	1.63	.33	2.99
1I	16.50	16.00	10.55	8.40	8.25	3.71	4.60	3.50	3.39	1.10	.50	1.53
2I	18.30	17.00	12.37	8.60	6.75	6.58	4.20	3.50	3.46	.57	.33	.69
3I	31.90	31.50	25.48	8.15	7.50	4.89	6.35	4.50	4.11	1.60	.67	1.93
Total	22.55	16.50	20.65	8.86	7.25	7.20	5.05	3.50	5.65	1.22	.50	1.94

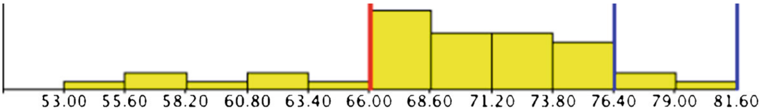


Fig. 1 Distribution of the Teams' Average Creativity Scores of the Participating Teams

- *MIX*: the team's MIX has one of the four possible nominal values, *0I*, *1I*, *2I*, or *3I*, where the digit of the value is the number of DIs the three-person team has.
- *CR*: as shown in Fig. 1, the Williams creativity scores were distributed so that scores in the range of 66 through 76.40 were the central part of the distribution. Therefore, each score was converted into a nominal value:
  - Low: for a score less than 66,
  - Medium: for a score between 66 and 76.40 inclusive, and
  - High: for a score greater than 76.40.
- *REXP*: based on the distribution shown in Fig. 2, each number was converted into a nominal value:
  - None: for a number equal to zero,
  - Low: for a number less than 0.67,
  - Medium: for a number between 0.67 and 1.33 inclusive, and
  - High: for a number greater than 1.33.
- *IREXP*: based on the distribution shown in Fig. 3, each number was converted into a nominal value:
  - None: for a number equal to zero,
  - Low: for a number less than 0.40,
  - Medium: for a number between 0.40 and 1.06 inclusive, and
  - High: for a number greater than 1.06.
- *IEXP*: based on the distribution shown in Fig. 4, each number was converted into a nominal value:
  - None: for a number equal to zero,
  - Low: for a number less than 0.67,
  - Medium: for a number between 0.67 and 1.33 inclusive, and
  - High: for a number greater than 1.33.
- *NCS*: the number, between 0 and 3, of members in the team who are CS students.
- *NSE*: the number, between 0 and 3, of members in the team who are SE students.
- *NGRAD*: the number, between 0 and 3, of members in the team who are graduate students.

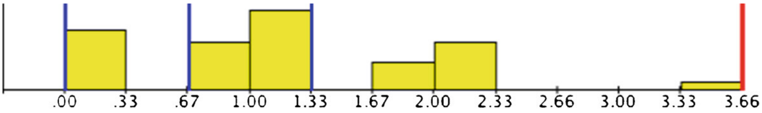


Fig. 2 Distribution of the Teams' Average RE Experience

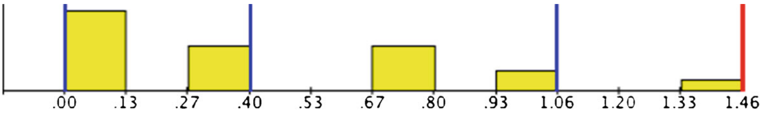


Fig. 3 Distribution of the Teams' Average Industrial RE Experience

Table 5 summarizes the variables of the experiment. (It includes variables introduced in later subsections.)

8.2 Data Normalization

In order to apply an ANOVA, the data needed to be normal. Table 6 shows the results of the two normalization tests, i.e., Kolmogorov-Smirnov and Shapiro-Wilk, indicating significant p-values of less than 0.05. Thus, none of the dependent variables are normally distributed. Therefore, an ANOVA officially cannot be used.

On the other hand, an ANOVA is not very sensitive to moderate deviations from normality. However, it has been shown that the severity of the affects of non-normality on an ANOVA is amplified by kurtosis and skewness of the data (Glass et al. 1972), which need to be considered beside normality. Table 7 shows that all standard skewness and kurtosis scores are outside of the acceptable ranges.

Because the dataset was surely non-normal, with extreme skewness and kurtosis, it needed to be transformed in order to use an ANOVA. Blom's formula (Blom 1960) is a rank-based method that can be used to normalize non-normally distributed data. Table 8 shows the normalized versions of the data in Table 4.

Table 9 shows that all the dependent variables, except NI, were successfully transformed into normal distributions. For NI, the Kolmogorov-Smirnov test result is 0.008 and the Shapiro-Wilk test result is 0.007, each of which is less than 0.05. Therefore, NI is not normalized.

Skewness and kurtosis are calculated once again. Table 10 shows that the skewness and kurtosis standard scores for all four dependent variables are within the acceptable range, even for NI.

Figure 5 shows on the left side, the plots for the original data for the dependent variables and on the right side, the plots for the normalized versions of the original data. It is evident that normalization has worked very well in transforming the data into normal distributions.

After normalization, each of the NRAW, NR, and NF distributions appears to more or less satisfy the normality requirement for an ANOVA. Although the normality tests showed that NI's distribution is not normal, it passes the skewness and kurtosis tests. Therefore, we decided to apply an ANOVA to all dependent variables, and then, as an insurance policy, to apply to the original unnormalized AVG_I data a non-parametric test, which does not require the data to be normally distributed.

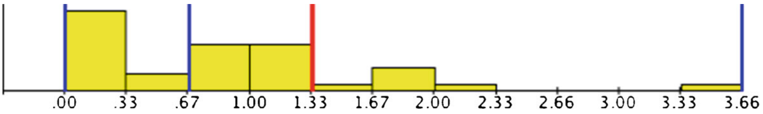


Fig. 4 Distribution of the Teams' Average Industrial Experience

Table 5 Variables of the Study

Name	Independent Variable About a Team	Values
MIX	Mix of domain familiarities	*0I*, *1I*, *2I*, *3I*
CR	Average creativity score level	Low, Medium, High
REXP	Average RE experience	None, Low, Medium, High
IREXP	Average industrial RE experience	None, Low, Medium, High
IEXP	Average industrial experience	None, Low, Medium, High
NCS	Number of participants with CS background	0, 1, 2, 3
NSE	Number of participants studying SE	0, 1, 2, 3
NGRAD	Number of graduate student participants	0, 1, 2, 3
Name	Dependent Variable About a Team	Values
RAW	Raw number of ideas	Numeric
NRAW	Normalized RAW	Numeric
AVG_R	Average number of relevant ideas	Numeric
NR	Normalized AVG_R	Numeric
AVG_F	Average number of feasible ideas	Numeric
NF	Normalized AVG_F	Numeric
AVG_I	Average number of innovative ideas	Numeric
NI	Normalized AVG_I	Numeric

Table 6 Test of Normality of the Dependent Variables

Dependent Variable	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	p	Statistic	df	p
RAW	.211	40	.000	.752	40	.000
AVG_R	.212	40	.000	.666	40	.000
AVG_F	.214	40	.000	.691	40	.000
AVG_I	.287	40	.000	.646	40	.000

Table 7 Skewness and Kurtosis Test Results of the Dependent Variables

	RAW	AVG_R	AVG_F	AVG_I
N	40	40	40	40
Skewness	2.304	3.319	3.152	2.708
Std. Error of Skewness	.374	.374	.374	.374
Std. Score of Skewness	6.160	8.874	8.428	7.241
Kurtosis	6.26	14.021	13.771	8.671
Std. Error of Kurtosis	.733	.733	.733	.733
Std. Score of Kurtosis	8.540	19.128	18.787	11.829

Table 8 Normalized Combined Data of the Generated Ideas

	Normalized No. of Raw Ideas			Normalized No. of Relevant Ideas			Normalized No. of Feasible Ideas			Normalized No. of Innovative Ideas		
Mix of Teams	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
0I	−0.01	−0.42	0.96	0.00	−0.02	0.99	−0.41	−0.64	1.11	0.09	−0.28	1.00
1I	−0.24	−0.05	0.82	0.16	0.28	0.87	0.06	0.06	0.82	−0.07	−0.03	0.98
2I	−0.13	0.06	0.94	−0.04	−0.17	0.84	−0.15	0.06	0.99	−0.26	−0.28	0.80
3I	0.38	0.70	1.18	−0.11	−0.02	1.27	0.51	0.25	0.80	0.34	0.22	0.85
Total	0.00	0.00	0.98	0.00	−0.02	0.97	0.00	0.00	0.97	0.03	−0.03	0.91

Table 9 Test of Normality of the Dependent Variables after Normalization

Dependent Variable	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	p	Statistic	df	p
NRAW	.041	40	.200	.997	40	1.000
NR	.054	40	.200	.994	40	.998
NF	.106	40	.200	.984	40	.844
NI	.165	40	.008	.919	40	.007

When the preconditions of ANOVA are not met, a non-parametric substitute for an ANOVA should be applied. The most common substitute is the Kruskal-Wallis test, which compares independent samples using medians instead of means, as does the ANOVA test. When the distributions are normal, one-way ANOVA is the most powerful method to detect differences among the means (BBN Technologies 2015). The Kruskal-Wallis test is a powerful method to detect differences among the medians. To be thorough, with analyses based on both the means and the medians, we normalized the data and applied ANOVA to the normalized data. Then, as a totally separate analysis, we applied the Kruskal-Wallis test to the original non-normalized data.

8.3 Outliers

Boxplots are used to detect potential outliers. Figure 6 shows the boxplots of the four dependent variables grouped by the main independent variable of the study, MIX. Figure 6a shows that the values of RAW for Teams 8, 16, 32, and 34 are outliers. Figure 6b shows that the values of AVG_R for Teams 32 and 34 are outliers. Figure 6c shows that the value of AVG_F for Team 34 is an outlier. Figure 6d shows that the values of AVG_I for Teams 19, 20, 24, and 34 are outliers. A deeper study of the outliers, strengthening these conclusions, can be found in Online Appendix A.

The analysis described hereafter was done on two sets of data: 1) on the data including the outliers, and 2) on the data without the outliers. Whenever outliers were removed prior to a study, the results are marked as “Filtered”. Otherwise, the results are marked as “Unfiltered”.

Table 10 Skewness and Kurtosis Test Results for the Dependent Variables after Normalization

	NRAW	NR	NF	NI
N	40	40	40	40
Skewness	.003	.005	.047	.367
Std. Error of Skewness	.374	.374	.374	.374
Std. Score of Skewness	.008	.013	.126	.981
Kurtosis	−.279	−.28	−.321	−.653
Std. Error of Kurtosis	.733	.733	.733	.733
Std. Score of Kurtosis	−.381	−.382	−.438	−.891

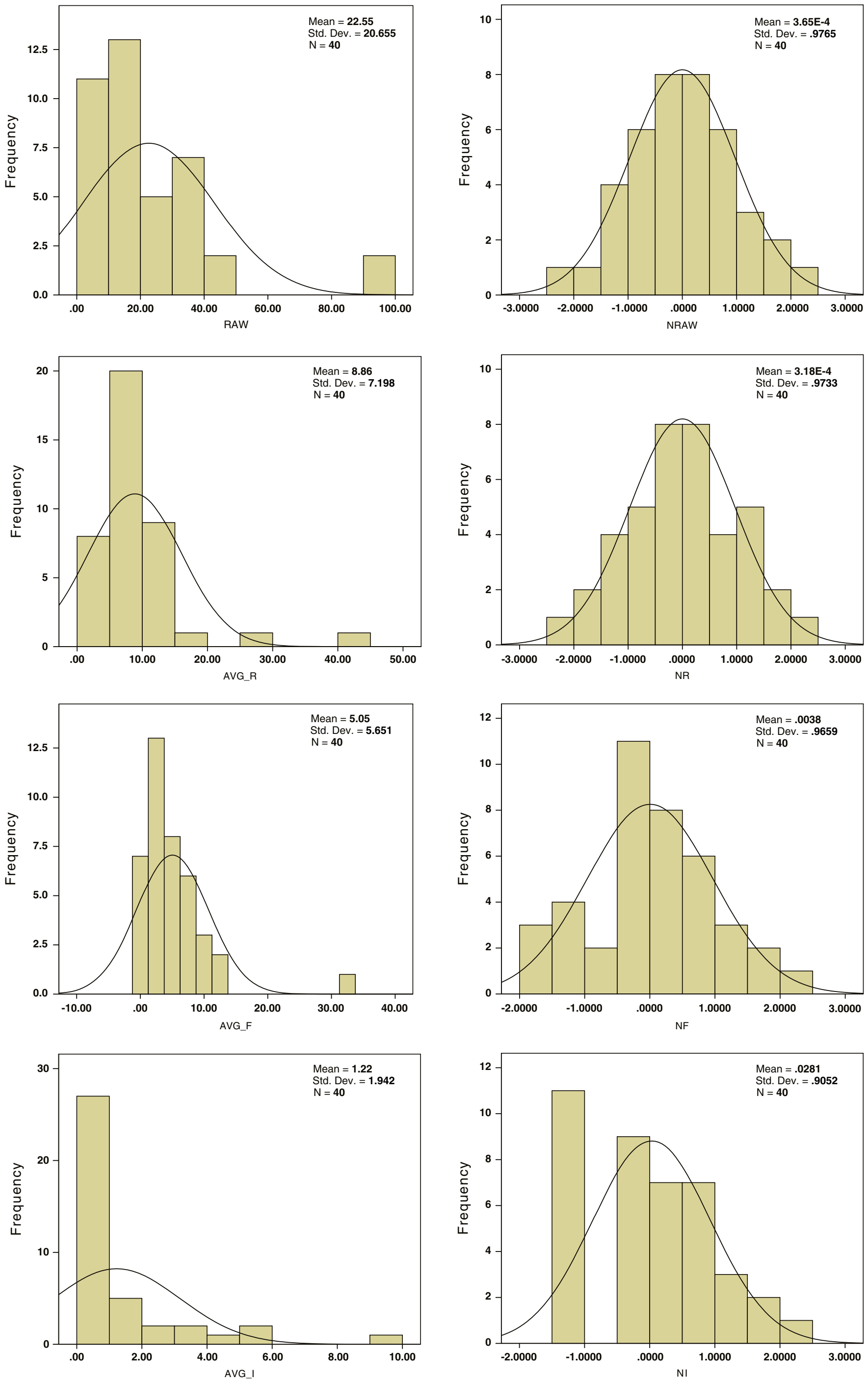
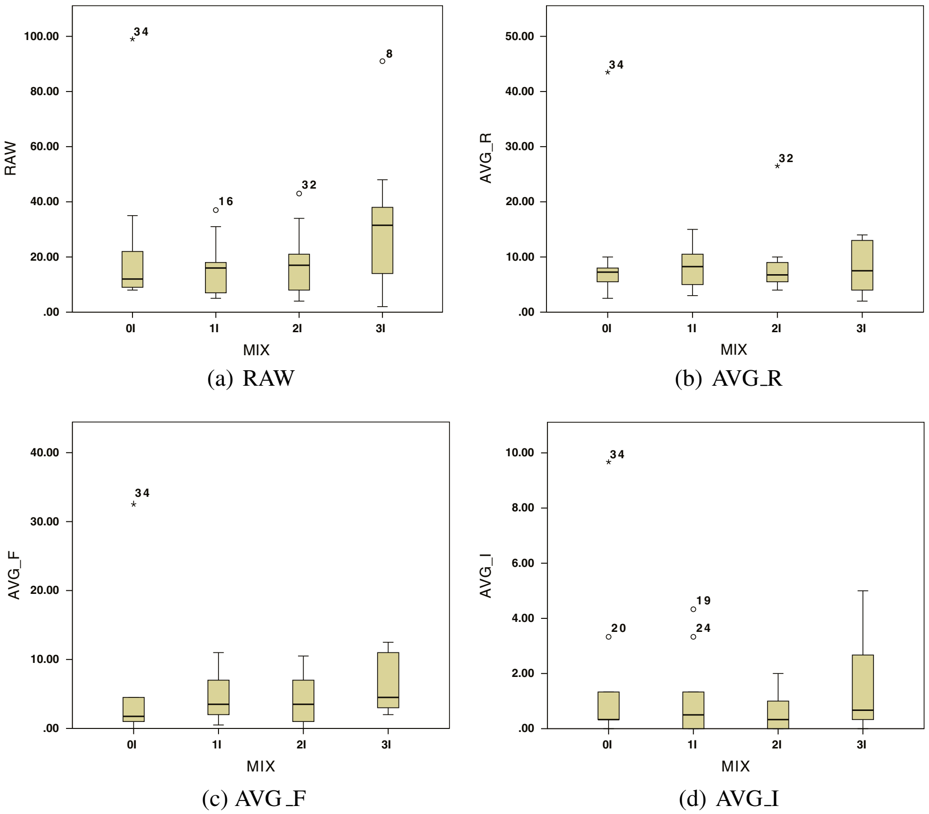


Fig. 5 Normality Plots of the Dependent Variables



(a) RAW

(b) AVG_R

(c) AVG_F

(d) AVG_I

Fig. 6 Boxplots of the Dependent Variables

Table 11 Rotated Factor Matrix

Independent Variables	Factor	
	1	2
CR	.147	.225
REXP	−.410	.625
IREXP	.055	.851
IEXP	.261	.705
NSE	.951	.278
NGRAD	−.877	.050
NCS	.783	.145

¯Extraction Method: Principal Axis Factoring.

¯Rotation Method: Equamax with Kaiser Normalization.

Table 12 Final List of Hypotheses

Identifier	Hypothesis
H_{MIX_1}	The effectiveness of a team in requirement idea generation is affected by the team's mix of domain familiarities.
H_{MIX_0}	The effectiveness of a team in requirement idea generation is not affected by the team's mix of domain familiarities.
H_{CR_1}	The effectiveness of a team in requirement idea generation is affected by the team's creativity level.
H_{CR_0}	The effectiveness of a team in requirement idea generation is not affected by the team's creativity level.
H_{EXP_1}	The effectiveness of a team in requirement idea generation is affected by the team's EXP value.
H_{EXP_0}	The effectiveness of a team in requirement idea generation is not affected by the team's EXP value.
H_{REXP_1}	The effectiveness of a team in requirement idea generation is affected by the team's average number of academic and industrial RE projects the team members have done in the past.
H_{REXP_0}	The effectiveness of a team in requirement idea generation is not affected by the team's average number of academic and industrial RE projects the team members have done in the past.
H_{IREXP_1}	The effectiveness of a team in requirement idea generation is affected by the team's average number of industrial RE projects the team members have done in the past.
H_{IREXP_0}	The effectiveness of a team in requirement idea generation is not affected by the team's average number of industrial RE projects the team members have done in the past.
H_{IEXP_1}	The effectiveness of a team in requirement idea generation is affected by the team's average number of years of industrial software development experience of the team members.
H_{IEXP_0}	The effectiveness of a team in requirement idea generation is not affected by the team's average number of years of industrial software development experience of the team members.
H_{EDU_1}	The effectiveness of a team in requirement idea generation is affected by the team's EDU value.
H_{EDU_0}	The effectiveness of a team in requirement idea generation is not affected by the team's EDU value.
H_{NCS_1}	The effectiveness of a team in requirement idea generation is affected by the team's number of CS student members.
H_{NCS_0}	The effectiveness of a team in requirement idea generation is not affected by the team's number of CS student members.
H_{NSE_1}	The effectiveness of a team in requirement idea generation is affected by the team's number of SE student members.
H_{NSE_0}	The effectiveness of a team in requirement idea generation is not affected by the team's number of SE student members.
H_{NGRAD_1}	The effectiveness of a team in requirement idea generation is affected by the team's number of graduate student members.
H_{NGRAD_0}	The effectiveness of a team in requirement idea generation is not affected by the team's number of graduate student members.

Table 13 Results of the Levene Test for MIX (Unfiltered)

Dependent Variable	Levene Statistic [a]	df1 [b]	df2 [c]	p [d]
NRAW	.450	3	36	.719
NR	1.838	3	36	.158
NF	.174	3	36	.913
NI	.427	3	36	.735

[a] Numeric Levene test results
[b] Degrees of freedom 1
[c] Degrees of freedom 2
[d] p-value

8.4 Factor Analysis

As a statistical method, factor analysis is used to shrink a large number of independent variables to a potentially smaller set of non-observed variables called *factors*[3]. to be the main driver behind the dependent variables (Hanebutte et al. 2003). Omitted from the set is any so-called independent variable that is found to be dependent on others.

There are eight independent variables in this study, listed in Table 5. Since MIX is the main variable of the study, it was left out of the factor analysis, and the analysis was performed on the remaining seven variables. After the factor analysis, MIX will be added to those variables that are grouped by the analysis to be further studied in depth.

First, it is necessary to test whether the chosen set of variables is adequate for factor analysis (Hinton et al. 2004). One such test is the Kaiser-Meyer-Olkin (KMO) measure. The KMO measure of the chosen set of independent variables is 0.656, which is greater than 0.5. Therefore, the set is adequate for factor analysis.

Principal Factor Analysis (PFA)[4] is the most common method used in social sciences (Warner 2012) to find a smaller number of factors to examine.

The results of the factor analysis are shown in Table 11. The two factors indicated in Table 11 as Factor 1 and Factor 2 are those identified by the analysis. The numbers in Table 11 are the loading values of each variable on each of the two identified factors. The presence of a higher loading value of a variable on a factor f means that the variable loads more strongly on f and loads more weakly on the other factor. The values closer to 1 have the most impact on a factor. Therefore, REXP, IEXP, and IREXP, have the most impact on Factor 2, while NSE and NCS have the most impact on Factor 1.

The two new factors that are defined based on the results of the factor analysis are:

1. *Experience (EXP)*: the sum of REXP, IREXP, and IEXP. The resulting value is in the range of 0 to 9. This value is binned into:
   - Low: for values 0 to 3,
   - Medium: for values 4 to 6, and
   - High: for values 6 to 9.
2. *Education (EDU)*: the sum of NSE and NCS. The resulting value is in the range of 0 to 6. This value is binned into:
   - Low: values 0 to 3, and
   - High: values 4 to 6.

[3] Factors the statistical analyses.

[4] Also called “principal axis factoring” or “common factor analysis”.

Table 14 Results of the Levene Test for MIX (Filtered)

Dependent Variable	Levene Statistic	df1	df2	p
NRAW	1.143	3	32	.347
NR	4.789	3	34	**.007**
NF	.697	3	35	.560
NI	9.361	3	32	**.000**

Because factor analysis identified two factors, a three-way ANOVA is necessary to test the effect of these factors and the main variable of the study, MIX. The three-way ANOVA is given in Section 10.3. The two factors are studied individually in detail also by means of a one-way ANOVA in the rest of Section 10 and in Online Appendix B.

The factor analysis suggests that the independent variables REXP, IREXP, and IEXP *could be* replaced by EXP and that NCS and NSE *could be* replaced by EDU in the analysis, thus analyzing the effects of only five independent variables, MIX, CR, EXP, EDU, and NGRAD. However, it is prudent to analyze the effects of all ten original and constructed independent variables, just to be sure that there are no surprises. Indeed, it turned out that NSE has a significant effect that is not observable in the analysis of the effect of EDU.

9 Hypotheses

Based on the independent variables listed in Section 6.5 and the two factors identified in Section 8.4, the main hypothesis described in Section 4 is broken down into several pairs of subhypotheses, one for each original or constructed independent variable X. The second of each pair is a null hypothesis, labeled H_{X_0}, and the first is the corresponding non-null hypothesis, labeled H_{X_1}. These are shown in Table 12. In this table, the hypotheses about independent variables that could be replaced by a constructed independent variables are indented under the hypotheses for the constructed independent variable.

Of the full set of hypotheses, only H_{MIX_1}, H_{MIX_0}, H_{CR_1}, H_{CR_0}, H_{REXP_1}, H_{REXP_0}, H_{IEXP_1}, and H_{IEXP_0} were tested in E1, as hypotheses $H1_1$, $H1_0$, $H2_1$, $H2_0$, $H3_1$, $H3_0$, $H4_1$, and $H4_0$, respectively (Niknafs and Berry 2012).

Table 15 Results of the One-Way ANOVA of the Effect of MIX (Unfiltered)

Dependent Variable	Sum of Squares [a]	df [b]	Mean Squares	F [c]	p [d]	Partial η^2 [e]	Observed Power
NRAW	2.228	3	.743	.765	.521	.060	.197
NR	.397	3	.132	.130	.941	.011	.072
NF	4.548	3	1.516	1.714	.181	.125	.41
NI	1.943	3	.648	.777	.515	.061	.200

[a] Type III sum of squares

[b] Degrees of freedom

[c] Value of the ANOVA's F-test

[d] p-value of the F-test

[e] Measure of effect size

Table 16 Results of the One-Way ANOVA of the Effect of MIX (Filtered)

Dependent Variable	Sum of Squares	df	Mean Squares	F	p	Partial η^2	Observed Power
NRAW	3.099	4	.775	1.049	.398	.116	.292
NF	7.218	4	1.804	2.576	.054	.227	.664

Later, as the effects on dependent variables of interactions between independent variables are explored and found to be significant, other hypotheses are added and logically named in terms of the independent variables involved, e.g., $H_{\text{MIX}*\text{EXP}*\text{EDU}_1}$ is the hypothesis that the effectiveness of a team in requirement idea generation is affected by the interaction between the team's MIX, EXP, and EDU values, and $H_{\text{MIX}*\text{EXP}*\text{EDU}_0}$ is the corresponding null hypothesis.

10 Statistical Analysis

This section describes a set of ANOVA and Kruskal-Wallis tests (Warner 2012) conducted on each of the independent variables and the two factors identified in Section 8.4 to test the hypotheses given in Section 9. Recall that each factor is considered an independent variable.

To test these hypotheses thoroughly, there needs to be an attempt to do an ANOVA to assess the impact of a chosen set of independent variables, IV_1, …, and IV_n, of a team on the team's unfiltered and filtered versions of the four dependent variables, for a total of eight dependent variables: RAW, AVG_R, AVG_F, AVG_I, NRAW, NR, NF, and NI. In most cases, the chosen set of independent variables is a singleton set, containing only one independent variable, for a one-way ANOVA. However, there is a three-way ANOVA with a set of three independent variables. So, this formulation is in terms of a chosen set of independent variables.

To be able to safely do this ANOVA, it is necessary to do a Levene test on each of the unfiltered and filtered versions of the four dependent variables of a team plotted against the team's chosen set of independent variables in order to ensure that the variances of the values of the dependent variable in the plots are homogeneous. When the result of the Levene test for any particular dependent variable DV, plotted against the chosen set of independent variables, is greater than 0.05, then an ANOVA assessing the impact of the chosen set of independent variables on DV is reliable.

- For the subset of a team's dependent variables for which an ANOVA is determined to be reliable, the ANOVA itself is done to assess the effect of the chosen set of independent variables of a team on the dependent variables in the subset. Then, for each of a team's dependent variables that the ANOVA test finds to be significantly affected by the chosen set of independent variables, a Tukey HSD Pairwise Comparison Test (Wikipedia 2013) is conducted to compare all possible pairs of means of the dependent variable to show which means are significantly different from each other.

Table 17 Results of the Kruskal-Wallis Test of the Effect of MIX (Unfiltered)

Dependent Variable	p
AVG_I	.555

Table 18 Results of the Kruskal-Wallis Test of the Effect of MIX (Filtered)

Dependent Variable	p
AVG_R	.697
AVG_I	.264

- For each of a team's dependent variables for which an ANOVA is determined not to be reliable, and for AVG_I, the dependent variable that was not normalized, a Kruskal-Wallis test is done to assess the effect of the chosen set of independent variables of the team on the dependent variable. Then, for each of a team's dependent variables for which the Kruskal-Wallis test is found to be significantly affected by the chosen set of independent variables, a Dunn-Bonferroni Pairwise Comparison Test (IBM Corp. 2013) is conducted to compare all possible pairs of medians of the dependent variable to show which medians are significantly different from each other.

These analyses are very lengthy and repetitive. To save space, this paper gives in Sections 10.1, 10.2, and 10.3, only a sampling of the detailed analyses, for (1) the one-way ANOVAs for only the MIX and NSE independent variables and (2) the one three-way ANOVA for the MIX, EXP, and EDU combination of independent variables. The MIX independent variable was chosen again for full treatment in this paper, because it is the subject of the main hypotheses derived from the RQs. The NSE independent variable was chosen for full treatment because its analysis required every kind of test mentioned above. Finally, the three-way ANOVA was chosen for full treatment, because it is totally different from the one-way ANOVAs. The rest of the detailed analyses are given in the subsections of Online Appendix B. Tables summarizing all of the analyses are provided in Section 10.4 of this paper.

Based on this plan, each of Sections 10.1, 10.2, and 10.3 and each subsection of Online Appendix B gives the following in short order with no further explanation:

1. Levene tests in the form of two tables, one for the unfiltered dependent variables and one for the filtered dependent variables: Each row of each table shows the results of the test for one dependent variable. When a row's p-value is greater than 0.05, the variances of the row's dependent variable are shown to be equal.
2. ANOVA tests in the form of two tables, one for the unfiltered dependent variables and one for the filtered dependent variables: Each row of each table shows the results of the test for one dependent variable. When a row's p-value is less than 0.05, the chosen set of independent variables is shown to have a significant effect on the row's dependent variable.
3. Tukey HSD Pairwise Comparison Tests in the form of a table for each significantly affected dependent variable: Each row of the table shows the results of the test for one

Table 19 Results of the Levene Test for NSE (Unfiltered)

Dependent Variable	Levene Statistic	df1	df2	p
NRAW	.141	3	36	.935
NR	1.354	3	36	.272
NF	1.106	3	36	.359
NI	.771	3	36	.518

Table 20 Results of the Levene Test for NSE (Filtered)

Dependent Variable	Levene Statistic	df1	df2	p
NRAW	.944	3	32	.431
NR	3.446	3	34	**.027**
NF	2.102	3	35	.118
NI	1.287	3	32	.296

pair of values of the affected dependent variable. When a row's p-value is less than 0.05, the difference between the pair of values in the row is shown to be significant.

4. Kruskal-Wallis tests in the form of two tables, one for the unfiltered dependent variables and one for the filtered dependent variables: Each row of each table shows the test results for one dependent variable. When a row's p-value is less than 0.05, the chosen set of independent variables is shown to have a significant effect on the row's dependent variable.
5. Dunn-Bonferroni Pairwise Comparison Tests are given in the form of a table for each significantly affected dependent variable: Each row of the table shows the test results for one pair of values of the affected dependent variable. When a row's p-value is less than 0.05, the difference between the pair of values in the row is shown to be significant.

Then, the subsection draws its conclusions relative to the hypotheses being tested.

10.1 One-Way ANOVA on MIX

Table 13 shows that the Levene test result of the unfiltered dependent variables plotted against MIX is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these unfiltered variables. Table 14 shows that the Levene test result of the filtered dependent variables plotted against MIX is not significant for each of NRAW and NF, but is significant for each of NR and NI. Thus, an ANOVA is applicable to the filtered NRAW and NF, but is not applicable to the filtered NR and NI.

Table 15 shows the results of the ANOVA test of the unfiltered dependent variables plotted against MIX. The test shows no significant effect of the team's MIX on any of these variables. Table 16 shows the results of the ANOVA test of the filtered NRAW and NF plotted against MIX. The test shows no significant effect of the team's MIX on any of these variables.

Table 17 shows the results of the Kruskal-Wallis test of the effect of a team's MIX on the unfiltered AVG_I generated by the team. The test shows no significant effect of the team's MIX on this variable. Table 18 shows the results of the Kruskal-Wallis test of the effect of

Table 21 Results of the One-Way ANOVA of the Effect of NSE (Unfiltered)

Dependent Variable	Sum of Squares	df	Mean Square	F	p	η^2	Observed Power
NRAW	4.629	3	1.543	1.706	.183	.124	.408
NR	1.733	3	.578	.591	.625	.047	.160
NF	10.624	3	3.541	4.949	**.006**	.292	.879
NI	4.832	3	1.611	2.138	.112	.151	.500

Table 22 Results of the One-Way ANOVA of the Effect of NSE (Filtered)

Dependent Variable	Sum of Squares	df	Mean Square	F	p	Partial η^2	Observed Power
NRAW	5.947	4	1.487	2.288	.081	.222	.599
NF	13.499	4	3.375	6.477	**.001**	.425	.981
NI	8.637	4	2.159	4.829	**.004**	.376	.923

a team's MIX on the filtered AVG_R and AVG_I generated by the team. The test shows no significant effect of the team's MIX on any of these variables.

10.2 One-Way ANOVA on NSE

Table 19 shows that the Levene test result of the unfiltered dependent variables plotted against NSE is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these unfiltered variables. Table 20 shows that the Levene test result of the filtered dependent variables plotted against NSE is not significant for each of NRAW, NF, and NI, but is significant for NR. Thus, an ANOVA is applicable to the filtered NRAW, NF, and NI, but is not applicable to the filtered NR.

Table 21 shows the results of the ANOVA test of the unfiltered dependent variables plotted against NSE. The test shows no significant effect of the team's NSE on each of NRAW, NR, and NI, but shows a significant effect of the team's NSE on NF. Table 22 shows the results of the ANOVA test of the filtered dependent variables plotted against NSE. The test shows no significant effect of the team's NSE on NRAW, but shows a significant effect of the team's NSE on NF and NI.

Table 23 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team's NSE on the unfiltered NF generated by the team. The test shows that the difference between the means of the NF of the teams is significant when NSE = 0 is paired with NSE = 2 and when NSE = 0 is paired with NSE = 3.

Table 24 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team's NSE on the filtered NF generated by the team. The test shows that the difference between the means of the NF of the teams is significant when NSE = 0 is paired with NSE = 2, when NSE = 0 is paired with NSE = 3, and when NSE = 1 is paired with NSE = 3.

Table 25 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team's NSE on the filtered NI generated by the team. The test shows that the difference

Table 23 Results of the Tukey HSD Pairwise Comparison Test of the Effect of NSE on NF (Unfiltered)

Sample 1	Sample 2	Mean Difference	Std. Error	p
0	1	−.026	.409	1.000
	2	−1.039	.370	**.039**
	3	−1.040	.336	**.019**
1	2	−1.012	.457	.138
	3	−1.014	.429	.103
2	3	−.001	.393	1.000

Table 24 Results of the Tukey HSD Pairwise Comparison Test of the Effect of NSE on NF (Filtered)

Sample 1	Sample 2	Mean	Std. Error	p
0	1	−.215	.352	.928
	2	−1.228	.320	**.003**
	3	−1.229	.291	**.001**
1	2	−1.012	.390	.063
	3	−1.014	.366	**.042**
2	3	−.001	.335	1.000

between the means of the NI of the teams is significant when NSE = 0 is paired with NSE = 2 and when NSE = 0 is paired with NSE = 3.

Table 26 shows the results of the Kruskal-Wallis test of the effect of a team's NSE on the unfiltered AVG_I generated by the team. The test shows no significant effect of the team's NSE on this variable. Table 27 shows the results of the Kruskal-Wallis test of the effect of a team's NSE on each of the filtered AVG_R and AVG_I generated by the team. The test shows no significant effect of the team's NSE on AVG_R, but shows a significant effect of the team's NSE on AVG_I.

Table 28 shows the results of the Dunn-Bonferroni Pairwise Comparison Test of the effect of a team's NSE on the filtered AVG_I generated by the team. The test shows that the difference between the medians of the AVG_I of the teams is significant when NSE = 0 is paired with NSE = 3.

10.3 Three-Way ANOVA on MIX, EXP, and EDU

Table 29 shows that the Levene test result of the unfiltered dependent variables plotted against MIX, EXP, and EDU is not significant for each of NRAW, NR, and NF, but is significant for NI. Thus, an ANOVA is applicable to the unfiltered NRAW, NR, and NF, but is not applicable to the unfiltered NI. Table 30 shows that the Levene test result of the filtered dependent variables plotted against MIX, EXP, and EDU is not significant for each of NRAW and NR, but is significant for each of NF and NI. Thus, an ANOVA is applicable to the filtered NRAW and NR, but is not applicable to the filtered NF and NI.

The Kruskal-Wallis test, which is used whenever the dependent variables do not meet the conditions for using an ANOVA, is a substitute for only a one-way ANOVA. We could not find any robust non-parametric equivalent of the multiple-way ANOVA to apply on a non-singleton set of dependent variables that do not satisfy the conditions for use of ANOVA. Therefore, a three-way ANOVA is applied anyway to the set MIX, EXP, and EDU.

Table 25 Results of the Tukey HSD Pairwise Comparison Test of the Effect of NSE on NI (Filtered)

Sample 1	Sample 2	Mean	Std. Error	p
0	1	−.0489	.352	.999
	2	−.871	.313	**.043**
	3	−1.006	.274	**.005**
1	2	−.823	.392	.175
	3	−.957	.361	.057
2	3	−.134	.323	.975

Table 26 Results of the Kruskal-Wallis Test of the Effect of NSE (Unfiltered)

Dependent Variable	p
AVG_I	.069

Table 31 shows the results of the three-way ANOVA test of the unfiltered dependent variables plotted against MIX, EXP, and EDU. This ANOVA reveals that:

1. MIX, alone, does not significantly affect any type of ideas;
2. EXP, alone, significantly affects only NI. However, the ANOVA results on NI are not reliable, since NI did not pass the Levene test;
3. EDU, alone, significantly affects all types of ideas;
4. the interaction of MIX, EXP, and EDU does significantly affect NRAW and NR; and
5. the rest of the interactions do not significantly affect any type of ideas.

Therefore, this ANOVA reveals that the interaction between MIX, EXP, and EDU on the unfiltered NRAW and NR is significant.

Table 32 shows the results of the three-way ANOVA test of the filtered dependent variables plotted against MIX, EXP, and EDU. This ANOVA reveals that:

1. MIX, alone, does not significantly affect any type of ideas;
2. EXP, alone, significantly affects only NI;
3. EDU, alone, significantly affects NF and NI;
4. the interaction of EXP and EDU does significantly affect NRAW;
5. the number of data points is not enough to calculate three-way interactions, e.g., the group with MIX=1, EDU=2, and EXP=1 has only one instance, i.e., the group's standard deviation is zero and degrees of freedom become zero; and
6. the rest of the interactions do not significantly affect any type of ideas.

Therefore, this ANOVA reveals that the interaction between EXP and EDU on the filtered NRAW is significant.

*10.3.1 MIX * EXP * EDU (Unfiltered)*

Figure 7 shows the interactions between three independent variables of MIX, EXP, and EDU on the unfiltered RAW and AVG_R. It is not possible to show interactions of three independent variables in a single plot. Thus, one of the independent variables, EDU, is fixed and the plots are provided for each value of EDU.

An issue with the sub-plots of Fig. 7 is that there are not enough data points to show the interactions between all values of the affecting independent variables. Also, comparing Fig. 7a with Fig. 7b and Fig. 7c with Fig. 7d, the correlations seem to be contradictory for EXP = "Low" and EXP = "High". All in all, the plots do not show anything interesting.

Table 27 Results of the Kruskal-Wallis Test of the Effect of NSE (Filtered)

Dependent Variable	p
AVG_R	.538
AVG_I	**.005**

Table 28 Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of NSE on AVG_I (Filtered)

Sample 1	Sample 2	Test Statistic	Std. Error	Std. Test Statistic	p
3	2	−.370	4.960	−.075	1.000
	1	−11.727	5.534	−2.119	.204
	0	−12.535	4.203	−2.982	**.017**
2	1	−11.357	6.007	−1.891	.352
	0	−12.165	4.810	−2.529	.069
1	0	−.808	5.399	−.150	1.000

One possible explanation for the interactions shown in Fig. 7 is that the less educated in CS a team is, the more a higher level of overall experience helps in generating raw requirement ideas. Conversely the more educated in CS a team is, the less a higher level of overall experience helps in generating raw requirement ideas.

*10.3.2 EXP * EDU (Filtered)*

Figure 8 shows the interactions between two independent variables of EXP and EDU on the filtered RAW. The plot shows that the medians of the filtered RAW generated by teams with EDU = "Low" is positively correlated with the teams' EXP. On the other hand, the medians of the filtered RAW generated by teams' with EDU = "High" is negatively correlated with the teams' EXP.

10.4 Summary of Statistical Analyses

Tables 33 and 34 summarize the statistical results. Table 33 summarizes the one-way ANOVAs and Table 34 summarizes the three-way ANOVA. Table 34 should actually be a section of Table 33, but the tables require different headers; so it is easier to make Table 34 a separate table, while marking the place in Table 33 in which Table 34's data would appear.

Table 33 is divided into three parts, each of which is what fits on one physical page. The legend explaining how to read the column headers is found at the bottom of Part III. A section of this table, which is about one independent variable, is the ten rows including and below each row with a value in the left-most column, namely the independent variable that the section is about. A subsection of this table is either the first five rows of a section or the last five rows of a section. The first subsection of any section is about unfiltered dependent variables and the second subsection of any section is about filtered dependent

Table 29 Results of the Levene Test for MIX, EXP, and EDU (Unfiltered)

Dependent Variable	Levene Statistic	df1	df2	p
NRAW	1.245	14	25	.306
NR	1.408	14	25	.220
NF	1.448	14	25	.203
NI	2.880	14	25	**.010**

Table 30 Results of the Levene Test for MIX, EXP, and EDU (Filtered)

Dependent Variable	Levene Statistic	df1	df2	p
NRAW	1.283	12	23	.292
NR	1.620	13	24	.148
NF	2.249	14	24	**.039**
NI	2.722	13	22	**.019**

variables. Notice that after the first column, the header is split into two rows. The upper row is the header that applies to the first row of any subsection. The lower row is the header that applies to the remaining four rows of any subsection.

Table 31 Results of the Three-Way ANOVA of the Effect of MIX, EXP, and EDU (Unfiltered)

Source	Dependent Variable	Sum of Square	df	Mean Square	F	p	Partial η^2	Observed Power
MIX	NRAW	.445	3	.148	.201	.894	.024	.082
	NR	1.879	3	.626	.665	.582	.074	.169
	NF	.474	3	.158	.213	.887	.025	.084
	NI	2.147	3	.716	1.168	.342	.123	.275
EXP	NRAW	.072	2	.036	.049	.953	.004	.057
	NR	.288	2	.144	.153	.859	.012	.071
	NF	.540	2	.270	.363	.669	.028	.102
	NI	4.496	2	2.248	3.670	**.040**	.227	.621
EDU	NRAW	6.170	1	6.170	8.384	**.008**	.251	.795
	NR	4.069	1	4.069	4.317	**.048**	.147	.515
	NF	6.832	1	6.832	9.192	**.006**	.269	.830
	NI	4.392	1	4.392	7.169	**.013**	.223	.730
MIX * EXP [a]	NRAW	1.545	4	.386	.525	.718	.077	.154
	NR	3.677	4	.919	.975	.439	.135	.263
	NF	1.152	4	.288	.387	.816	.058	.124
	NI	.817	4	.204	.334	.853	.051	.113
MIX * EDU	NRAW	1.097	1	1.097	1.491	.233	.056	.217
	NR	.080	1	.080	.085	.773	.003	.059
	NF	.977	1	.977	1.315	.262	.050	.197
	NI	.215	1	.215	.351	.559	.014	.088
EXP * EDU	NRAW	.025	1	.025	.034	.855	.001	.054
	NR	.160	1	.160	.170	.684	.007	.068
	NF	.068	1	.068	.092	.764	.004	.060
	NI	.250	1	.250	.407	.529	.016	.094
MIX * EXP * EDU	NRAW	3.733	1	3.733	5.073	**.033**	.169	.581
	NR	4.662	1	4.662	4.946	**.035**	.165	.571
	NF	1.639	1	1.639	2.205	.150	.081	.298
	NI	1.218	1	1.218	1.988	.171	.074	.273

[a] X * Y denotes the interaction of X and Y

Table 32 Results of the Three-Way ANOVA of the Effect of MIX, EXP, and EDU (Filtered)

Source	Dependent Variable	Sum of Squares	df	Mean Square	F	p	Partial η^2	Observed Power
MIX	NRAW	2.179	3	.726	1.279	.305	.143	.296
	NR	2.453	3	.818	1.090	.372	.120	.257
	NF	.793	3	.264	.508	.680	.060	.138
	NI	.486	3	.162	.494	.690	.063	.134
EXP	NRAW	.318	2	.159	.280	.759	.024	.089
	NR	1.697	2	.848	1.131	.339	.086	.225
	NF	.342	2	.171	.328	.723	.027	.096
	NI	4.704	2	2.352	7.168	**.004**	.395	.895
EDU	NRAW	1.214	1	1.214	2.139	.157	.085	.289
	NR	.316	1	.316	.421	.522	.017	.096
	NF	6.832	1	6.832	13.131	**.001**	.354	.935
	NI	2.507	1	2.507	7.641	**.011**	.258	.752
MIX * EXP	NRAW	4.467	3	1.489	2.622	.075	.255	.565
	NR	5.204	4	1.301	1.735	.175	.224	.450
	NF	1.118	4	.280	.537	.710	.082	.156
	NI	1.813	4	.453	1.382	.273	.201	.357
MIX * EDU	NRAW	.385	1	.385	.679	.418	.029	.124
	NR	1.733	1	1.733	2.310	.142	.088	.309
	NF	.977	1	.977	1.878	.183	.073	.260
	NI	8.087E-006	1	8.087E-006	.000	.996	.000	.050
EXP * EDU	NRAW	2.732	1	2.732	4.811	**.011**	.173	.556
	NR	1.933	1	1.933	2.578	.121	.097	.338
	NF	.068	1	.068	.132	.720	.005	.064
	NI	.152	1	.152	.464	.503	.021	.100
MIX * EXP * EDU	NRAW	.000 [a]	0	.	.	.	.000	.
	NR	.000	0	.	.	.	.000	.
	NF	1.639	1	1.639	3.151	.089	.116	.399
	NI	.000	0	.	.	.	.000	.

[a] When the number of data points needed to calculate the effect of a variable or interactions of some variables is not enough, SPSS outputs a value of 0 for sum of squares and degrees of freedom and "." for the other fields.

The ten rows of a section is about the independent variable, IV, that is displayed in the section's first row in the column headed by "IV". For the independent variable, IV, of a section:

– The five rows of either subsection of the section for IV is about the relationship between IV and four dependent variables, which are of *filtration* unfiltered or filtered as indicated by the value, "U" or "F", respectively, in the subsection's second row, in the column headed by "Filt'd?" in the lower header row. For the independent variable, IV, and the dependent variables of the subsection's filtration:

Table 33 Summary of Statistical Analysis: One-Way ANOVA

IV	Levene_T#			ANOVA_T#					K-W_T#		
	Filt'd?	DV	App'le?	SigEff?	Tukey_T#	WhenSig?	DV	Need?	SigEff?	DB_T#	WhenSig?
MIX	13			15					17		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
	14			16					18		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	No	–	–	–	AVG_R	Yes	No	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	No	–	–	–	AVG_I	Yes	No	–	–
CR	36			38					40		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
	37			39					41		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	No	–	–	–	AVG_R	Yes	No	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
REXP	42			44					47		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	Yes	46	Med:High	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
	43			45					48		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
IREXP	49			51					53		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
	50			52					54		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–

Table 33 (continued)

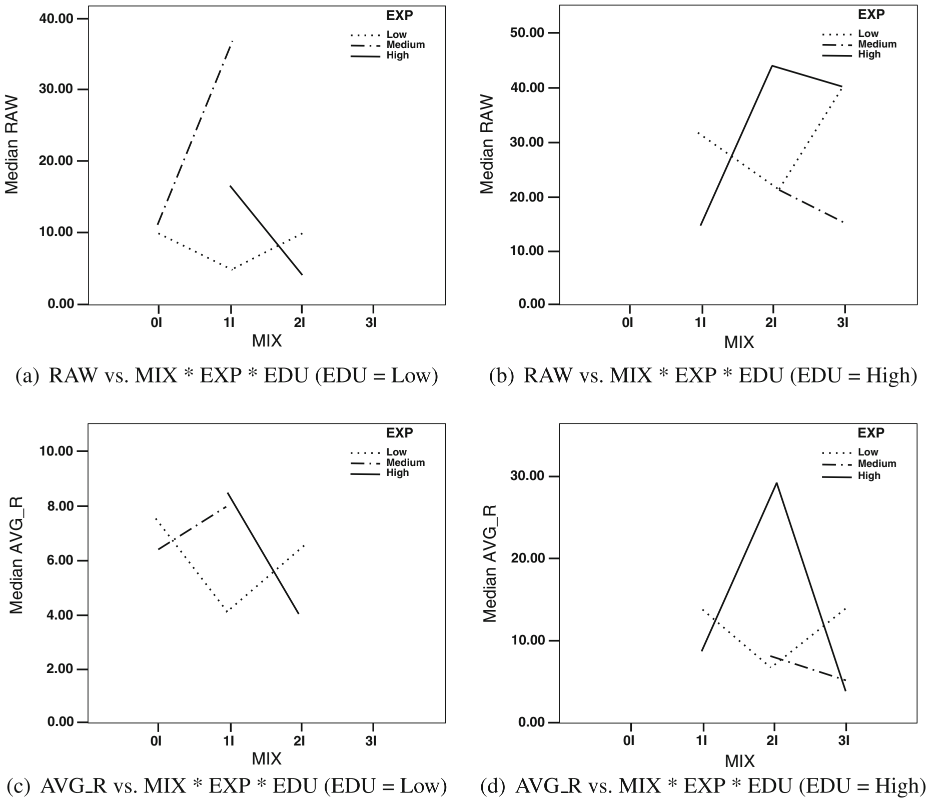
IV	Levene_T#			ANOVA_T#					K-W_T#		
	Filt'd?	DV	App'le?	SigEff?	Tukey_T#	WhenSig?	DV	Need?	SigEff?	DB_T#	WhenSig?
IEXP	55			57					59		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
	56			58					60		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
NCS	61			63					67		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
	62			64					68		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	Yes	65	0:3	AVG_F	No	–	–	–
		NI	Yes	Yes	66	0:3&1:3	AVG_I	Yes	Yes	69	0:3
NSE	19			21					26		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	Yes	23	0:2&0:3	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
	20			22					27		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	No	–	–	–	AVG_R	Yes	No	–	–
		NF	Yes	Yes	24	0:2&0:3&1:3	AVG_F	No	–	–	–
		NI	Yes	Yes	25	0:2&0:3	AVG_I	Yes	Yes	28	0:3
NGRAD	70			72					77		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
	71			73					78		
	F	NRAW	Yes	Yes	74	0:3	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	Yes	75	0:3	AVG_F	No	–	–	–
		NI	Yes	Yes	76	0:3	AVG_I	Yes	Yes	79	0:3

Table 33 (continued)

IV	Levene_T#			ANOVA_T#					K-W_T#		
	Filt'd?	DV	App'le?	SigEff?	Tukey_T#	WhenSig?	DV	Need?	SigEff?	DB_T#	WhenSig?
The 3-way ANOVA data would come here, but they require different headings. See Table 34 for these data											
EDU	80			82					84		
	U	NRAW	Yes	Yes	.	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	Yes	.	–	AVG_F	No	–	–	–
		NI	Yes	Yes	.	–	AVG_I	Yes	Yes	*	–
	81			83					85		
	F	NRAW	Yes	Yes	.	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	Yes	.	–	AVG_F	No	–	–	–
		NI	Yes	Yes	.	–	AVG_I	Yes	Yes	*	–
EXP	86			88					92		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	Yes	90	Low:Med	AVG_I	Yes	Yes	94	Low:Med
	87			89					93		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	Yes	91	Low:Med& Med:High	AVG_I	Yes	Yes	95	Low:Med

Legend

section	A section of this table is the 10 rows including and below each row with a value in the left-most column
subsection	A subsection of this table is either the first five rows of a section or the last five rows of a section
IV	the independent variable that is considered in the current section
Levene_T#	The Levene test results for the next four rows are found in the table whose number is given
ANOVA_T#	The ANOVA results for the next four rows are found in the table whose number is given
K-W_T#	The Kruskal-Wallis test results for the next four rows are found in the table whose number is given
Filt'd?	Are the DVs in the current subsection filtered (denoted "F") or unfiltered (denoted "U")?
DV	the dependent variable that is considered in the current row and in the next four columns
App'le?	Is the ANOVA applicable to the DV in the current row?
SigEff?	According to the test of the column, does the IV in the current section have a significant effect on the DV in the current row?
Tukey_T#	The Tukey HSD Pairwise Comparison Test results for the next four rows are found in the table whose number is given
WhenSig?	For which pairs of IV values are the DV variable values significantly different from each other?
Need?	Is the original non-normalized DV of the current row not normally distributed so that Kruskal-Wallis test is needed?
DB_T#	The Dunn-Bonferroni Pairwise Comparison Test results for the next four rows are found in the table whose number is given
*	Since EDU has only two values, the Tukey HSD Pairwise Comparison Test results would be the same as one-way ANOVA results; so no Tukey HSD Pairwise Comparison Test was done



(a) RAW vs. MIX * EXP * EDU (EDU = Low)

(b) RAW vs. MIX * EXP * EDU (EDU = High)

(c) AVG_R vs. MIX * EXP * EDU (EDU = Low)

(d) AVG_R vs. MIX * EXP * EDU (EDU = High)

Fig. 7 Ideas vs. MIX * EXP * EDU (Unfiltered)

- The first row of the subsection gives in the columns headed by "Levene_T#", "ANOVA_T#", and "K-W_T#" in the upper header row, the numbers of the tables, in Sections 10.1, 10.2, or 10.3 or in Online Appendix B, in which results can be found of the Levene test, the ANOVA, and the Kruskal-Wallis test for IV and the dependent variables of the given filtration.
- Each of the four other rows of the subsections is about the relationship between IV and the relevant filtration version of the dependent variables, DV_n and DV_u, displayed in one of the columns headed by "DV" in the lower header row. The value of a dependent variable is the number of one kind of requirement ideas generated, normalized or not.

  A segment of a row is the portion of the row lying between two adjacent vertical lines.

  The left segment of the row is about DV_n, the normalized version of some dependent variable, DV.

  The right segment of the row is about DV_u, the unnormalized version of the *same* dependent variable, DV.

Table 34 fits on one physical page. The legend explaining how to read the column headers is found at the bottom of the table. The only section of this table, which is about one triple of

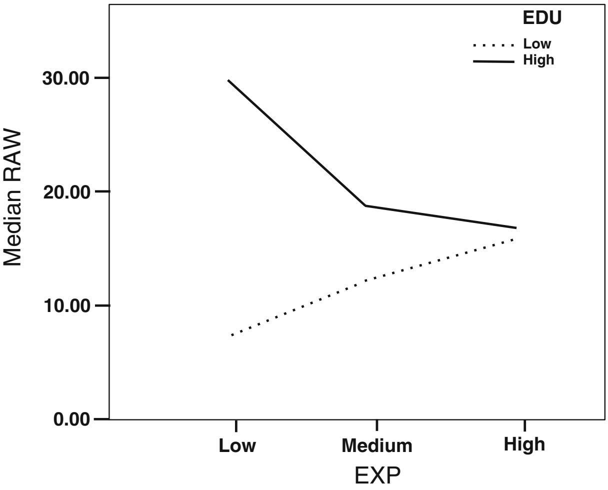


Fig. 8 RAW vs. EXP * EDU (Filtered)

independent variables, is the ten rows lying beneath the header. The section is about a triple of independent variables. A subsection of this table is either the first five rows of the section or the last five rows of the section. The first subsection of the section is about unfiltered dependent variables and the second subsection of any section is about filtered dependent variables. Notice that after the first column, the header is split into two rows. The upper row is the header that applies to the first row of any subsection. The lower row is the header that applies to the remaining four rows of any subsection.

The ten rows of the section is about the triple, MIX*EXP*EDU, of independent variables that is abbreviated as "MEE" in the section's first row, in the column headed by "IV". For the triple, MIX*EXP*EDU, of independent variables of the section:

- The five rows of either subsection of the section is about the relationship between MIX*EXP*EDU and four dependent variables, which are of *filtration* unfiltered or filtered as indicated by the value, "U" or "F", respectively, in the the subsection's second row, in the column headed by "Filt'd?" in the lower header row. For the triple of independent variables, MIX*EXP*EDU, and the dependent variables of the subsection's filtration:
  - The first row of the subsection gives in the columns headed by "Levene_T#" and "ANOVA_T#" in the upper header row, the numbers of the tables in which results can be found of the Levene test and the three-way ANOVA for MIX*EXP*EDU and the normalized dependent variables of the given filtration.
  - Each of the four other rows of the subsections is about the relationship between MIX*EXP*EDU and the relevant filtration version of the normalized dependent variable DV displayed in the column headed by "DV" in the lower header row.
  - Each row has only one segment in the sense of in Table 33 because of the various tests done, only the ANOVA, which needs normalized variables, works in the three-way mode.

Table 34 Summary of Statistical Analysis: Three-Way ANOVA

The postponed 3-way ANOVA data come here.

IV	Levene_T#			ANOVA_T#						
	Filt'd?	DV	App'le?	ALLsig?	MIX*EXPsig?	MIX*EDUsig?	EXP*EDUsig?	MIXsig?	EXPsig?	EDUsig?
MEE	29			31						
	U	NRAW	Yes	Yes	No	No	No	No	No	Yes
		NR	Yes	Yes	No	No	No	No	No	Yes
		NF	Yes	No	No	No	No	No	No	Yes
		NI	No*	No	No	No	No	No	Yes	Yes
	30			32						
	F	NRAW	Yes	.	No	No	Yes	No	No	No
		NR	Yes	.	No	No	No	No	No	No
		NF	No*	No	No	No	No	No	No	Yes
		NI	No*	.	No	No	No	No	Yes	Yes

Legend

section A section of this table is the 10 rows including and below each row with a value in the left-most column

subsection A subsection of this table is either the first five rows of a section or the last five rows of a section

IV the independent variable that is considered in the current section

Levene_T# The Levene test results for the next four rows are found in the table whose number is given

ANOVA_T# The ANOVA results for the next four rows are found in the table whose number is given

Filt'd? Are the DVs in the current subsection filtered (denoted "F") or unfiltered (denoted "U")?

DV the dependent variable that is considered in the current row

App'le? Is the ANOVA applicable to the DV in the current row?

ALLsig? According to the three-way ANOVA, do the three IVs, MIX, EXP, and EDU, together have a significant effect on the DV in the current row?

MIX*EXPsig? According to the three-way ANOVA, do two of the IVs, MIX and EXP, together have a significant effect on the DV in the current row?

MIX*EDUsig? According to the three-way ANOVA, do two of the IVs, MIX and EDU, together have a significant effect on the DV in the current row?

EXP*EDUsig? According to the three-way ANOVA, do two of the IVs, EXP and EDU, together have a significant effect on the DV in the current row?

MIXsig? According to the three-way ANOVA, does one of the IVs, MIX, alone have a significant effect on the DV in the current row?

EXPsig? According to the three-way ANOVA, does one of the IVs, EXP, alone have a significant effect on the DV in the current row?

EDUsig? According to the three-way ANOVA, does one of the IVs, EDU, alone have a significant effect on the DV in the current row?

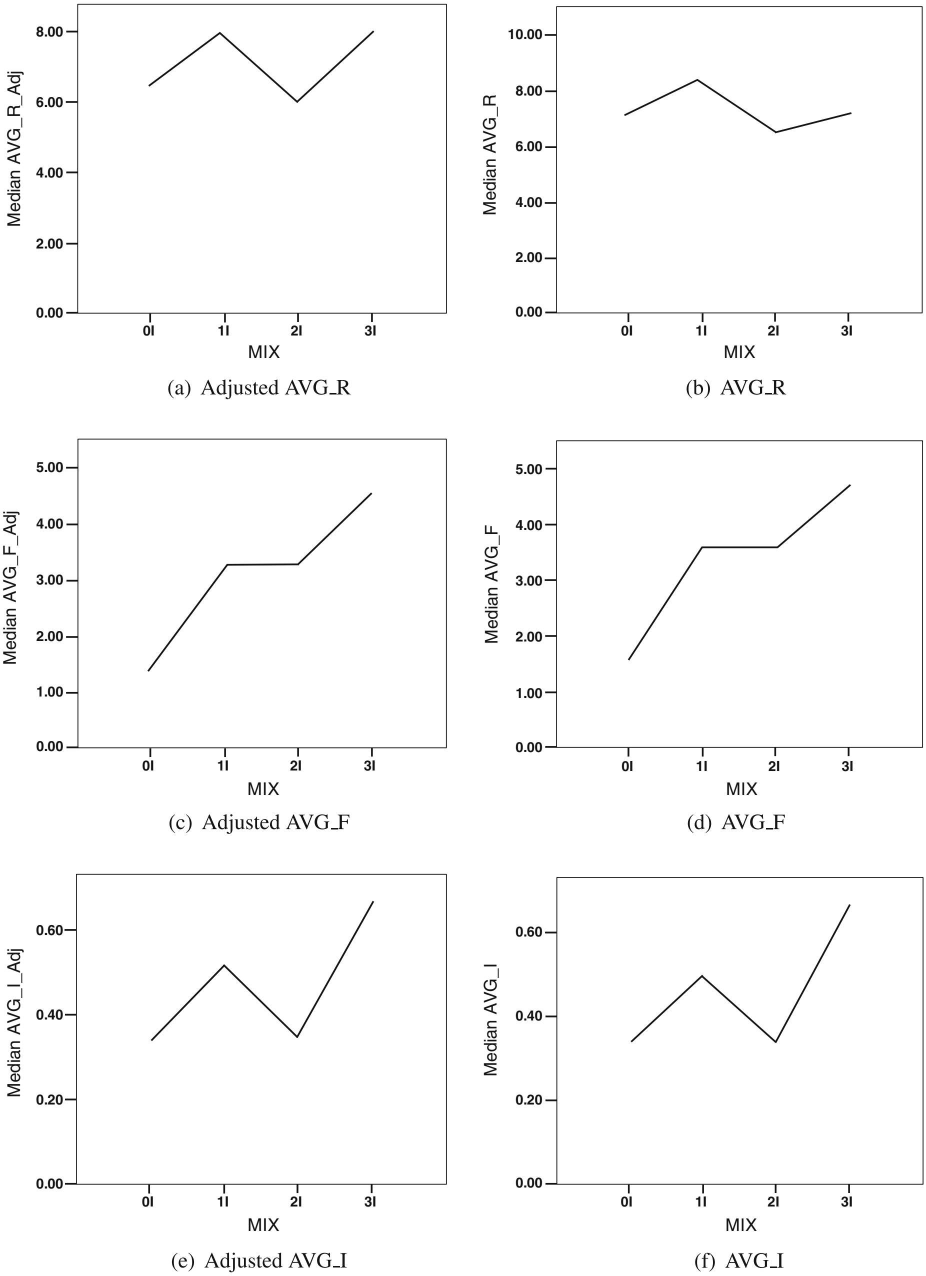
MEE MIX*EXP*EDU

"No*" Even though the ANOVA is not applicable for the row's DV, the ANOVA is done anyway, because there is no alternative test that works in three-way mode

"." (Period) There were not enough data points to calculate the effect of the IV of the current section on the dependent variable of the current row

11 Threats to Validity

This study is trying to provide practical results that are of high industrial relevance. Therefore, the more realistic the experiments are, the more useful the results are for practitioners.



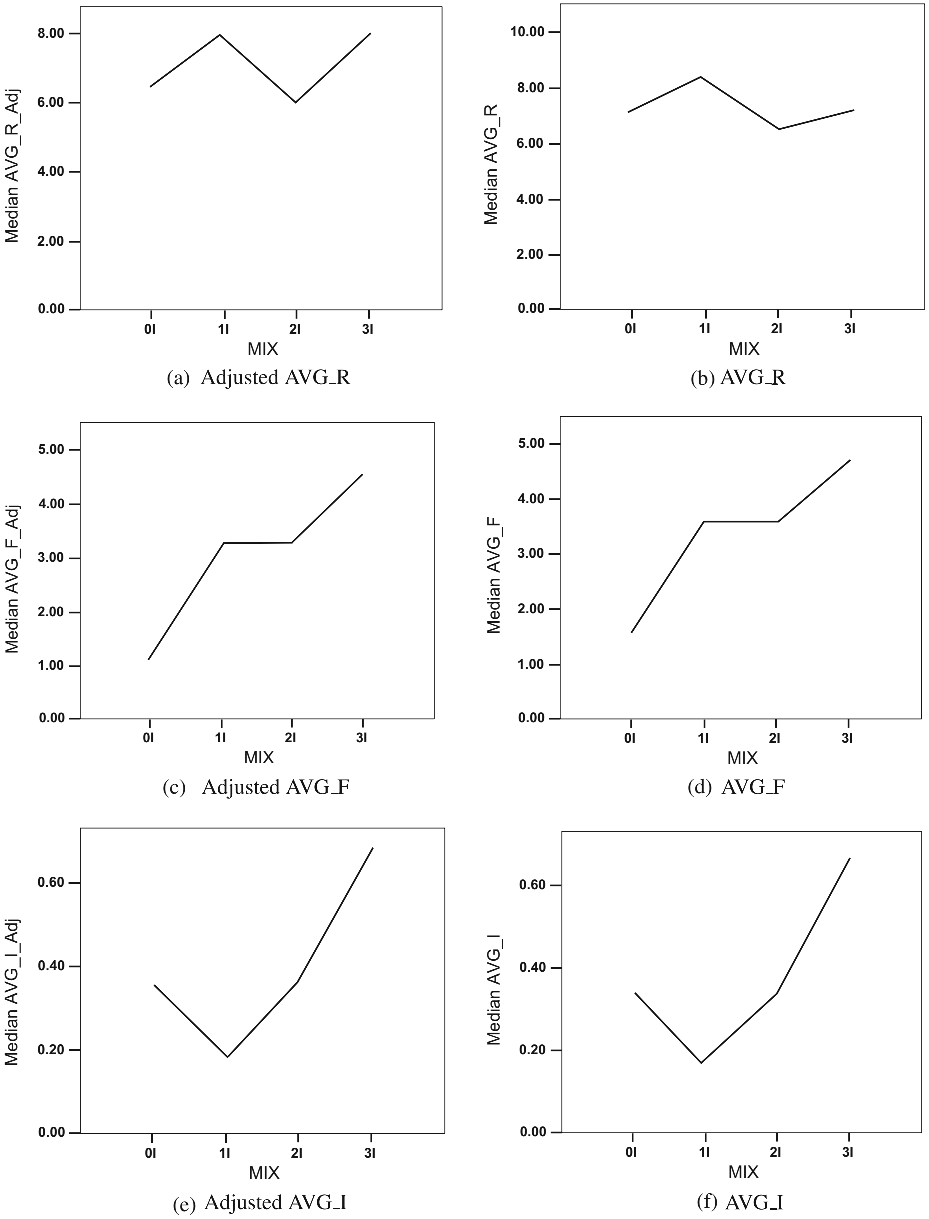
(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

Fig. 9 Adjusted Ideas vs. MIX – Ideas vs. MIX (Unfiltered)

Unfortunately, controlled experiments on real-world projects are not easy since many aspects of the project need to be controlled in order to conduct a well-designed experiment and obtain valid results. Real-world projects are usually constrained by real-world concerns that work against experimental validity.



(a) Adjusted AVG_R

(b) AVG_R

(c) Adjusted AVG_F

(d) AVG_F

(e) Adjusted AVG_I

(f) AVG_I

Fig. 10 Adjusted Ideas vs. MIX – Ideas vs. MIX (Filtered)

More feasible are controlled experiments with student participants and with realistically sized, but nevertheless contrived artifacts, such as the experiments described in this paper. Such a controlled experiment faces many threats to the validity of its results, which can be mitigated, if not eliminated, by careful design of the experiments.

There are four main types of validity of the experiments that are subject to threats: conclusion, internal, construct and external (Wohlin et al. 2000). The first author's PhD thesis (Niknafs 2014) identifies many threats in the experiments and explains the adopted mitigations. In most cases, the threat is quite typical and the adopted mitigation was standard. Due to space limitations, this paper addresses only the most salient of these threats.

11.1 Threats to Conclusion Validity

Conclusion validity addresses whether the conclusions about the hypotheses follow from the results of the experiment (Feldt and Magazinius 2010). The biggest conclusion validity threats for the experiments that we conducted concern

1. possible low statistical power, i.e., too few data points,
2. possible violations of the assumptions of the statistical tests used, and
3. the use of subjective measures for the quality of generated requirement ideas.

We used standard techniques to address these threats.

1. A post-hoc power analysis was performed to detect the minimum sample size required to achieve the standard minimum power value of 0.8. The analysis (Niknafs 2014) showed that the minimum needed sample size is 35. In this case, the sample size is the total number of teams, which is 40, well above 35.
2. Prior to performing ANOVAs, all the data were normalized. When necessary, other tests, which are more suitable for non-normal data, were run. In addition, outliers were identified, and tests were run both with and without the outlier data.
3. For the qualitative classifications of ideas, at least two and in some cases, three, classifiers were used. Moreover, statistical tests were used to show that there was high agreement among the classifiers.

11.2 Threats to Internal Validity

Internal validity addresses whether confounding factors within the experiment design are controlled so that the outcome of the experiment shows the causal relationship between the treatment and outcome. Typical internal validity threats include

1. possible learning effects and
2. possible instrument changes.

These threats were avoided by simply

1. using no participant more than once in the experiment and
2. conducting every run of the experiment according to the same plan and using the same requirement idea classification procedure each time.

Nevertheless, because there were two distinct collections of runs in two experiments E1 and E2, and each experiment had its own classification, there is a chance that the classifications of ideas in the two experiments might be different from what they would be if there had been only *one* classification of *all* the ideas at once. Section 7 shows that this chance became reality. Examination of the ratios between the numbers of relevant, feasible, and innovative ideas and the numbers of raw ideas in E1 and E2 showed significant differences between the E1 and E2 ratios for the relevant and feasible ideas. In order to determine if these differences affected the results, we tried adjusting the E2 data to equalize the ratios between the two experiments. Therefore, the number of ideas of each type of idea from E2, *T*, was multiplied by

$$\frac{\text{the ratio of the number of } T \text{ ideas to the number of } \textit{raw} \text{ ideas for E1}}{\text{the ratio of the number of } T \text{ ideas to the number of } \textit{raw} \text{ ideas for E2}}.$$

For example,

- the number of *relevant* ideas in E2 was multiplied by (27.5/58 = .474),
- the number of *feasible* ideas in E2 was multiplied by (20/26.5 = .755), and
- the number of *innovative* ideas in E2 was multiplied by (3.5/5 = 1.167).

It is possible to produce a plot of each unfiltered and filtered dependent variable against each independent variable, and it is possible to redo these plots with the adjusted data. To save space in this paper, the corresponding plots for only the MIX independent variable, the most important of the independent variables, are shown in Figs. 9 and 10. The corresponding plots for the rest of the unfiltered and filtered independent variables are shown in Online Appendix C. Specifically, these corresponding plots show that correlations between the medians of the adjusted data generated by teams and each of the teams' dependent variables either

- have no significant difference or
- have a slight difference in strength but are in the same direction as the corresponding plots of the unadjusted data.

Therefore, it is unlikely that a more detailed analysis would show any real difference.

What follows is evidence that the difference between the ratios of the ideas in E1 and E2 is due to the changes in the participants, not in the classifiers. Naturally, DAs are better in generating relevant and feasible ideas. The ratio of DAs to DIs in E1 is 0.32 and in E2 is 0.68. Since E2 had significantly more DAs, it is anticipated that the data of E2 had more relevant and feasible ideas. Besides, experience with classifying E1 data showed that classifying innovative ideas is more subjective than classifying relevant and feasible ideas. However, the ratios shown in Table 2 indicate that the changes on the less subjective data, i.e., for the relevant and feasible ideas, were large and the changes on the more subjective data, i.e., for the innovative ideas, were almost zero. The same conclusion follows an examination of the multipliers introduced in this section. Thus, the large differences in the ratios are in the more objective classifications for which the classifiers are not likely to change. Thus, it appears that the classifiers were very consistent between the two experiments, since they performed almost exactly the same on the more subjective data.

11.3 Threats to Construct Validity

Construct validity addresses whether the artifacts and procedures of the experimental plan ensure that the measures measure what they are intended to measure and that the results imply what they are intended to imply. The construct validity threats present in these experiments are

1. too few independent variables to discover true effects,
2. too few measures to discover true effects,
3. too few values of variables to discover true effects,

4. inaccurate or meaningless values to variables, and
5. bias towards confirming results in classifications.

Elements of the experimental procedure were designed specifically to address these threats.

1. We tested many more independent variables about properties of the participants than are needed to test the main hypothesis about the effect of the mix of teams' domain familiarities in case these properties proved to have more of an effect than the mix.
2. We used both easy-to-calculate, objective, quantitative data and difficult-to-evaluate, subjective, qualitative data about the requirement ideas generated.
3. When possible, we used more than just "present" or "absent" as the value of a variable, e.g., for REXP, the values were ranges of numbers of past RE projects.
4. We carefully chose as the application about which to generate requirement ideas, an application whose domain sharply divides the population by domain familiarity and for which it is easy to determine each participant's domain familiarity. The BDWP domain is quite rare in this respect, and finding it was a lucky strike.
5. Our method of having the experimenters evaluate generated requirement ideas, described in Section 6.3, ensures that no classifier knew from which team any idea came and thus that each classifier could focus on applying his or her expertise to evaluate all ideas accurately and uniformly.

11.4 Threats to External Validity

External validity addresses whether the results of the experiment with its highly controlled context generalize to the highly uncontrolled real-world context in which the RQs were asked. The three main possible threats to external validity are

1. the use of student as participants in the experiment rather than practicing requirements analysts who do requirement idea generation as part of their jobs,
2. the use of non-CS and non-SE, but nevertheless high-technology students as participants in the experiment rather than only CS or SE students who are learning to do the sorts of things that requirements analysts do, and
3. the use of the medium-sized application of a BDWP as the application about which to generate requirement ideas.

These threats require more attention than most of the threats to internal validity.

1. The goal of most empirical studies in SE is to draw conclusions valid for practitioners. However, convincing companies to allow their practitioner employees time off to participate in experiments is difficult. Therefore, these kinds of experiments are usually performed with students as participants. It is still not universally accepted that conclusions about software development professionals can be drawn from the results of a study done on software development students. However, Höst et al. (Höst et al. 2000) conducted some experiments using both students and professionals as participants and showed that the student participants did perform as well as the professional participants with no major difference, although they emphasize that their student participants possessed a good knowledge of SE. Note that the purpose of their experiments was to identify the factors affecting the lead time of software development projects.

For the experiments described in this paper, the plan was to use only CS and SE students as participants. The CS and SE education at the University of Waterloo (UW) includes courses that cover software requirements and specification. Moreover, almost all UW undergraduate CS and SE students are co-op students who get one term of industrial experience per year of study, and the co-op experience of many of these students includes software development. The CS and SE education at UW includes some courses for which a significant portion of the grade comes from a term-long group software development project. Finally, the purpose of the REXP, IREXP, and IEXP independent variables is to measure the extent to which the values of these variables say that these assumptions about the student participants are correct.

2. Recall that participants in E1 were all CS and SE students. In order to be able to get 10 teams of each mix over E1 and E2, for E2, we had to allow participants in high technology fields of study other than CS and SE. Doing so forced the introduction of new variables, namely NCS and NSE, to the study in order to be able assess whether this change affected the results. As shown in Sections 10.2 and 10.4 and as discussed in Section 13, the results *were* affected. So the threat materialized, but it was taken into account in the analysis.
3. While the BDWP is not a super-sized application requiring hundreds of developers, it is a real, medium-sized application, and there are several real products, e.g., TextEdit for Mac OS X (Wikipedia 2014), in the market supplying its functionality with varying degrees of success. With each such product, in the opinions of these authors, there *are* features that are missing or that could be changed. Thus, requirements elicitation for a BDWP is a real problem. Moreover, the one-half-hour duration of the requirement idea generation session, is realistic and matches what would be in an industrial one-hour brainstorming session that includes both an idea-generation step and an idea-pruning-and-refining step (Osborn 1953).

12 Conclusions and Discussion

The data of the aggregated results of the combined controlled experiments were analyzed to find any statistically significant results:

1. A factor analysis was conducted first to reveal the most influential variables. The found factors replaced five independent variables to give the final set of four independent variables.
2. Statistical analyses were performed on the eight original independent variables plus the two factors identified by the factor analysis.

Table 35 summarizes the statistical analysis results of Section 10. In this table, in any row, if the independent variable (IV) or set thereof of the row has a significant effect on any dependent variable (DV), then only those affected dependent variables are listed; the independent variable or set thereof has no significant effect on any other dependent variable, unfiltered or filtered.

The results of the statistical analysis on the full set of data for forty teams are taken into account to accept or reject the hypotheses:

H_{MIX}: The statistical analysis did not show any significant effect of MIX on any dependent variable. Therefore, H_{MIX_1}, the main hypothesis, is rejected, and H_{MIX_0} is accepted.

Table 35 Summary of the Statistical Analysis Results

IV or	DVs Significantly	Hypothesis H_{IV_1} Supported?			Corresponding	Overall
Set of IVs	Affected	Unfiltered	Filtered	Overall	E1 Hypothesis	Agrees w/E1?
MIX	None	No	No	No	$H1_1$	No
CR	None	No	No	No	$H2_1$	Yes
REXP	unfiltered NR	Very Weakly	No	No	$H3_1$	Yes
IREXP	None	No	No	No	None	N.A.
IEXP	None	No	No	No	$H4_1$	No
NCS	filtered NF, NI	No	Yes	Weakly	None	N.A.
NSE	unfiltered NF, filtered NF, NI	Very Weakly	Yes	Weakly	None	N.A.
NGRAD	filtered NRAW, NF, NI	No	Yes	Weakly	None	N.A.
EDU	unfiltered NRAW, NF, NI filtered NRAW, NF, NI	Yes	Yes	Yes	None	N.A.
EXP	unfiltered NI, filtered NI	Very Weakly	Very Weakly	Very Weakly	None	N.A.
MIX*EXP*EDU	unfiltered NRAW, NR	Yes	No	Weakly	None	N.A.

H_{CR}: The statistical analysis did not show any significant effect of CR on any dependent variable. Therefore, H_{CR_1} is rejected and H_{CR_0} is accepted.

H_{EDU}: A team's EDU incorporates two separate variables, NCS and NSE. While the statistical analysis showed that each of NCS and NSE has a significant effect on only a minority of the dependent variables, the analysis showed that EDU has a significant effect on a majority of the filtered and unfiltered dependent variables. Therefore, H_{EDU_1} is accepted and H_{EDU_0} is rejected.

H_{NGRAD}: The statistical analysis showed that NGRAD has a significant effect on a majority of the filtered dependent variables. Therefore, H_{NGRAD_1} is weakly accepted and H_{NGRAD_0} is weakly rejected.

H_{EXP}: A team's EXP incorporates three separate variables, REXP, IREXP, and IEXP. The statistical analysis did not show any significant effect of each of IEXP and IREXP on any dependent variable. The analysis showed that REXP has a significant effect on only one unfiltered dependent variable. In the end, the analysis showed that EXP has a significant effect on only one filtered and unfiltered dependent variable. Therefore, H_{EXP_1} is very weakly accepted and H_{EXP_0} is very weakly rejected.

$H_{\mathrm{MIX*EXP*EDU}}$: The statistical analysis showed that MIX in conjunction with EXP and EDU has a significant effect on half of the unfiltered dependent variables. Therefore, $H_{\mathrm{MIX*EXP*EDU}_1}$ is weakly accepted, and $H_{\mathrm{MIX*EXP*EDU}_0}$ is weakly rejected.

In addition, we had conducted a corroborating case study in a software-producing company of the idea generation part of a brainstorming for requirement ideas for a new software product (Niknafs and Berry 2013). The brainstorming team was staffed with four domain experts supplied by the company and with four domain ignorants supplied by us, all CS or SE people from UW. In e-interviews after the session, the interviewed company members of the team agreed that this session was more effective at generating innovative requirement ideas than their usual session. Moreover, Niknafs showed by analyzing traces from the origin of an ideas to its exhaustion, the out-of-box ideas originated with domain ignorants. Thus, the case study results support the conclusion that having a team consisting of a mix of domain experts and domain ignorants, all in CS or SE, improves the effectiveness of the idea generation part of requirement idea brainstorming.

13 Comparing Results of E1 and E1+E2

The last two columns of Table 35 shows how well the present results from E1+E2 agree with those reported for E1 in the conference paper by the same authors (Niknafs and Berry 2012). In E1, each of the participants was a CS or SE student. The results reported for E1 suggest that teams with a mix of domain familiarities are more effective at generating requirement ideas than are teams composed of only one domain familiarity. That is, H_{1_1} was weakly supported. However, E1 suffered from too few teams and unequal numbers of teams with different mixes of domain familiarities, and therefore, the statistical analysis results were weak.

E2 was conducted using the same plan used for E1, with the goal of having an equal number of teams of all mixes of domain familiarity, i.e., to have a balance among the mixes. To achieve this balance, it was necessary to include in E2 participants other than CS and SE students, who were nevertheless in some high technology fields. After combining the data of E1 and E2, there were an equal number of teams with the different mixes of domain familiarities, and therefore, the statistical analysis would be more reliable.

The statistical analysis of the combined data shows some differences with the statistical analysis of the E1 data. The statistical analysis performed on the combined data did not show any significant effect of mix of domain familiarities. However, that analysis revealed that there are other factors that are affecting the results. The main such factor was the educational background of the participants. Thus, while the statistical analysis of the E1 data showed some support for accepting the hypothesis H_{1_1} about the effect of MIX, the statistical analysis of the combined E1+E2 data did not provide *any* support for accepting the hypothesis H_{MIX_1}.

The natural question to ask is "Why do the two statistical analyses yield such different conclusions?" One possibility is that there was one of the two kinds of experimental error:

1. a Type I error occurred during E1, i.e., the null hypothesis is in fact true and there is really no effect of the mix of domain familiarities. In this case, the hypothesis would be wrong.
2. a Type II error occurred during the combined E1 and E2, i.e., the null hypothesis is really false, and the effectiveness of a team is really affected by the team's mix of domain familiarities. In this case, there would be factors besides the ones tested that are affecting the results and causing the Type II error. One possible such factor is personality traits, e.g. self-esteem. A DI might need to have high self-esteem to be effective. A DI should not be shy about showing his ignorance when it is useful, because he should know that doing so makes him more useful to a project. Also, he should know that he is competent in general and not ignorant about lots of other things. Thus, by revealing his ignorance about something, he should not be bothered. A person with low self-esteem, who conflates ignorance with stupidity or incompetence, may find it difficult to participate fully for fear of being thought stupid or incompetent. Since no data were collected about self-esteem, there is no way to determine if self-esteem, or lack thereof, affected the results. If another experiment is done in the future, these data can be gathered.

Another possibility is that there was no experimental error and the change in the educational backgrounds, from CS or SE to other high technology fields, of the participants affected the results. Certainly, the results of Section 10.3 say that the educational backgrounds of the members of a team affects the number of ideas generated by the team. The reality is that the main hypothesis carries an assumption that all analysts involved in idea generation *are competent in their CS-or-SE-related professions*. This assumption was so strong, that we never thought to introduce and measure any educational background variable in E1 and started doing so only in E2. It was fortunate that we knew the value of the variable for each subject of E1; he was in either CS or SE, which one being determinable from the course from which he was recruited. Thus, in having to to use participants from outside CS and SE, we have ended up demonstrating the importance of this assumption. Clearly, one possible item of future work would be to redo E2, using only CS and SE students to see if the results are more in line with those of E1.

14 Future Work

As for any controlled experiment, more data points will improve the strength of the results. Also, replication of the controlled experiment on different domains will improve the validity of its results. The more factors are controlled, the more precisely the effectiveness of domain ignorance might be studied. Because of the issues discussed in Section 13, replicating E2 with only CS or SE participants looks necessary.

There are several ways to extend this study. Testing the participants' level of domain familiarity is an important thing missing in this study. This study focused on the mere presence or absence of knowledge of a particular domain in participants. It might be a good idea to divide the participants into more categories.

1. Domain Expert (DE): those who are experts in the domain, e.g, from having implemented a CBS in the domain,
2. Domain Generalist (DG): those who have general knowledge of the domain arising, e.g., from regular use of a CBS in the domain,
3. Domain Novice (DN): those who have a limited knowledge of the domain arising, e.g., from only some use of a CBS in the domain or from use of a CBS in a similar domain, and
4. Domain Ignorant (DI): those who have no knowledge of the domain whatsoever.

Then, form teams of different combinations of DEs, DGs, DNs, and DIs and compare their effectiveness. The main issue with such a design is that it requires a large number of participants in order to be able to form a reasonable number of teams so as to achieve statistically valid results. It was fortunate that the domain used in E1 and E2 so sharply divided the population of participants. Basically, *every* E1 and E2 DI was thoroughly ignorant of the domain, *every* E1 and E2 DA was a regular user of a BDWP and was thus a DG, and there were no DEs and DNs. Among the classifiers, Niknafs and the third classifier were DGs, and Berry, having implemented some BDWPs (Habusha and Berry 1990; Berry 1999), was a DE.

Another way to extend the study is to run the experiment with domains other than that of BDWPs, to be sure that the results are not confined to only the BDWP domain. Doing so would allow using an industrial company's domain. Replication within industry is very valuable for improving the validity of the experiment. Surveys and examination of project histories are also other ways of finding evidence for the hypothesis, although with less significance than with controlled experiments.

Of course, use of domains other than that of BDWPs is potentially a double-edged sword. As we learned in the pilot studies, many other domains fail to sharply divide the participants by domain familiarity. An important element of control over the main independent variable would be lost.

Still another way to extend the study is to investigate the impact of participants' knowledge of domains different from the domain of the CBS under study. An idea that is common in one domain might be totally new to another domain. Thus, injecting knowledge of different domains fosters the creativity of the whole team. However, one of the issues with such a design is how to discover domains that participants are knowledgeable of. Also, it would require a large number of participants with the same domain knowledge to be able to form different combinations of teams and analyze the results.

Finally, while this work focused on RE tasks, the findings might be applicable to other RE and SE tasks. The experiment could be run using other tasks in RE and using software development tasks not in RE.

15 Implications for Practitioners, Researchers, and Educators

In this section, it is assumed that all employees involved are trained in CS or SE so that the weakly supported hypothesis $H_{\mathrm{MIX}*\mathrm{EXP}*\mathrm{EDU}_1}$ is relevant when considering an employee who is ignorant in the domain of the CBS being developed.

The most obvious use of these results is in the situation that was tested in the experiments, in brainstorming sessions for requirement ideas. When it is necessary to staff a group for brainstorming for requirement ideas, it cannot hurt to, and it will probably help to put at least one DI in the group. At the very least, doing so would help find a productive use for new employees who have not yet learned the employer's domain, and who would, at best, be useless and, at worst, be a drain on other more experienced employees (DeMarco and Lister 1987; Brooks 1995).

The results can be used by employers when considering initial tasks for a new employee, who is generally ignorant in the employer's domain on arrival at the new job. Specifically, assign the new employee to a knowledge-intensive task, such as brainstorming for requirement ideas, so that while the new employee slowly learns the domain without being a drain on other employees, the new employee parlays his or her domain ignorance in a way that is helpful to the other employees on the task.

Of course, following these staffing suggestions requires that management understand the value of domain ignorance and that it behave, at least in the near future, differently from the norm. Even if the management understands, there may be resistance from employees that are labeled "ignorant". On the other hand, once these employees see that their *temporary* ignorance is being used productively, and they are experiencing a more gradual and less stressful immigration into a new company, all the while being productive from day one, they should begin to appreciate the positive value of the temporary label and cooperate enthusiastically (Dagenais et al. 2010; Mehrotra 2011). We do know of at least one company, C, in North America that has learned the value of domain ignorance and is actively exploiting it. C's business is a service that is amenable to a high degree of automation. C has set up an Innovation and Creativity Center (ICC) whose role is come up with innovative and creative product ideas. C is sending at least some of its new hires to work in the ICC until they learned enough about C's domain to be useful elsewhere in C (Mehrotra 2011).

These results can be used by requirements-analysis consulting firms to turn their lack of knowledge about a client's domain into a competitive advantage over other firms with expertise in the client's domain. After all, why hire a consultant to provide domain expertise when there are plenty of domain experts in house?

These results may be applicable also with respect to CS and SE domain knowledge. In a typical instance of the context described in Section 3, the client-supplied DA who knows a lot about the domain of the client's CBS is probably ignorant in many aspects of CS or SE. Conversely, the requirement analyst DI who knows very little about the domain of the CBS may be overly fixed on past and current computing and software technology. Perhaps the DA can serve as a computing-and-software ignorant to help the requirements elicitation team be creative about technology.

If one is willing to generalize a bit from the task of requirement idea generation, there are other RE tasks that might benefit from domain ignorance. According to Mehrotra (2011), one such RE task is requirements specification inspection. Requirements specification inspection is basically brainstorming for signs of defects in the inspected requirements specification. Testing that this benefit actually occurs is of course a topic for additional research.

More generally, one of the expected benefits of domain ignorance is the ability of a DI to bring out any existing tacit assumptions. Thus, any discipline that needs tacit assumptions to be surfaced could potentially benefit from domain ignorance. The literature shows that a few of the disciplines that benefit from domain knowledge are cross-functional communication (Damian et al. 2013), data mining (Kopanas et al. 2002; Anand et al. 1995), and exploratory software testing (Itkonen et al. 2013). Another discipline that requires studying the effect of domain ignorance is knowledge management. The main goal of knowledge management is to codify the knowledge of an organization (Frappaolo 2008). While codifying explicit knowledge would be a straightforward task (e.g. by interviewing domain experts), codifying tacit knowledge is much harder. Tacit knowledge needs to identified, converted to explicit knowledge, and then codified. Thus, potentially, DIs could be very beneficial in an effort to extract tacit knowledge in a knowledge-management task. Here too, testing that these benefits actually occur is a topic for additional research.

SE education traditionally teaches that domain awareness and even domain expertise is needed to adequately conduct requirements analysis (Lauesen 2001; Berenbach et al. 2009). It is time that a more balanced view about the effects of domain knowledge be taught, as does a more recent RE textbook (Laplante 2014).

16 Contributions

The main contributions of this paper are:

- a domain, namely that of BDWPs, and a general experiment design for controlled experiments dealing with the effects of domain ignorance in RE and SE,
- support for concluding that *among teams consisting of mostly computer scientists and software engineers*, a team with a mix of domain familiarities is more effective, quantitatively and qualitatively, at requirement idea generation than a team with only one kind of domain familiarity, and
- ideas on how this conclusion can be applied by practitioners, researchers, and educators.

Acknowledgments The authors thank the anonymous reviewers of the earlier version of this paper for their comments. Ali Niknafs's and Daniel Berry's work were supported in parts by Canada's NSERC grant NSERC-RGPIN227055-00 and by Canada's NSERC–Scotia Bank Industrial Research Chair NSERC-IRCPJ365473-05.

References

Al-Rawas A, Easterbrook S (1996) Communication problems in requirements engineering: A field study. In: Proceedings of the First Westminster Conference on Professional Awareness in Software Engineering (PACE), pp 47–60

Anand SS, Bell DA, Hughes JG (1995) The role of domain knowledge in data mining. In: Proceedings of the Fourth International Conference on Information and Knowledge Management (CIKM), pp 37–43

Apfelbaum EP, Phillips KW, Richeson JA (2014) Rethinking the baseline in diversity research: Should we be explaining the effects of homogeneity? Perspect Psychol Sci 9(3):235–244

Basili VR, Caldiera G, Rombach DH (1994) The goal question metric approach. In: Marciniak JJ (ed) Encyclopedia of software engineering, vol I. Wiley

BBN Technologies (2015) Prophet statguide: Possible alternatives if your data violate one-way anova assumptions. [Online; accessed 18-May-2015]. http://www.basic.northwestern.edu/statguidefiles/oneway_anova_alts.html

Berenbach B, Paulish DJ, Kazmeier J, Rudorfer A (2009) Software & systems requirements engineering: in practice. McGraw-Hill, New York

Berry DM (1995) The importance of ignorance in requirements engineering. J Syst Softw 28(2):179–184

Berry DM (1999) Stretching letter and slanted-baseline formatting for arabic, hebrew, and persian with ditroff/ffortid and dynamic postscript fonts. Softw Pract Experience 29(15):1417–1457

Berry DM (2002) The importance of ignorance in requirements engineering: an earlier sighting and a revisitation. J Syst Softw 60(1):83–85

Blom G (1960) Statistical estimates and transformed beta-variables. Inc Stat 10(1):53–55

Brooks FP (1995) The mythical man-month: essays on software engineering, 20th anniversary edition. Addison-Wesley Professional, Boston

Carver JC, Nagappan N, Page A (2008) The impact of educational background on the effectiveness of requirements inspections: an empirical study. IEEE Trans Softw Eng 34(6):800–812

Dagenais B, Ossher H, Bellamy RKE, Robillard MP, de Vries JP (2010) Moving into a new software project landscape. In: Proceedings of the International Conference on Software Engineering (ICSE), vol 1, pp 275–284

Damian D, Helms R, Kwan I, Marczak S, Koelewijn B (2013) The role of domain knowledge and cross-functional communication in socio-technical coordination. In: Proceedings of the 2013 International Conference on Software Engineering (ICSE), pp 442–451

DeMarco T, Lister T (1987) Peopleware: productive projects and teams. Dorset House, New York

Dieste O, Juristo N, Shull F (2008) Understanding the customer: what do we know about requirements elicitation? IEEE Softw 25(2):11–13

Dunbar K (1999) How scientists build models invivo science as a window on the science mind. In: Magnani L, Nersessian N, Thagard P (eds) Model-based reasoning in scientific discovery. Kluwer Academic/Plenum Publishers, New York, pp 85–99

Feldt R, Magazinius A (2010) Validity threats in empirical software engineering research—an initial survey. In: Proceedings of the International Conference on Software Engineering and Knowledge Engineering, pp 374–379

Ferrari R, Madhavji NH (2007) The impact of requirements knowledge and experience on software architecting: An empirical study. In: Proceedings of the Sixth Working IEEE/IFIP Conference on Software Architecture (WICSA)

Finkelstein A (1994) Requirements engineering: a review and research agenda. In: Proceedings of the First Asia-Pacific Software Engineering Conference, pp 10–19

Firestein S (2013) Ignorance (Course). http://bioweb.biology.columbia.edu/firestein/?page_id=36

Fischer G (1999) Symmetry of igorance, social creativity, and meta-design. In: Proceedings of the 3rd Conference on Creativity & Cognition (C&C), pp 116–123

Frappaolo C (2008) Implicit knowledge. Knowl Manag Res Pract 6(1):23–25

Glass GV, Peckham PD, Sanders JR (1972) Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev Educ Res 42(3):237–288

Habusha U, Berry D (1990) vi.iv, a bi-directional version of the vi full-screen editor. Electron Publ — Origination Dissemination, Des 3(2):65–91

Hadar I, Soffer P, Kenzi K (2014) The role of domain knowledge in requirements elicitation via interviews: an exploratory study. J Requir Eng 19(2):143–149

Hanebutte N, Taylor CS, Dumke RR (2003) Techniques of successful application of factor analysis in software measurement. Empir Softw Eng 8(1):43–57

Hinton PR, McMurray I, Brownlow C (2004) SPSS Explained. Routledge, East Sussex

Höst M, Regnell B, Wohlin C (2000) Using students as subjects — a comparative study of students and professionals in lead-time impact assessment. Empir Softw Eng 5(3):201–214

IBM Corp. (2013) Post hoc comparisons for the Kruskal-Wallis test. http://www-01.ibm.com/support/docview.wss?uid=swg21477370. [Online; accessed 11-Sep-2013]

Itkonen J, Mantyla MV, Lassenius C (2013) The role of the tester's knowledge in exploratory software testing. IEEE Trans Softw Eng 39(5):707–724

Jarke M, Jr JAB, Rolland C, Sutcliffe AG, Vassiliou Y (1993) Theories underlying requirements engineering: an overview of NATURE at Genesis. In: Proceedings of the IEEE International Symposium on Requirements Engineering (RE), pp 19–31

Kopanas I, Avouris NM, Daskalaki S (2002) The role of domain knowledge in a large scale data mining project. In: Vlahavas CDS, Ioannis P (eds) Methods and Applications of Artificial Intelligence, Lecture Notes in Computer Science, vol 2308. Springer, Berlin, pp 288–299

Kristensson P, Gustafsson A, Archer T (2004) Harnessing the creative potential among users. J Prod Innov Manag 21(1):4–14

Laplante PA (2014) Requirements engineering for software and systems, 2nd Edn. Taylor & Francis Group, Boca Raton

Lauesen S (2001) Software requirements: styles & techniques. Pearson Education, Harlow

Lehrer J (2009) Accept defeat: The neuroscience of screwing up. http://www.wired.com/2009/12/fail_accept_defeat. [Online; accessed 6-May-2014]

Luchins AS (1942) Mechanization in problem solving: the effect of einstellung. Psychol Monogr 54(6):i–95

Luchins AS, Luchins EH (1950) New experimental attempts at preventing mechanization in problem solving. J Gen Psychol 42:1335–1342

McAllister CA (2006) Requirements determination of information systems: User and developer perceptions of factors contributing to misunderstandings. Ph.D. thesis. Capella University, Minneapolis. http://search.proquest.com/docview/304908259

Mehrotra G (2011) Role of domain ignorance in software development. Master's thesis, University of Waterloo, Waterloo. http://se.uwaterloo.ca/~dberry/FTP_SITE/students.theses/gaurav.mehrotra/gauravMehrotraThesis.pdf

Naur P, Randell B (1969) Software engineering: report of a conference sponsored by the NATO science committee. Scientific Affairs Division, NATO, Brussels

Niknafs A (2014) The impact of domain knowledge on the effectiveness of requirements engineering activities. Ph.D. thesis. University of Waterloo, Waterloo. https://uwspace.uwaterloo.ca/handle/10012/8470

Niknafs A, Berry DM (2012) The impact of domain knowledge on the effectiveness of requirements idea generation during requirements elicitation. In: Proceedings of the 20th IEEE International Requirements Engineering Conference (RE), pp 181–190

Niknafs A, Berry DM (2013) An industrial case study of the impact of domain ignorance on the effectiveness of requirements idea generation during requirements elicitation. In: Proceedings of the 21st IEEE International Requirements Engineering Conference (RE), pp 279–283

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

Pascal B, Krailsheimer AJ (1968) Pensees: Translated with an Introduction by A.J. Krailsheimer. Penguin, London

Rose P, Kumar M, Ajmeri N, Agrawal M, Sivakumar V, Ghaisas S (2009) A method and framework for domain knowledge assisted requirements evolution (K-RE). In: Proceedings of CONSEG-09: International Conference on Software Engineering, pp 87–97

Sharp H (1991) The role of domain knowledge in software design. Behav Inform Technol 10(5):383–401

Taylor CW, Williams FE (1965) Instructional Media and Creativity: The Proceedings of the Sixth Utah Creativity Research Conference. Distributed by ERIC Clearinghouse, Washington. http://nla.gov.au/nla.cat-vn5184417

Thagard P (1997) Collaborative knowledge. Noûs 31(2):242–261

Warner RM (2012) Applied statistics: from bivariate through multivariate techniques: from bivariate through multivariate techniques. Sage Publications, Thousand Oaks

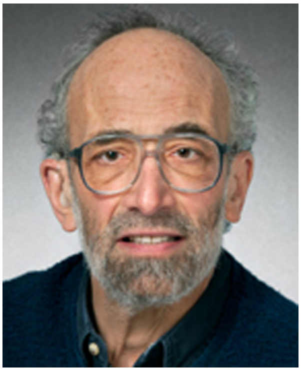
Wikipedia (2013) Tukey's range test — Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Tukey%27s_range_test. [Online; accessed 1-Sept-2013]

Wikipedia (2014) Textedit— Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/TextEdit. [Online; accessed 2-May-2014]

Wiley J (1998) Expertise as mental set: the effects of domain knowledge in creative problem solving. Mem Cogn 26(4):716–730

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

Ali Niknafs got his Ph.D. in Computer Science from the University of Waterloo, Canada in 2014, his M.S. in Computer Engineering from Sharif University of Technology, Iran in 2008, and his B.S. in Computer Engineering from Shomal University, Iran in 2006. His research interests lie in the area of software engineering, with a focus on software requirements engineering and the software development lifecycle.



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