Noname manuscript No.
(will be inserted by the editor)

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

Ali Niknafs · **Daniel Berry**

Received: date / Accepted: date

Abstract [Background] 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 might overlook obvious issues.
[Objective] 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?"
[Method] 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.
[Results] The results, although not conclusive, provide some support for the positive effect of the mix on effectiveness in idea generation. The results also showed a significant effect of other independent variables, especially educational background.
[Conclusion] The main conclusion is that the presence of a domain ignorant with a computer science or software engineering background improves the effectiveness of a requirement idea generation team.

Ali Niknafs
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
E-mail: niknafs@gmail.com

Daniel Berry
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
E-mail: dberry@uwaterloo.ca

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 [9]. 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 [19]. 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 [1, 30, 40].

However, deep knowledge of the problem domain seems to lead to falling into the tacit assumption tarpit [6]. Lack of domain knowledge might, in fact, have some benefits in RE activities. One such benefit has been observed by Berry [6], 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 [21]. This surfacing can lead to the necessary shared understanding of the topics of the tacit assumptions.

Section 2 of this paper describes related work. Sections 3 through 7 describe the experiment's design and method. Sections 8 through 13 describe and discuss the results of the experiment, including the threats. Sections 14 and 16 conclude the paper with a wrap up of the results, a comparison with results of previous work, and a description of future work.

2 Background and Related Work

Very few studies have investigated the impact of domain knowledge on software engineering (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 [6] 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 he was quite ignorant. As he noted later

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

[7], an even earlier observation of the impact of ignorance is from Burkinshaw's statement during the second NATO conference on SE in 1969 [35]:

> 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.

From a survey on requirements elicitation techniques, Dieste et al. [14] 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.

Kenzi et al. [25] 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.

Ferrari et al. [18] 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.

Carver et al. [10] 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.

In an experiment conducted on software design, Sharp [41] 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.

Mehrotra [34] 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. [12] 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.

Kristensson et al. [32] 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.

Stuart Firestein [20] teaches a course called *Ignorance* at the University of Columbia. 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 [39]:

> 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.

Dunbar [15] studied how scientists study things in practice. He found that over half of the data that scientists find are unexpected. What they do with the unexpected data? They find an excuse and ignore it altogether. Lehrer 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 [33]. 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 [15, 44]. This situation is called also *distributed reasoning* [15]. 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.

Apfelbaum et al. [3] 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 the 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 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 [4], 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?*
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 with 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 Desired Contributions

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 of DIs and DAs for requirement idea generation,
- providing at least one RE activity for which domain ignorance is at least helpful, and
- 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.

7 Experiment Design

This section explains the design of controlled experiments [52] that aimed to answer the RQs.

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 [37]. E1's results were that there was some support for accepting the main hypothesis. However, E1 suffered from (1) the small number of teams and (2) an imbalance in the numbers of teams with each mix of domain familiarity, the main independent variable, resulting in a reduction in the statistical strength of the results. E2 was conducted to provide more provide more teams and to balance the number of teams with each mix of domain familiarity.

7.1 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 are required to be taken into consideration. Thus, before conducting the actual experiment,

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, in the proverbial shower, 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.

7.2 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 were *3I*, consisting of only DIs. For E2, we decided to allow participants other than CS and SE students. We knew that this decision might introduce new variables to the study, but it was the only option left at the time: We had exhausted the pool of potential volunteer participants and would have to wait another year for a new batch of students to arrive. We did insist that each participant be in some high technology field of study.

7.3 Evaluation 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 [5], 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.
Feasibility: An idea is considered feasible if it is relevant and it is correct, well presented, and implementable.
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 evaluator.

We decided to use ourselves, both experts in the BDWP domain, as idea evaluators. To eliminate any bias in classifying an idea that might arise from an evaluator'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 evaluator would then classify the ideas in the full list. Once both evaluations are done, each evaluator'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 in each classification, as determined by the classifiers, is used as the value of the classification.

Later, we added a third evaluator, an Arabic and Hebrew speaker. We had discovered almost unanimous agreement over which ideas were relevant and feasible, but some disagreement over which feasible ideas were innovative. So, to save money while getting the most bang for each buck, we had the third evaluator evaluate for innovativeness only the union of our feasible ideas.

7.4 Procedure

As described in Figure 1, 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 [42] 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".

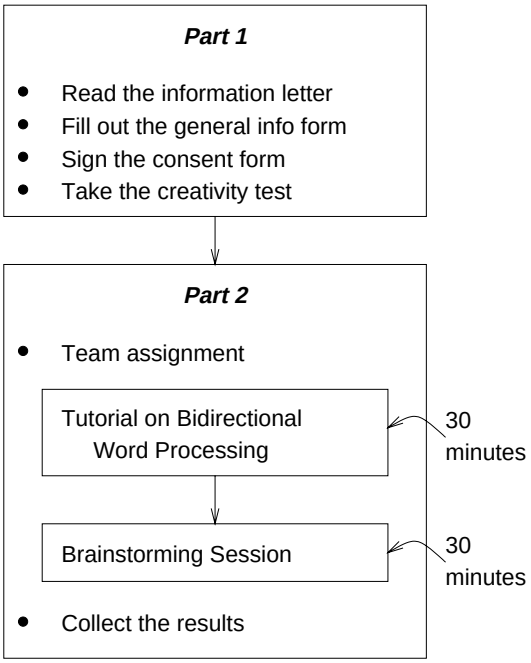


Fig. 1: Steps of the Controlled Experiment (Refined)

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/tech.reports/NiknafsBerryMaterials/`.

7.5 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.

7.5.1 Independent Variables About a Team

The independent variables about a *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*.

- *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.

7.5.2 Dependent Variables About a Team

The dependent variables about a team are based on the classifications of the requirement ideas described in Section 7.3:

- *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 evaluators 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 evaluators 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 evaluators 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.

7.6 Statistical Analyses

When using statistical methods to describe an observation, two kinds of errors can happen:

1. A Type I error occurs, with probability α, when a null hypothesis that should be accepted is rejected.

2. A Type II error occurs, with probability β, when a null hypothesis that should be rejected is accepted.

In order to test the hypothesis, we first need to define an acceptable probability for each of these two errors. The typical value for α is 0.05 and for β is 0.20 [16]. The value of $1-\beta$ for a statistical test is referred to as the power of the statistical test.

The differences between the teams are determined by means of an analysis of variance (ANOVA) [50]. In order to be allowed to apply an ANOVA, the data should be verified to meet the three prerequisite assumptions of the ANOVA test:

1. *Dependent variables are normally distributed*: Not normally distributed variables increase the chance of a false positive result. To check whether the dependent variables are normally distributed, the Shapiro-Wilk test of normality is used.
2. *Homogeneity of variances*: The variance should be the same for all observations, due to the huge dependence of the F-test on within-group variances. A Levene test of homogeneity of variances is carried out to check this assumption. If the Levene test results are not significant ($p > 0.05$), the assumption is valid that variances are equal enough, and it is safe to use the F-test in an ANOVA.
3. *All observations are independent*: By the design of the experiment, the teams have no interaction with each other. Therefore, the observations about the teams are totally independent of each other.

When the preconditions of an 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 k independent samples using medians instead of means as does the ANOVA test.

An ANOVA test shows only that the tested means are not equal to each other. In the same way, the Kruskal-Wallis test shows only that the tested medians are not equal to each other. In order to distinguish which means or medians differ significantly from which of the other means or medians, a pairwise comparison test needs to be carried out.

8 Gathered Data

For each experiment, E1 or E2, a single list of all ideas generated by all teams was created. Two domain experts classified the ideas with the classification procedure presented in Section 7.3. The experience in E1 with classifying ideas showed that classifying innovativeness of the ideas was more subjective than classifying relevance and feasibility of the ideas, for which the agreement between the two classifiers was 89.2%. Therefore, a third domain-expert classifier was employed to classify only the feasible ideas found by the first two classifiers for innovativeness. The third classifier classified both E1 and E2 data. When the classifications were done, 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 a strong correlation with each of the two other classifiers.

	Ideas				
			Innovative		
	Relevant	Feasible	(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

Table 1: Correlation Between the Classifiers' Classifications of Ideas

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

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. As shown in Table 2, the differences between the E1 and E2 ratios for the relevant and feasible ideas are clearly significant. Perhaps, the evaluators 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 13.

9 Data Preparation for Statistical analysis

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

9.1 Data Normalization

The values of some of the independent variables about a team were converted into nominal values and others were left unchanged.

Type of Teams	*No. of Teams*	*Creativity Score*		*RE Experience*		*Industrial RE Experience*		*Industrial Experience*		*No. CS Participants*		*No. SE Participants*		*No. Graduate Participants*	
		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 3: Combined Data about the Teams

Category	*Number of Raw Ideas*			*Number of Relevant Ideas*			*Number of Feasible Ideas*			*Number of Innovative Ideas*		
	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

Table 4: Combined Data of the Generated Ideas

- *MIX*: the team's MIX has one of the four possible nominal values, *0I*, *1I*, *2I*, or *3I*.
- *CR*: as shown in Figure 2, 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.

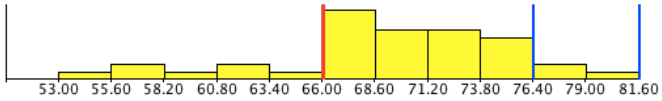


Fig. 2: Distribution of the Teams' Average Creativity Scores of the Participating Teams

- *REXP*: based on the distribution shown in Figure 3, 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.

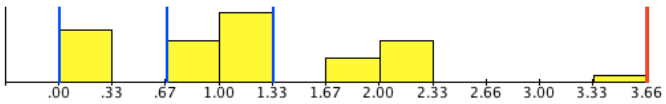


Fig. 3: Distribution of the Teams' Average RE Experience

- *IREXP*: based on the distribution shown in Figure 4, 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.

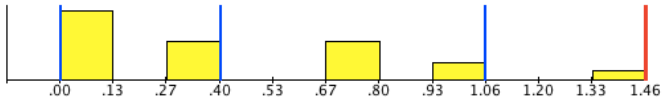


Fig. 4: Distribution of the Teams' Average Industrial RE Experience

- *IEXP*: based on the distribution shown in Figure 5, 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.

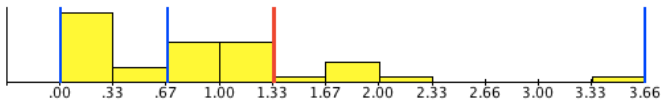


Fig. 5: Distribution of the Teams' Average Industrial Experience

- *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.

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

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 5: Variables of the Study

9.2 Data Normalization

As mentioned in Section 7.6, 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.

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 6: Test of Normality of the Dependent Variables

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 [23], which need to be considered beside normality.

1. *Skewness* is the extent by which a distribution leans to one side of the mean. That is, in a skewed distribution, the mean is not in the middle. When a distribution is

Nature of Distribution	*Skewness*	*Kurtosis*
Normal	0	2.90
Slightly Skewed	.45	3.53
Square Root Trans.	0	2.91
Moderately Skewed	.64	3.53
Logarithm Trans.	0	2.82
Extremely Skewed	2.04	9.54
Reciprocal Trans.	.03	2.88
Leptokurtic	0	9.16
Rectangular	0	1.80

Table 7: Acceptable Levels of Skewness (Adopted from [23])

	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: Skewness and Kurtosis Test Results of the Dependent Variables

skewed to the left, with what is called "negative skew", the mean is greater than the median. On the other hand, when a distribution is skewed to the right, with what is called "positive skew", the mean is smaller than the median [48]. SPSS generates for any distribution, its signed score of skewness and the standard error associated with the score. A skewness score is standardized by dividing it by its standard error. Table 7 shows that a distribution with a standard skewness score of greater than 2 is considered to be extremely skewed and therefore needs attention before applying an ANOVA. The optimal standard value for skewness is 0, but a score between -2 and $+2$ is considered acceptable.

2. *Kurtosis* is a measure of the peakedness versus flatness of a distribution [47]. It shows whether a distribution has a greater or less than normal proportion of extreme scores in each tail [45]. A more peaked than normal distribution has a negative kurtosis score and a flatter than normal distribution has a positive kurtosis score. As with skewness, a kurtosis score is standardized by dividing it by its standard error. While a standard kurtosis score near 0 is optimal, a standard kurtosis score between -2 and $+2$ is considered acceptable.

To avoid the bad effects of non-normality, in skewed distributions, the median is used instead of the mean. In a perfectly symmetric distribution, the mean is equal to the median, and therefore the skewness is 0. Table 8 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.

Category	***Normalized No. of Raw Ideas***			***Normalized No. of Relevant Ideas***			***Normalized No. of Feasible Ideas***			***Normalized No. of Innovative Ideas***		
	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: Normalized Combined Data of the Generated Ideas

9.2.1 Transforming Data into Normal Distribution

Blom's formula [8] is a rank-based method that can be used to normalize non-normally distributed data. In order to transform a dataset so that its distribution is as close as possible to being normal, the data need to be shifted, without changing ordering, to be symmetric around a focal point, with no skewing either to the left or to the right, i.e., to produce a bell-shaped curve. Thus, as a result of normalization, the mean, median, and mode of the data will end up being equal to or close to the focal point. For this study, without loss of generality and as is typical, we have chosen to normalize the data around zero. Table 9 shows the normalized versions of the data in Table 4.

Table 10 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.

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

Table 10: Test of Normality of the Dependent Variables after Normalization

Skewness and kurtosis are calculated once again. Table 11 shows that the skewness and kurtosis standard scores for all four dependent variables are within the acceptable range, even for NI.

	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

Table 11: Skewness and Kurtosis Test Results for the Dependent Variables after Normalization

Figure 6 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.

Q-Q plots are another way of verifying the normality of a set of data. In a Q-Q plot of a dataset, the more the data points gather around a straight line, the more normal is their distribution. Figure 7 shows on the left side, the Q-Q plots for the original dependent variables and on the right side, the Q-Q plots for the normalized dependent data variables.

The Q-Q plot for the original non-normalized data shows a significant deviation from a straight line for each dependent variable, and the Q-Q plot for the normalized data shows only a very small deviation from a straight line for each dependent variable. Therefore, it can be said that after normalization, the distribution of each dependent variable is at least moderately normal.

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, and its Q-Q plot shows only a small deviation from normality. 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.

9.3 Outliers

Irregular values in the data, referred to as "outliers", increase sample variance, which in turn reduces the F value of an ANOVA test. The smaller the F value, the greater the chances of incorrectly rejecting a null hypothesis [43] and committing a type I error. Consequently, outliers decrease the chances of showing the effect of an independent variable. It is therefore necessary to detect and remove outliers before any data analysis. However, it is possible that an outlier is a legitimate observation, and therefore, it needs to be examined carefully [24]. One condition that requires an outlier to be removed from the sample is when it is the result of an incorrect measurement, which, in this experiment, is hardly the case.

Boxplots are used to detect potential outliers. Figure 8 shows the boxplot of the four dependent variables grouped by the main independent variable of the study,

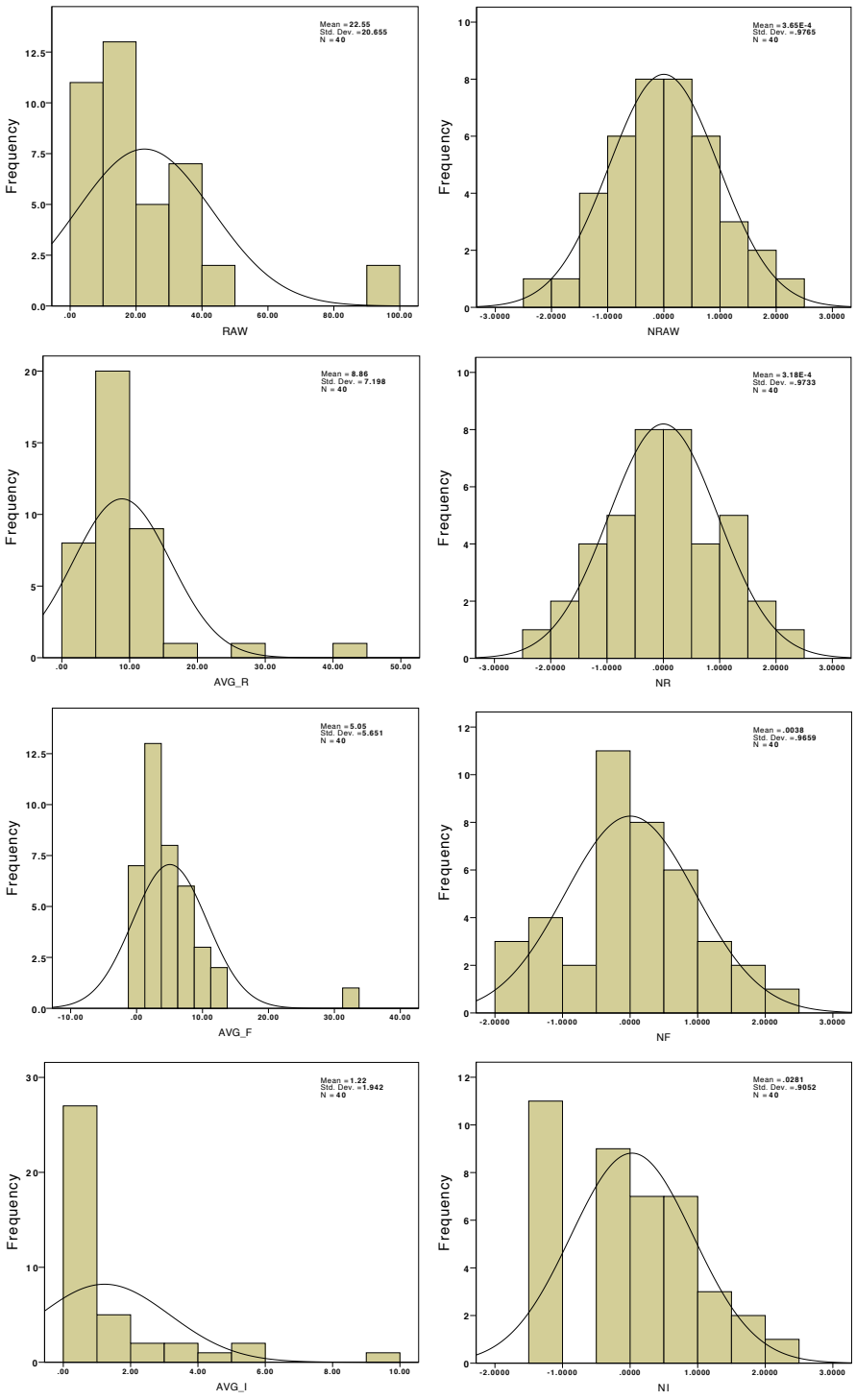
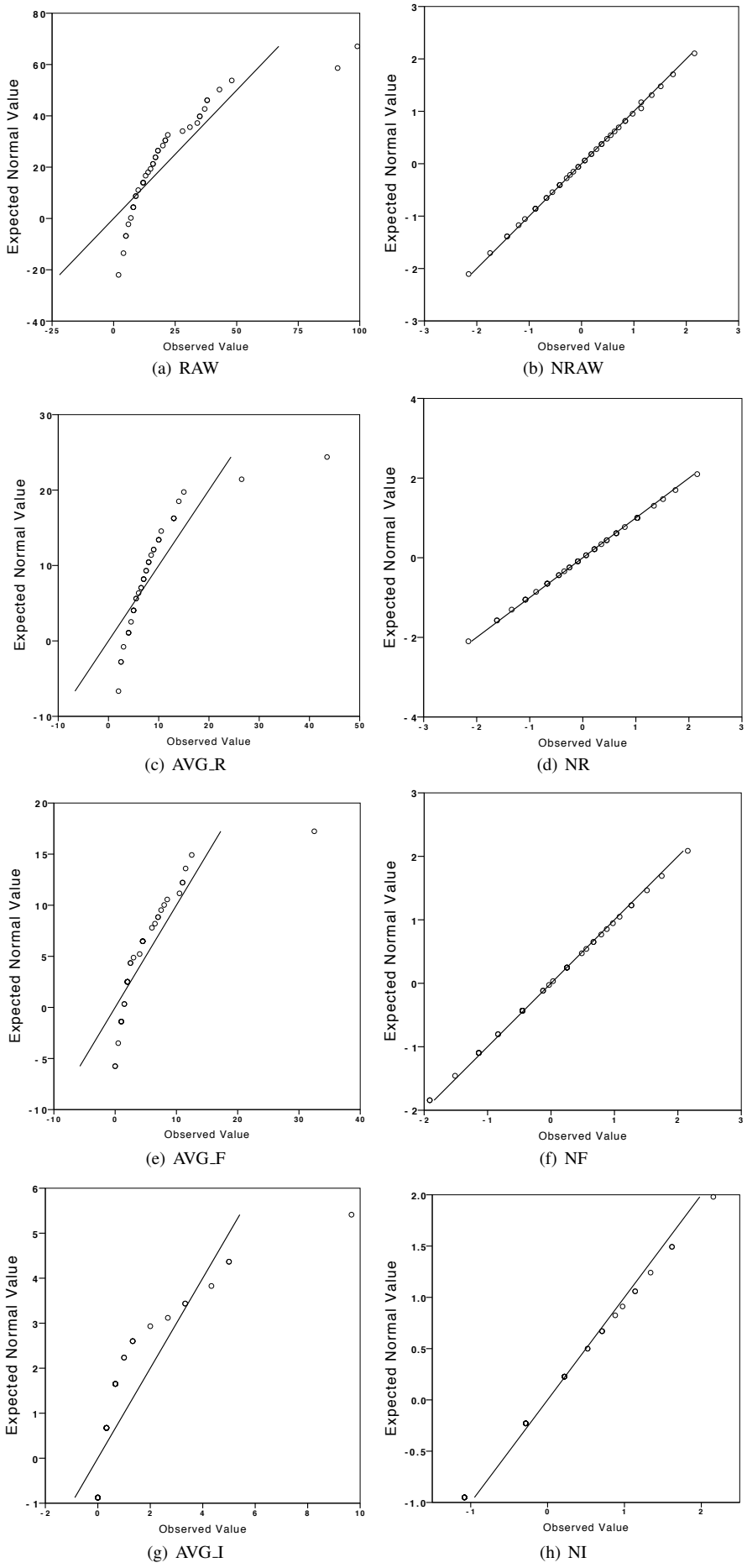


Fig. 6: Normality Plots of the Dependent Variables



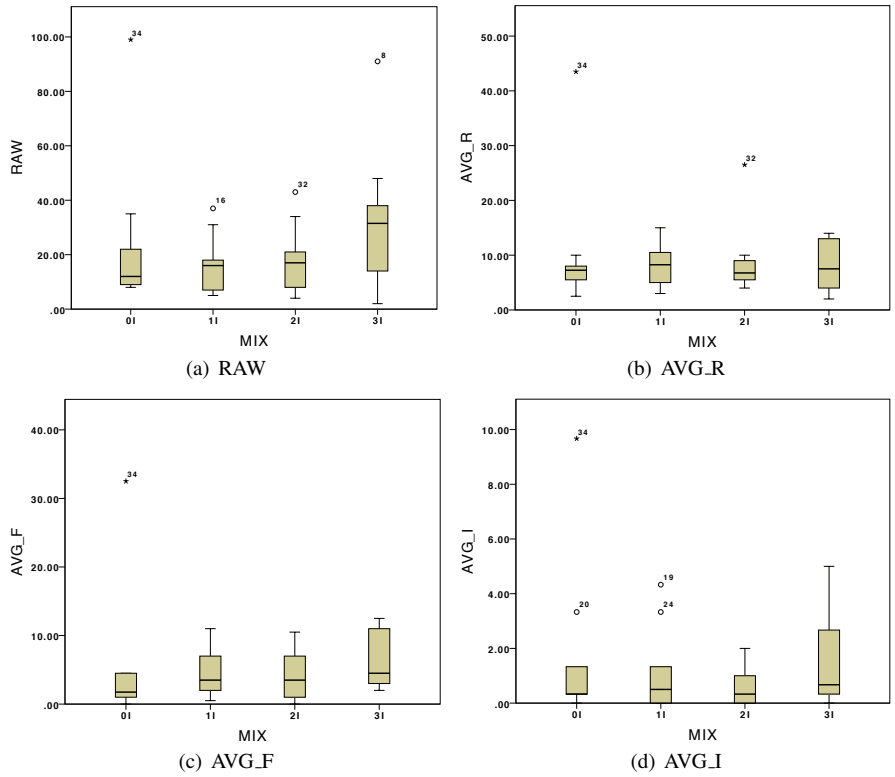
(a) RAW (b) NRAW

(c) AVG_R (d) NR

(e) AVG_F (f) NF

(g) AVG_I (h) NI

Fig. 7: Q-Q Plots of the Dependent Variables



(a) RAW (b) AVG_R

(c) AVG_F (d) AVG_I

Fig. 8: Boxplots of the Dependent Variables

MIX. Figure 8(a) shows that the values of RAW for Teams 8, 16, 32, and 34 are outliers. Figure 8(b) shows that the values of AVG_R for Teams 32 and 34 are outliers. Figure 8(c) shows that the value of AVG_F for Team 34 is an outlier. Figure 8(d) shows that the values of AVG_I for Teams 19, 20, 24, and 34 are outliers.

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", i.e., the study was done on the data including outliers.

9.3.1 Deeper Study of the Outliers

Outliers[2] produced about two times more RAW, AVG_R, AVG_F, and AVG_I than did non-outliers.

When forming teams for E1 and E2, the only independent variable, besides MIX, that was considered in forming teams, was the teams' CR values. The teams were

[2] Hereafter, an outlier is a team who has produced one or more values of dependent variables that are found to be outlier.

balanced by their CR values. It was not possible to balance also other independent variables. Therefore, there is a chance that teams are unbalanced in another independent variable that has a significant effect on the dependent variables, to the extent that some teams end up being outliers.

Compared to non-outliers, outliers had:

1. a higher average REXP, i.e., 2.29 for the outliers and 1.70 for the non-outliers,
2. a lower average IREXP, i.e., .71 for the outliers and 1.09 for the non-outliers,
3. a higher average NGRAD, i.e., 2.14 for the outliers and 1.73 for the non-outliers, and
4. a lower average NCS, i.e., 1.57 for the outliers and 2.06 for the non-outliers.

Other independent variables do not differ significantly. It turns out that the statistical analyses of Section 12 show that only two of these variables, NGRAD and NCS, have significant effects on the effectiveness of the participating teams. That is, teams with abnormal values for each of these two variables, are potentially outliers. Five out of the seven outlier teams have high levels of REXP. Therefore, it appears that REXP is the main factor causing the difference between outliers and non-outliers.

There is only one team, Team 34, for whom the value of each of the four dependent variables, i.e., RAW, AVG_R, AVG_F, and AVG_I, is an outlier. Team 34 is a *0I* team and the values of its independent variables are similar to the average values of independent variables of all outliers, including a high level of REXP. Therefore, Team 34 seems to be a real outlier.

For the teams whose value of AVG_R is an outlier and the teams whose value of RAW is an outlier, the average values of the independent variables do not differ significantly from the average values of the whole set of outliers.

9.4 Factor Analysis

As a statistical method, factor analysis is used to shrink a large number of independent variables to a potentially smaller set of unobserved variables called *factors*[3]. The produced set of factors is supposed to be the main driver behind the dependent variables [26]. 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.

Principal Factor Analysis (PFA)[4] is the most common method used in social sciences [46] to find a smaller number of factors to examine. The Kaiser-Meyer-Olkin (KMO) measure tests a set of variables for adequacy for factor analysis. When the

[3] Factors are treated as independent variables in the statistical analyses.

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

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.656
Bartlett Test of Sphericity	Approx. Chi-Square	141.694
	df	21
	p	.000

Table 12: KMO and Bartlett Test Results

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 13: Rotated Factor Matrix

KMO measure of a set of variables is greater than 0.5, factor analysis can be performed [27]. Table 12 shows that the KMO measure of the set of independent variables is 0.656, which is greater than 0.5. The other test result shown in Table 12 is the Bartlett test, which indicates whether there is any relationship among the tested variables. A p-value of less than 0.05 in a Bartlett test shows that there is a relationship, and, therefore, factor analysis makes sense. In this case, a p equal to 0.000 means that there is a very strong relationship among the variables.

The results of the factor analysis are shown in Table 13. The two factors indicated in Table 13 as Factor 1 and Factor 2 are the two factors identified by factor analysis. The numbers in Table 13 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 means that the variable loads more strongly on the factor and loads more weakly on the other factor.

Figure 9 plots the loading values of Table 13. 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:

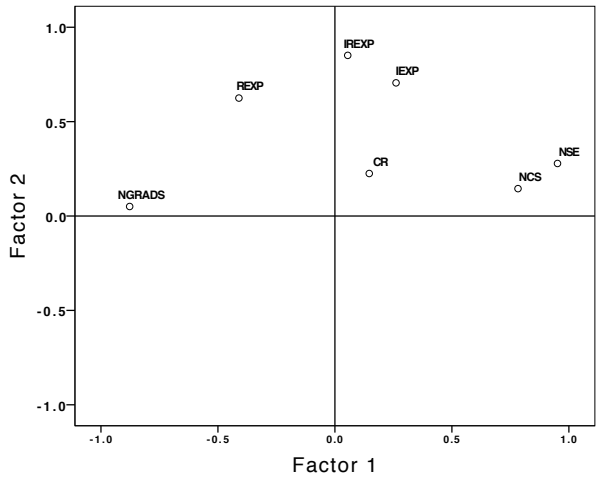


Fig. 9: Factor Loading Plot

- Low: values 0 to 3, and
- High: values 4 to 6.

Because the 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 12.9. The two factors are studied in detail also by means of a one-way ANOVA in Section 12.

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.

10 Hypotheses

Based on the independent variables listed in Section 7.5 and the two factors identified in Section 9.4, 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 14. 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 H_{1_1}, H_{1_0}, H_{2_1}, H_{2_0}, H_{3_1}, H_{3_0}, H_{4_1}, and H_{4_0}, respectively [37].

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 14: Final List of Hypotheses

11 Initial Observations

Initial assessments of support for the hypotheses are drawn from plots of the unfiltered and filtered dependent variables data against each of the independent variables.

Each subsection is about the effect on the unfiltered and filtered dependent variables of one independent variable for the purpose of testing one hypothesis and its null hypothesis. To assess this effect, the subsection gives:

1. a set of plots showing the median numbers of the kinds of ideas generated by teams including the outliers, i.e., the unfiltered dependent variables plotted against the new independent variables,
2. a set of plots showing the median numbers of the kinds of ideas generated by teams without the outliers, i.e., the filtered dependent variables plotted against the new independent variables, and
3. an interpretation of the plots.

Due to the skewness of the distributions of the data, demonstrated in Section 9.2, these plots plot the medians, instead of the means, of the data.

When the interpretations of the plots are written in careful, precise natural language, the narrative is, frankly, mind numbing. Thus, Table 96, summarizing the interpretations, is provided in the last subsection of this section, Section 11.11.

11.1 Impact of MIX

Figure 10(a) shows that the medians of the unfiltered RAW generated by teams are positively correlated with the teams' MIX. Figure 11(a) shows that the plot of the medians of the filtered RAW generated by teams is similar to that of Figure 10(a). Thus, for the medians of the RAW values, removal of the outliers makes no real difference.

Figure 10(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' MIX. Figure 11(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 10(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 10(c) shows that the medians of the unfiltered AVG_F generated by teams are positively correlated with the teams' MIX. Figure 11(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 10(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 10(d) shows that the medians of the unfiltered AVG_I generated by teams are not correlated with the teams' MIX. Figure 11(d) shows that the medians of the unfiltered AVG_I generated by teams are partially positively correlated with the teams' MIX. Thus, for the medians of the AVG_I values, removal of the outliers makes a difference.

These plots show that the medians of the unfiltered and filtered RAW, AVG_F, and AVG_I are highest for the teams with MIX = "*3I*", and that the medians of the unfiltered and filtered AVG_R are highest for the teams with MIX = "*1I*".

Overall, initially, as with E1, it appears that H_{MIX_1} is supported but H_{MIX_0} is not supported.

11.2 Impact of CR

Figure 12(a) shows that the medians of the unfiltered RAW generated by teams are partially negatively correlated with the teams' CR. Figure 13(a) shows that the medians of the filtered RAW generated by teams are not correlated with the teams' CR. Thus, for the medians of the RAW values, removal of the outliers makes a difference.

Figure 12(b) shows that the medians of the unfiltered AVG_R generated by teams are negatively correlated with the teams' CR. Figure 13(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 12(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 12(c) shows that the medians of the unfiltered AVG_F generated by teams are not correlated with the teams' CR. Figure 13(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 12(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 12(d) shows that the medians of the unfiltered AVG_I generated by teams are not correlated with the teams' CR. Figure 13(d) shows that the plot of the medians of the filtered AVG_I generated by teams is similar to that of Figure 12(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

These plots show that the medians of the unfiltered and filtered RAW, AVG_F, and AVG_I are highest for the teams with CR = "Medium", and that the medians of the unfiltered and filtered AVG_R are highest for the teams with CR = "Low".

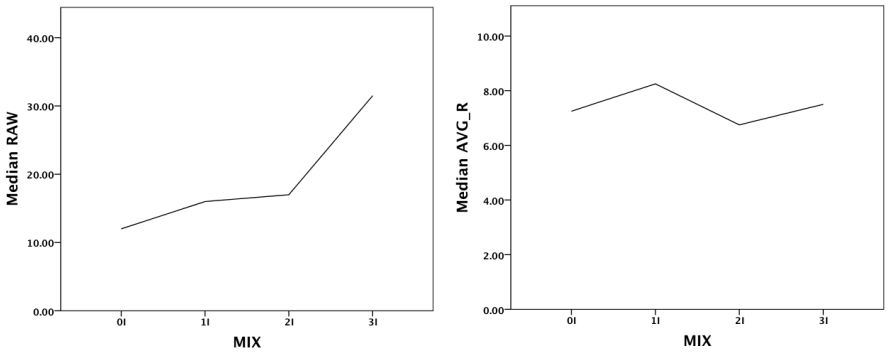
Overall, initially, as with E1, it appears that H_{CR_0} is supported and that hypothesis H_{CR_1} is not supported.

11.3 Impact of REXP

Figure 14(a) shows that the medians of the unfiltered RAW generated by teams are not correlated with the teams' REXP. Figure 15(a) shows that the plot of the medians of the filtered RAW generated by teams is quite similar to that of Figure 14(a). Thus, for the medians of the RAW values, removal of the outliers makes no real difference.

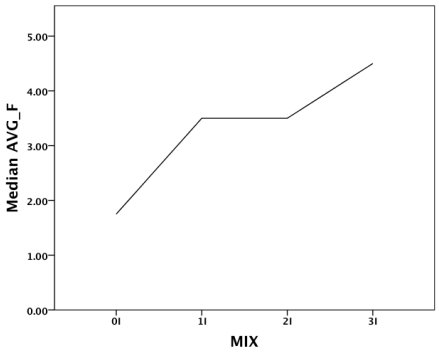
Figure 14(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' REXP. Figure 15(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 14(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 14(c) shows that the medians of the unfiltered AVG_F generated by teams are not correlated with the teams' REXP. Figure 15(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 14(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

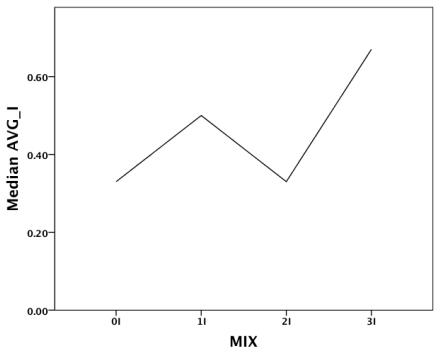


(a) RAW

(b) AVG_R

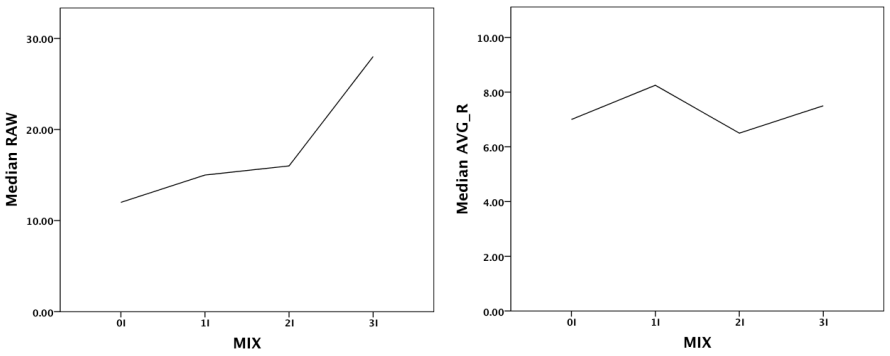


(c) AVG_F



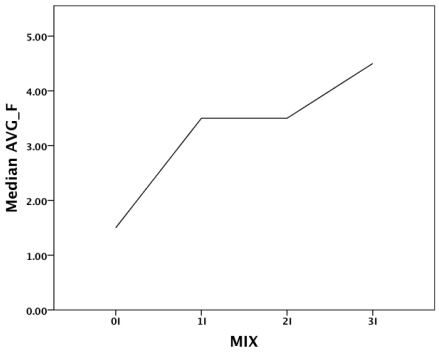
(d) AVG_I

Fig. 10: Ideas vs. MIX (Unfiltered)

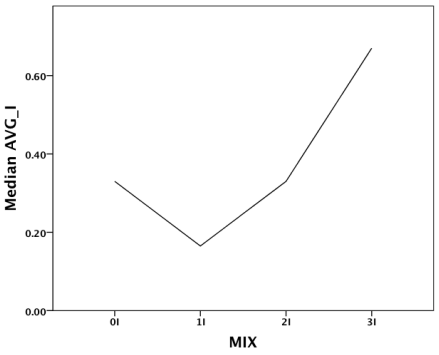


(a) RAW

(b) AVG_R

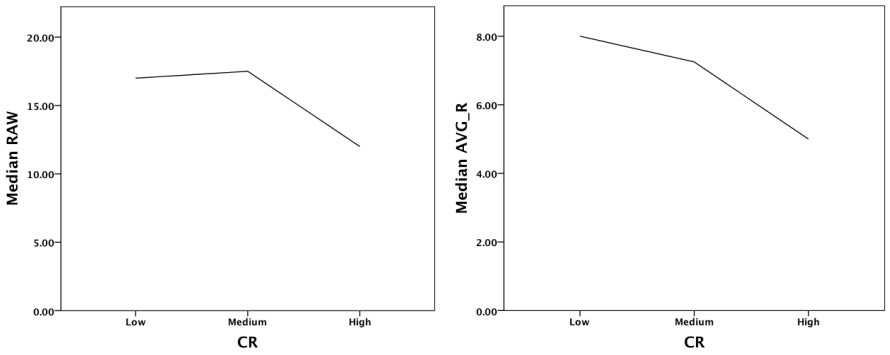


(c) AVG_F



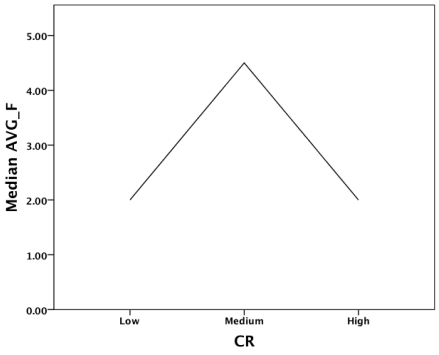
(d) AVG_I

Fig. 11: Ideas vs. MIX (Filtered)

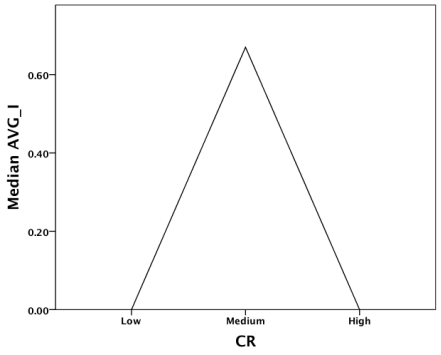


(a) RAW

(b) AVG_R

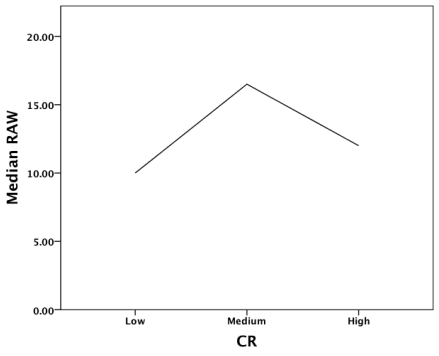


(c) AVG_F

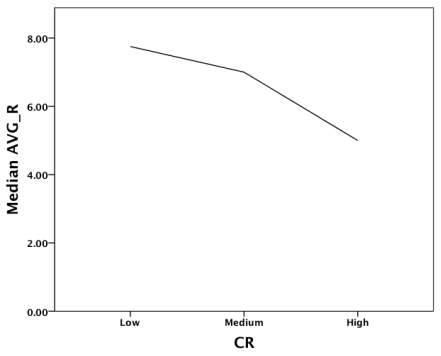


(d) AVG_I

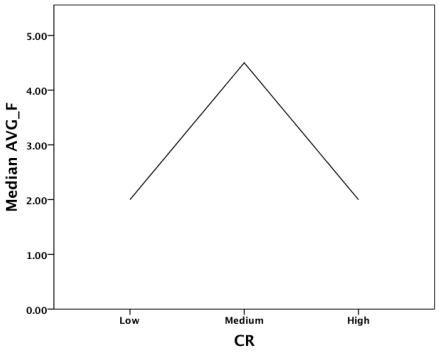
Fig. 12: Ideas vs. CR (Unfiltered)



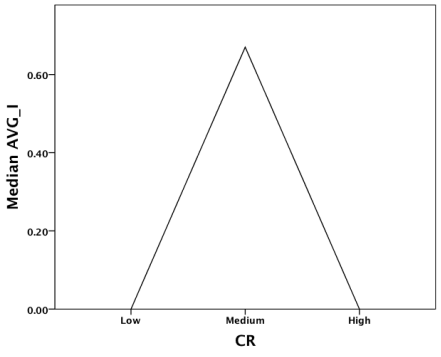
(a) RAW



(b) AVG_R



(c) AVG_F



(d) AVG_I

Fig. 13: Ideas vs. CR (Filtered)

Figure 14(d) shows that the medians of the unfiltered AVG_I generated by teams are partially positively correlated with the teams' REXP. Figure 15(d) shows that the medians of the filtered AVG_I generated by teams are not correlated with the teams' REXP. Thus, for the medians of the AVG_I values, removal of the outliers makes a difference.

These plots show that the medians of the unfiltered and filtered RAW, AVG_R, and AVG_F are highest for the teams with REXP = "None", that the median of the unfiltered AVG_I is highest for the teams with REXP = "High", and that the median of the filtered AVG_I is highest for the teams with REXP = "None" or REXP = "Medium".

Overall, initially, as with E1, it appears that H_{REXP_0} is supported and that hypothesis H_{REXP_1} is not supported.

11.4 Impact of IREXP

Figure 16(a) shows that the medians of the unfiltered RAW generated by teams are positively correlated with the teams' IREXP. Figure 17(a) shows that the medians of the filtered RAW generated by teams are partially positively correlated with the teams' IREXP. Thus, for the medians of the RAW values, removal of the outliers makes a slight difference.

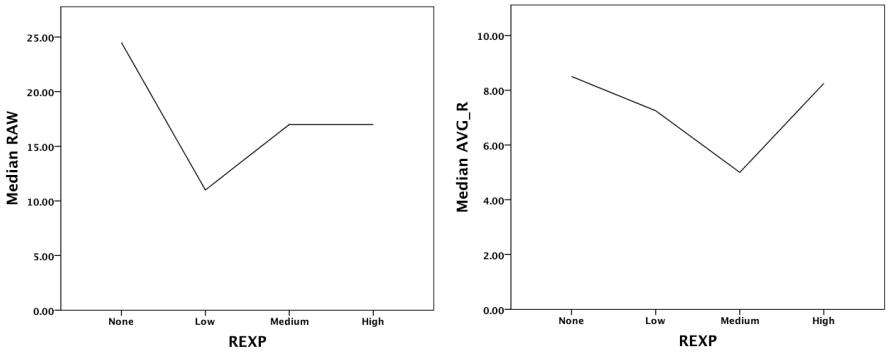
Figure 16(b) shows that the medians of the unfiltered AVG_R generated by teams are partially positively correlated with the teams' IREXP. Figure 17(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 16(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 16(c) shows that the medians of the unfiltered AVG_F generated by teams are partially positively correlated with the teams' IREXP. Figure 17(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 16(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 16(d) shows that the medians of the unfiltered AVG_I generated by teams are partially positively correlated with the teams' IREXP. Figure 17(d) shows that the plot of the medians of the filtered AVG_I generated by teams is similar to that of Figure 16(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

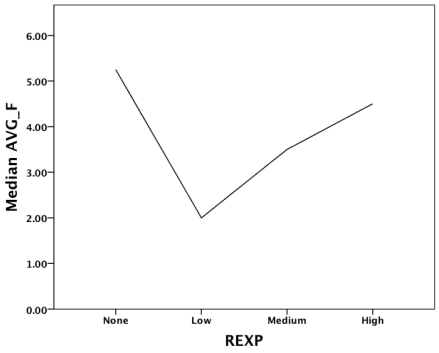
These plots show that the median of the unfiltered RAW is highest for the teams with IREXP = "High", that the median of the filtered RAW is highest for the teams with IREXP = "Medium", and that the medians of the unfiltered and filtered AVG_R, AVG_F, and AVG_I are highest for the teams with IREXP = "High".

Overall, initially, it appears that H_{IREXP_1} is supported and that hypothesis H_{IREXP_0} is not supported.

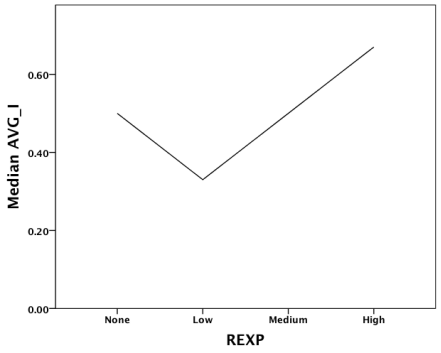


(a) RAW

(b) AVG_R

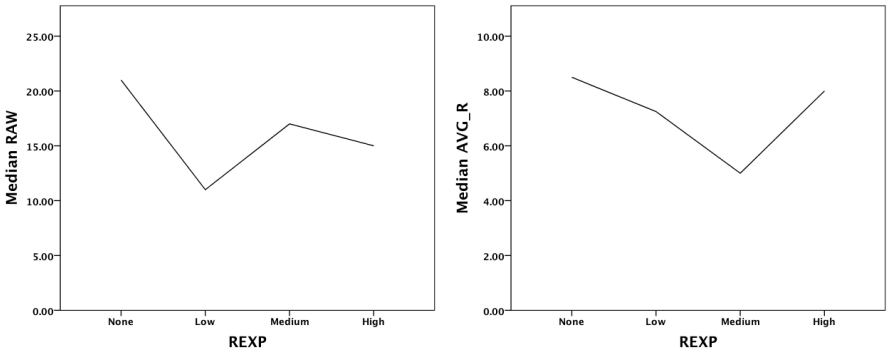


(c) AVG_F



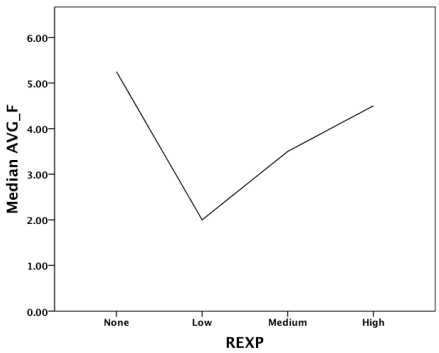
(d) AVG_I

Fig. 14: Ideas vs. REXP (Unfiltered)

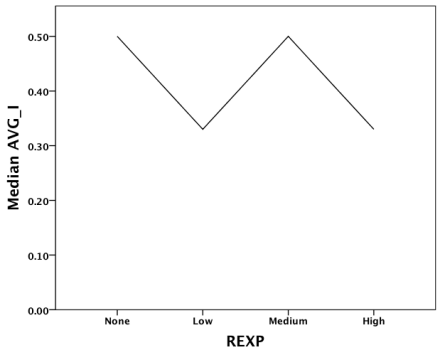


(a) RAW

(b) AVG_R

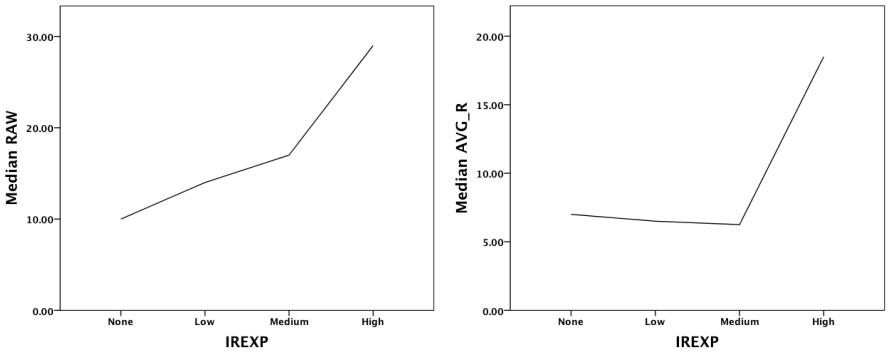


(c) AVG_F



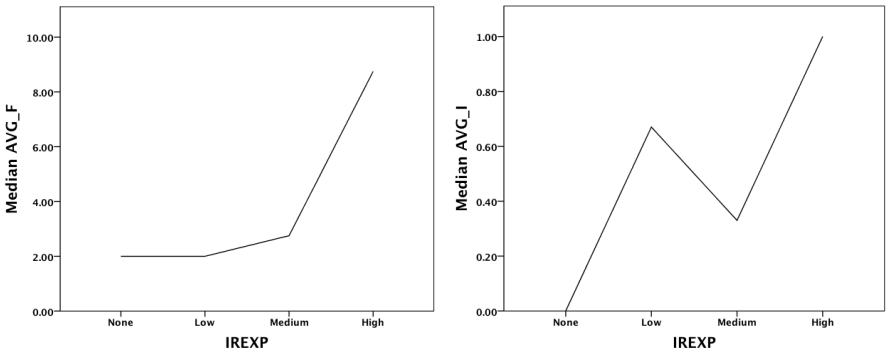
(d) AVG_I

Fig. 15: Ideas vs. REXP (Filtered)



(a) RAW

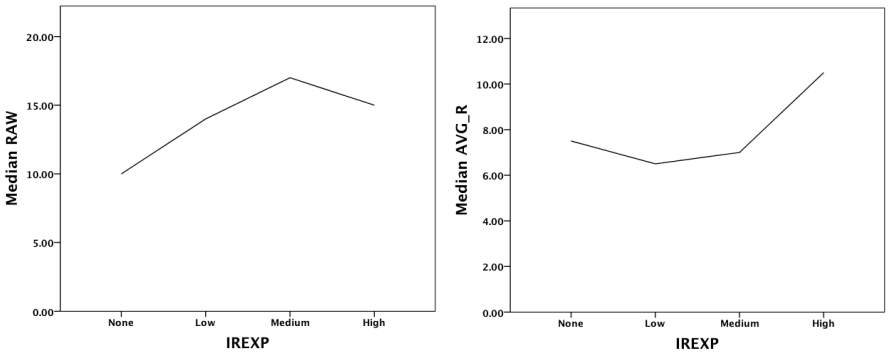
(b) AVG_R



(c) AVG_F

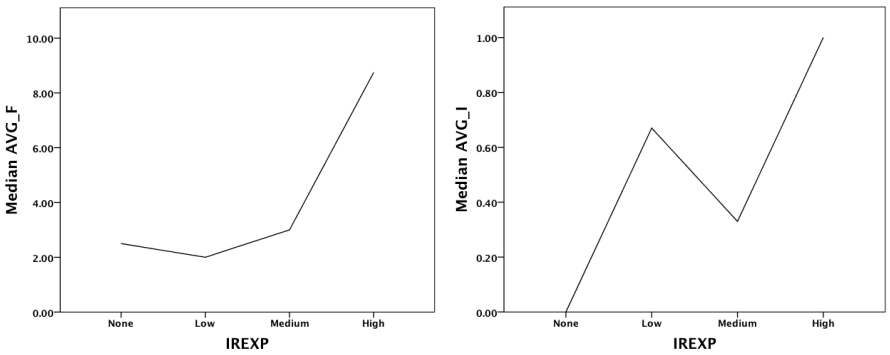
(d) AVG_I

Fig. 16: Ideas vs. IREXP (Unfiltered)



(a) RAW

(b) AVG_R



(c) AVG_F

(d) AVG_I

Fig. 17: Ideas vs. IREXP (Filtered)

11.5 Impact of IEXP

Figure 18(a) shows that the medians of the unfiltered RAW generated by teams are partially positively correlated with the teams' IEXP. Figure 19(a) shows that the medians of the filtered RAW generated by teams are partially positively correlated with the teams' IEXP. Nevertheless the plots are different enough to say that, for the medians of the RAW values, removal of the outliers makes a slight difference.

Figure 18(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' IEXP. Figure 19(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 18(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 18(c) shows that the medians of the unfiltered AVG_F generated by teams are partially positively correlated with the teams' IEXP. Figure 19(c) shows that the plot of the medians of the filtered AVG_F generated by teams is quite similar to that of Figure 18(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 18(d) shows that the medians of the unfiltered AVG_I generated by teams are partially positively correlated with the teams' IEXP. Figure 19(d) shows that the medians of the filtered AVG_I generated by teams are partially positively correlated with the teams' IEXP. Nevertheless the plots are different enough to say that, for the medians of the AVG_I values, removal of the outliers makes a slight difference.

These plots show that the medians of the unfiltered and filtered RAW and AVG_F are highest for the teams with IEXP = "Medium", that the medians of the unfiltered and filtered AVG_R are highest for the teams with IEXP = "Low", that the median of the unfiltered AVG_I is highest for the teams with IEXP = "Low" or IEXP = "Medium", and that the median of the filtered AVG_I is highest for the teams with IEXP = "Medium".

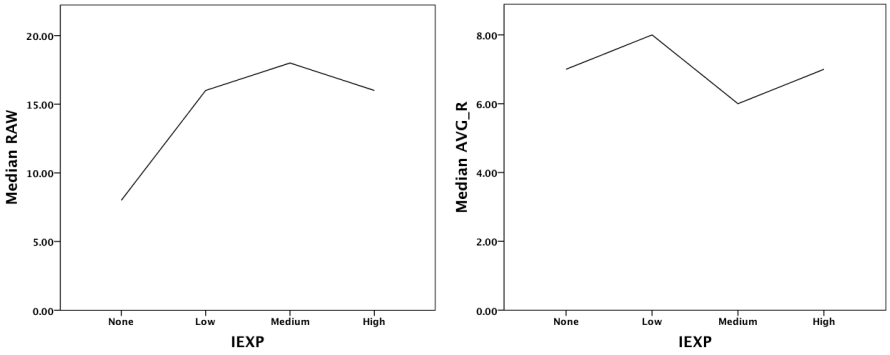
Overall, initially, as with E1, it appears that H_{IEXP_1} is supported and that hypothesis H_{IEXP_0} is not supported.

11.6 Impact of NCS

Figure 20(a) shows that the medians of the unfiltered RAW generated by teams are positively correlated with the teams' NCS. Figure 21(a) shows that the medians of the filtered RAW generated by teams are partially positively correlated with the teams' NCS. Thus, for the medians of the RAW values, removal of the outliers makes a slight difference.

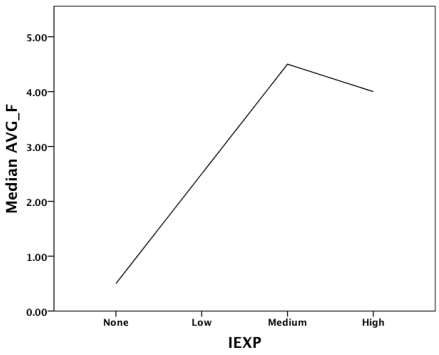
Figure 20(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' NCS. Figure 21(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 20(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 20(c) shows that the medians of the unfiltered AVG_F generated by teams are partially positively correlated with the teams' NCS. Figure 21(c) shows that the

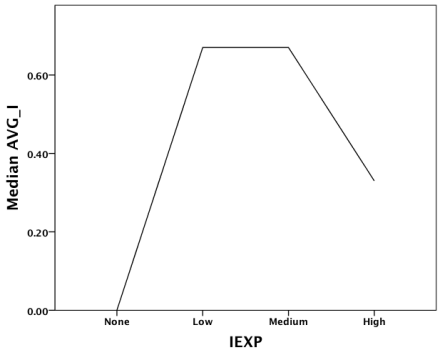


(a) RAW

(b) AVG_R

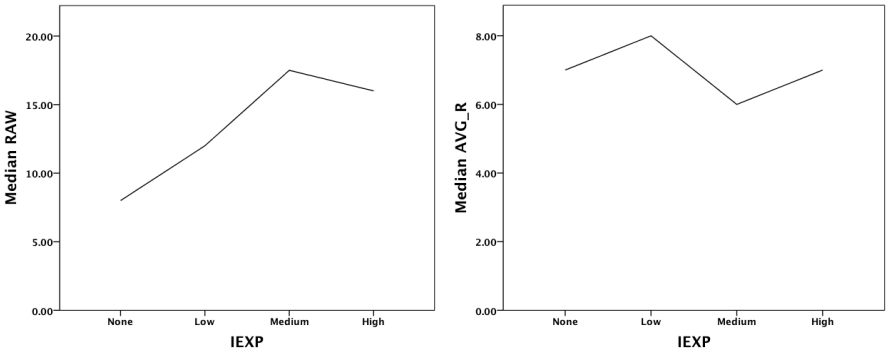


(c) AVG_F



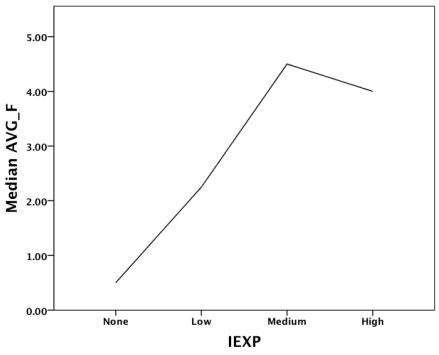
(d) AVG_I

Fig. 18: Ideas vs. IEXP (Unfiltered)

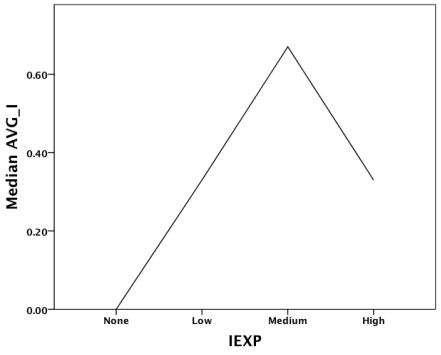


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Fig. 19: Ideas vs. IEXP (Filtered)

plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 20(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 20(d) shows that the medians of the unfiltered AVG_I generated by teams are partially positively correlated with the teams' NCS. Figure 21(d) shows that the plot of the medians of the filtered AVG_I generated by teams is similar to that of Figure 20(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

These plots show that the medians of the unfiltered and filtered RAW are highest for the teams with NCS = 3, that the medians of the unfiltered and filtered AVG_R are highest for the teams with NCS = 2, that the median of the unfiltered AVG_F is highest for the teams with NCS = 2 or NCS = 3, that the median of the filtered AVG_F is highest for the teams with NCS = 2, and that the medians of the unfiltered and filtered AVG_I are highest for the teams with NCS = 2 or NCS = 3.

Overall, initially, it appears that H_{NCS_1} is supported and that hypothesis H_{NCS_0} is not supported.

11.7 Impact of NSE

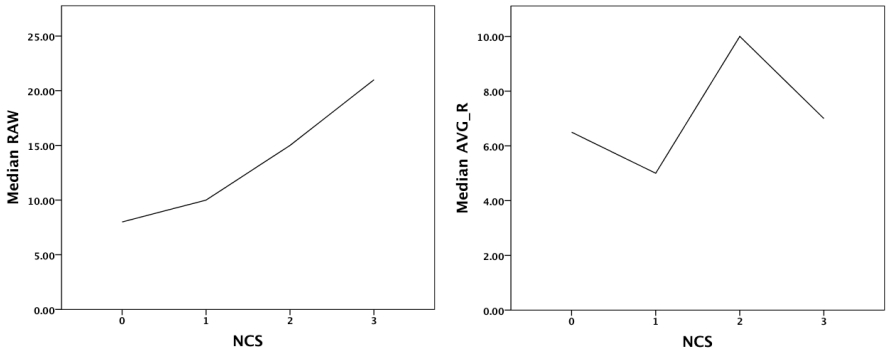
Figure 22(a) shows that the medians of the unfiltered RAW generated by teams are positively correlated with the teams' NSE. Figure 23(a) shows that the plot of the medians of the filtered RAW generated by teams is quite similar to that of Figure 22(a). Thus, for the medians of the RAW values, removal of the outliers makes no real difference.

Figure 22(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' NSE. Figure 23(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 22(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 22(c) shows that the medians of the unfiltered AVG_F generated by teams are partially positively correlated with the teams' NSE. Figure 23(c) shows that the plot of the medians of the filtered AVG_F generated by teams is quite similar to that of Figure 22(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

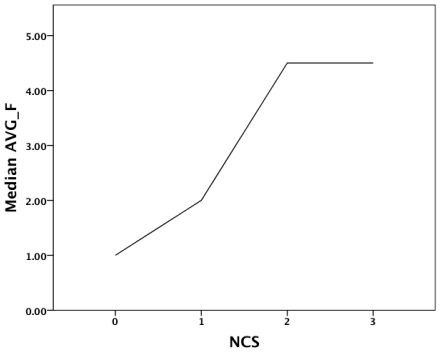
Figure 22(d) shows that the medians of the unfiltered AVG_I generated by teams are partially positively correlated with the teams' NSE. Figure 23(d) shows that the plot of the medians of the filtered AVG_I generated by teams is quite similar to that of Figure 22(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

These plots show that the medians of the unfiltered and filtered RAW are highest for the teams with NSE = 3, that the medians of the unfiltered and filtered AVG_R and AVG_F are highest for the teams with NSE = 2, that the median of the unfiltered AVG_I is highest for the teams with NSE = 2, that the median of the filtered AVG_I is highest for the teams with NSE = 2 or NSE = 3.

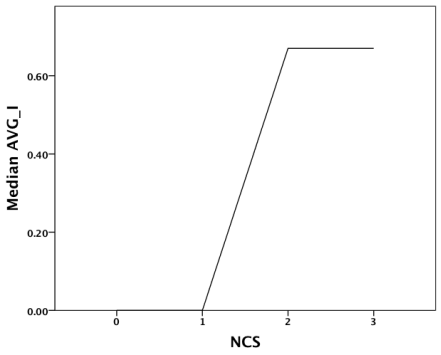


(a) RAW

(b) AVG_R

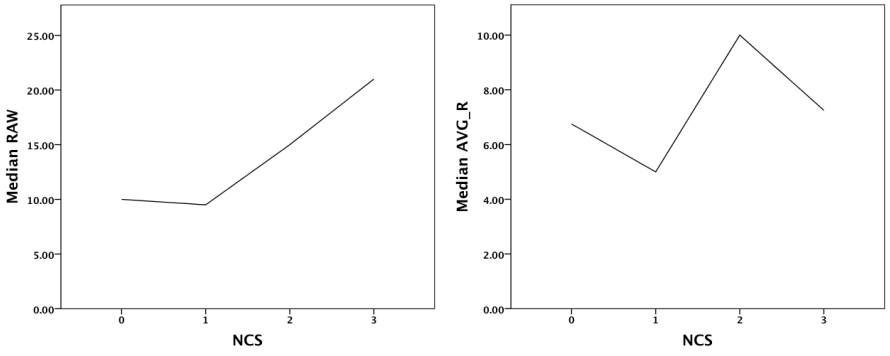


(c) AVG_F



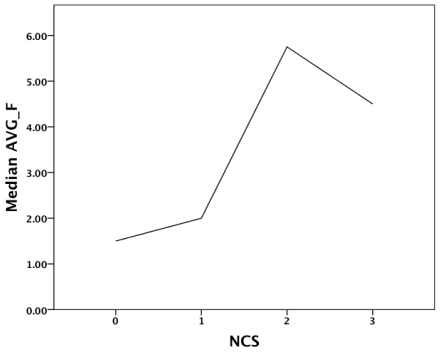
(d) AVG_I

Fig. 20: Ideas vs. NCS (Unfiltered)

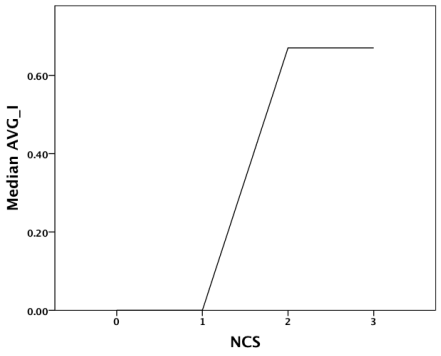


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Fig. 21: Ideas vs. NCS (Filtered)

Overall, initially, it appears that H_{NSE_1} is supported and that hypothesis H_{NSE_0} is not supported.

11.8 Impact of NGRAD

Figure 24(a) shows that the medians of the unfiltered RAW generated by teams are negatively correlated with the teams' NGRAD. Figure 25(a) shows that the medians of the filtered RAW generated by teams are partially negatively correlated with the teams' NGRAD. Thus, for the medians of the RAW values, removal of the outliers makes a slight difference.

Figure 24(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' NGRAD. Figure 25(b) shows that the plot of the medians of the filtered AVG_R generated by teams is quite similar to that of Figure 24(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 24(c) shows that the medians of the unfiltered AVG_F generated by teams are negatively correlated with the teams' NGRAD. Figure 25(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 24(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 24(d) shows that the medians of the unfiltered AVG_I generated by teams are partially negatively correlated with the teams' NGRAD. Figure 25(d) shows that the plot of the medians of the filtered AVG_I generated by teams is similar to that of Figure 24(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

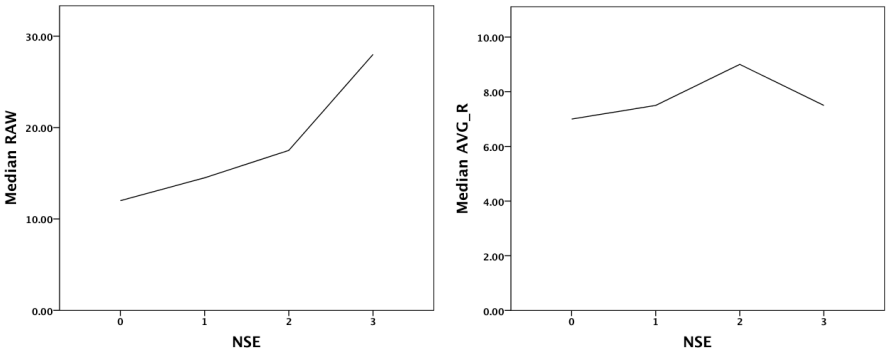
These plots show that the median of the unfiltered RAW is highest for the teams with NGRAD = 0, that the median of the filtered RAW is highest for the teams with NGRAD = 0 or NGRAD = 1, that the median of the unfiltered AVG_R is highest for the teams with NGRAD = 0 or NGRAD = 3, that the median of the filtered AVG_R is highest for the teams with NGRAD = 0, that the medians of the unfiltered and filtered AVG_F are highest for the teams with NGRAD = 0, and that the medians of the unfiltered and filtered AVG_I are highest for the teams with NGRAD = 1.

Overall, initially, it appears that H_{NGRAD_0} is supported and that hypothesis H_{NGRAD_1} is not supported.

11.9 Impact of EDU

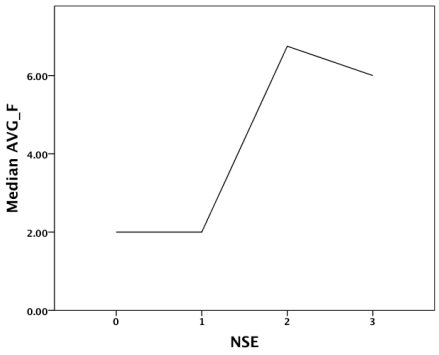
Figure 26(a) shows that the medians of the unfiltered RAW generated by teams are positively correlated with the teams' EDU. Figure 27(a) shows that the plot of the medians of the filtered RAW generated by teams is similar to that of Figure 26(a). Thus, for the medians of the RAW values, removal of the outliers makes no real difference.

Figure 26(b) shows that the medians of the unfiltered AVG_R generated by teams are positively correlated with the teams' EDU. Figure 27(b) shows that the plot of the

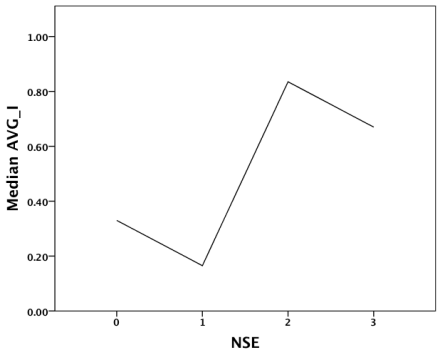


(a) RAW

(b) AVG_R

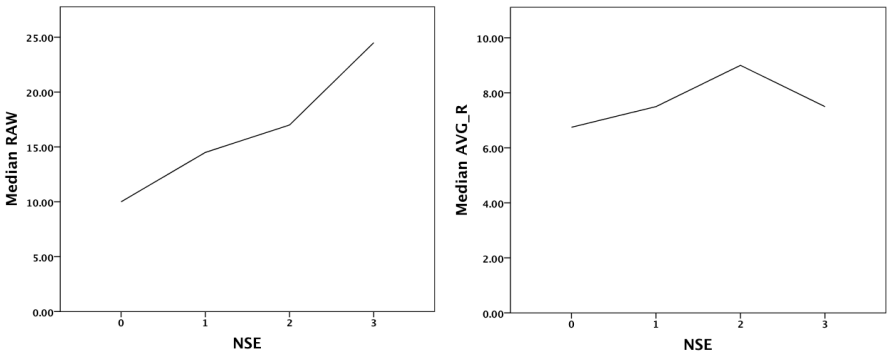


(c) AVG_F



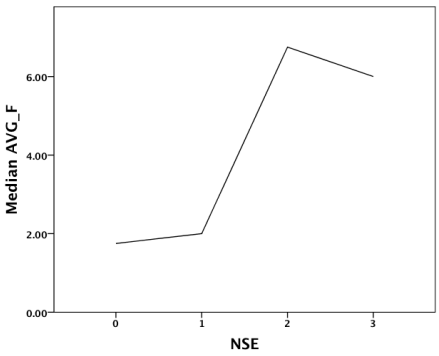
(d) AVG_I

Fig. 22: Ideas vs. NSE (Unfiltered)

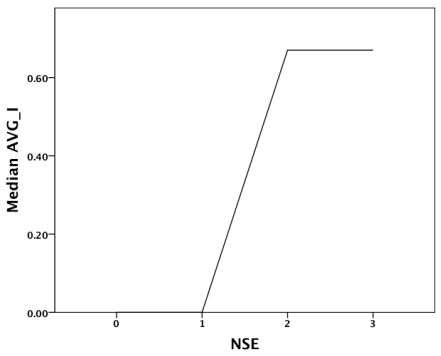


(a) RAW

(b) AVG_R

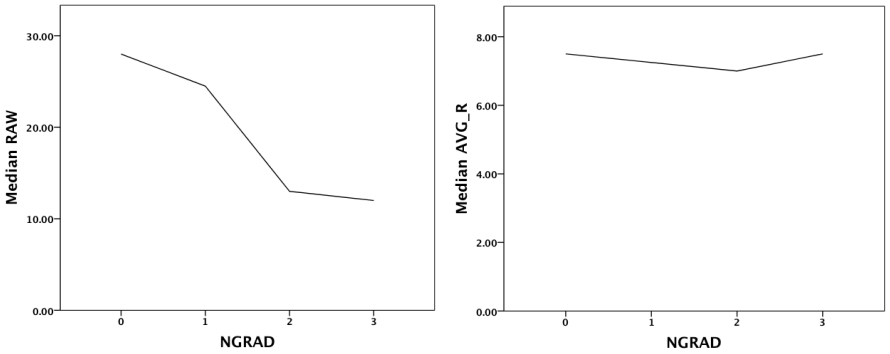


(c) AVG_F



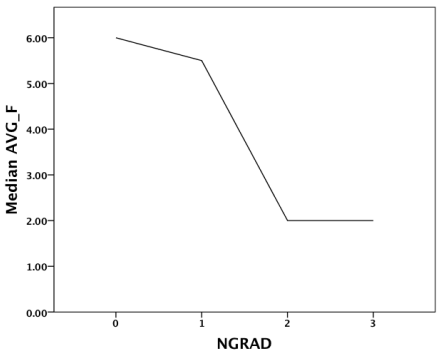
(d) AVG_I

Fig. 23: Ideas vs. NSE (Filtered)

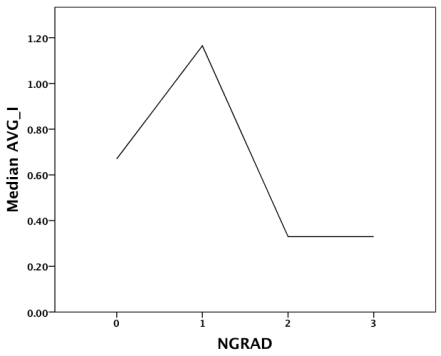


(a) RAW

(b) AVG_R

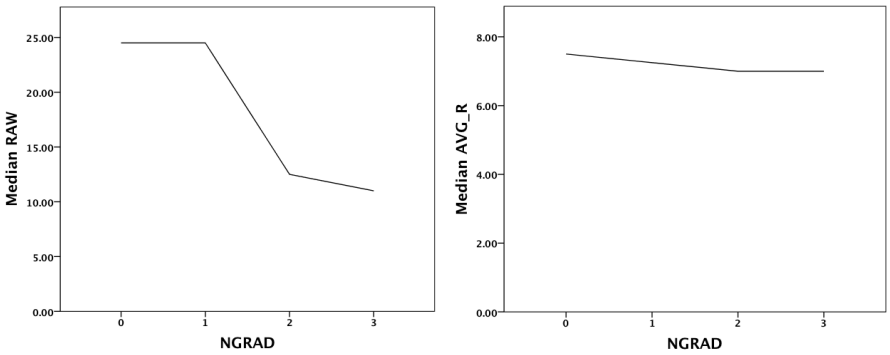


(c) AVG_F



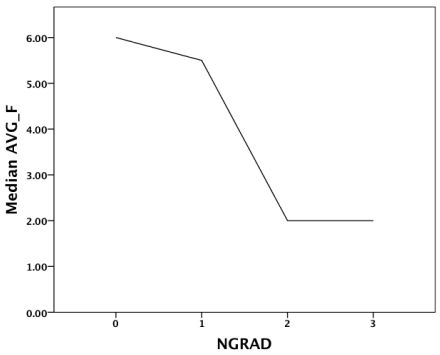
(d) AVG_I

Fig. 24: Ideas vs. NGRAD (Unfiltered)

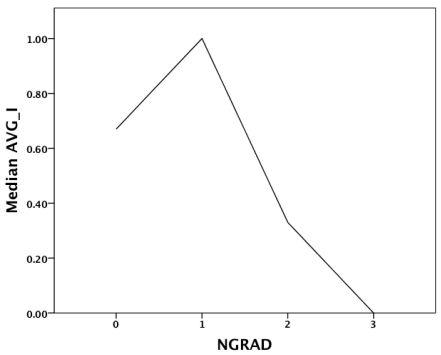


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Fig. 25: Ideas vs. NGRAD (Filtered)

medians of the filtered AVG_R generated by teams is similar to that of Figure 26(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 26(c) shows that the medians of the unfiltered AVG_F generated by teams are positively correlated with the teams' EDU. Figure 27(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 26(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 26(d) shows that the medians of the unfiltered AVG_I generated by teams are positively correlated with the teams' EDU. Figure 27(d) shows that the plot of the medians of the filtered AVG_R generated by teams is quite similar to that of Figure 26(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

These plots show that the medians of the unfiltered and filtered RAW, AVG_R, AVG_F, and AVG_I are highest for the teams with EDU = "High".

Overall, initially, it appears that H_{EDU_1} is supported and that hypothesis H_{EDU_0} is not supported.

11.10 Impact of EXP

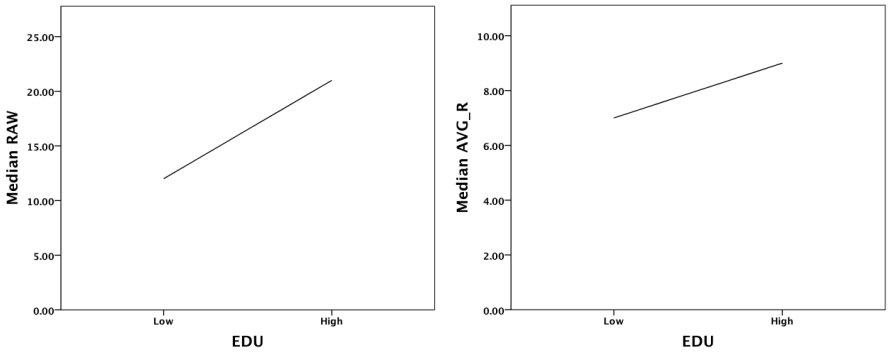
Figure 28(a) shows that the medians of the unfiltered RAW generated by teams are partially positively correlated with the teams' EXP. Figure 29(a) shows that the plot of the medians of the filtered RAW generated by teams is quite similar to that of Figure 28(a). Thus, for the medians of the RAW values, removal of the outliers makes no real difference.

Figure 28(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' EXP. Figure 29(b) shows that the plot of the medians of the filtered AVG_R generated by teams is quite similar to that of Figure 28(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 28(c) shows that the medians of the unfiltered AVG_F generated by teams are partially positively correlated with the teams' EXP. Figure 29(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 28(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

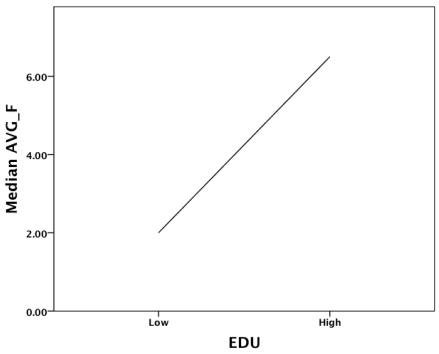
Figure 28(d) shows that the medians of the unfiltered AVG_I generated by teams are not correlated with the teams' EXP. Figure 29(d) shows that the plot of the medians of the filtered AVG_I generated by teams is quite similar to that of Figure 28(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

These plots show that the median of the unfiltered RAW is highest for the teams with EXP = "Medium", that the median of the filtered RAW is highest for the teams with EXP = "Medium" or EXP = "High", that the medians of the unfiltered and filtered AVG_R are highest for the teams with EXP = "High", and that the medians

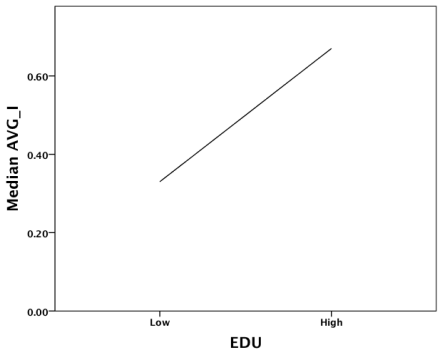


(a) RAW

(b) AVG_R

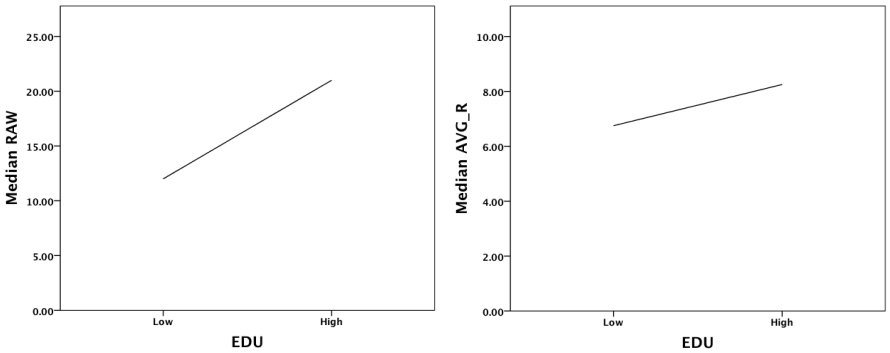


(c) AVG_F



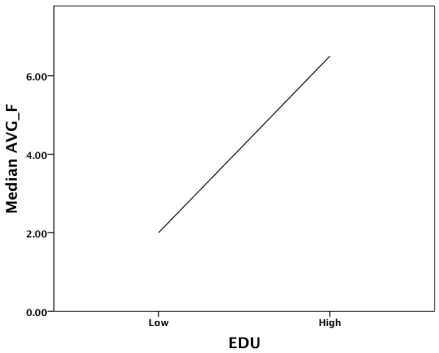
(d) AVG_I

Fig. 26: Ideas vs. EDU (Unfiltered)

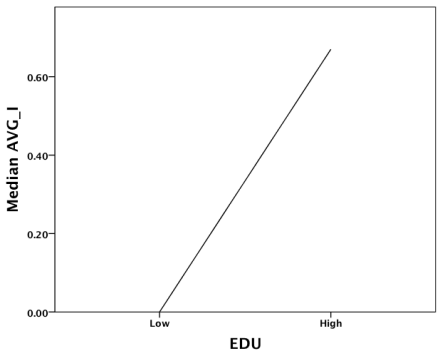


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Fig. 27: Ideas vs. EDU (Filtered)

of the unfiltered and filtered AVG_F, and AVG_I are highest for the teams with EXP = "Medium".

Overall, initially, it appears that H_{EXP_0} is supported and that hypothesis H_{EXP_1} is not supported.

11.11 Summary of Initial Observations

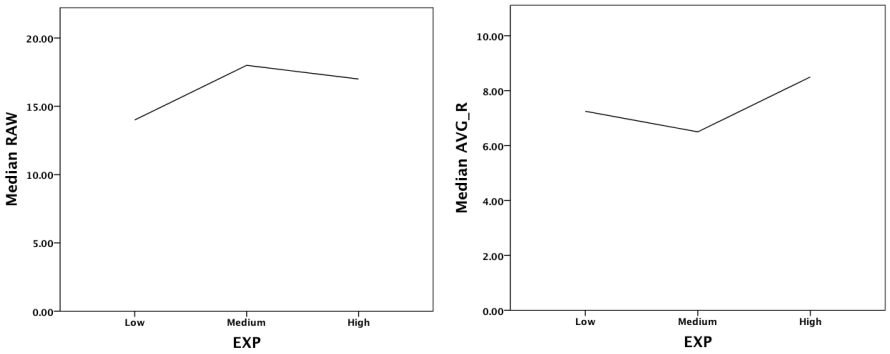
Table 96, at the end of the paper, summarizes the plots in Figures 10 through 29. The table is divided into two parts, each of which is what fits on one physical page. The legend explaining how to read the column headers is found at the bottom of Part II. A section of this table is the eight rows lying between two consecutive double horizontal lines. There is one section per independent variable. A subsection of this table is the two rows lying between two consecutive single horizontal lines, which are not the full width of the table. There is one subsection per dependent variable, each of whose value is the number of one kind of requirement ideas generated.

The eight rows of a section is about the independent variable, *IV*, that is displayed in the section's vertical middle in the column headed by "IV". For the independent variable, *IV*, of a section:

- The two rows of a subsection is about the two plots that plot against *IV*, the dependent variable, *DV*, that is displayed in the subsection's vertical middle in the column headed by "DV". For the independent variable, *IV*, and the dependent variable, *DV*, of a subsection:
  - According to the values displayed in the columns headed by "Fig#", "Filt'd?", "Corr?", and "WhenMax?", a row is about one plot that is shown in the indicated figure, which
    - plots against *IV*, the unfiltered or filtered values of *DV*,
    - shows whether or not this plot exhibits a correlation, and if so, what kind, and
    - indicates for which values of *IV* is the value of the unfiltered or filtered *DV* the highest.
  - The value that is displayed in the subsection's vertical middle in the column headed by "Diff?" is telling whether there is a real difference between the two plots of the subsection, one for unfiltered data and the other for filtered data.
- The values given the section's vertical middle in the columns headed by "H{IV}1", "H{IV}0", and "E1match?" assess whether the data given in the section's subsections support the hypotheses H_{IV_1} and H_{IV_0} about *IV*, and whether the support for these hypotheses matches the results reported for E1 for *IV* [37].

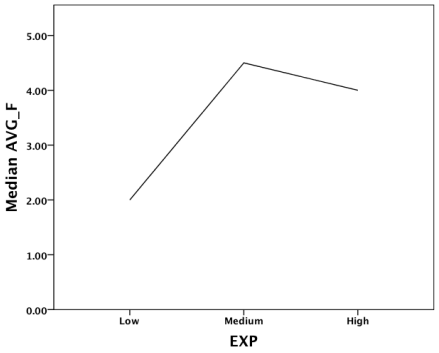
12 Statistical Analysis

This section presents a set of ANOVA and Kruskal-Wallis tests conducted on each of the independent variables and the two factors identified in Section 9.4 to test the hypotheses given in Section 10. Recall that each factor is considered an independent variable.

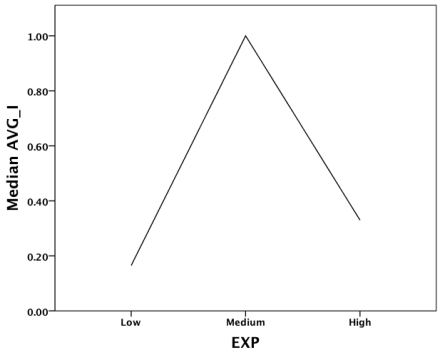


(a) RAW

(b) AVG_R

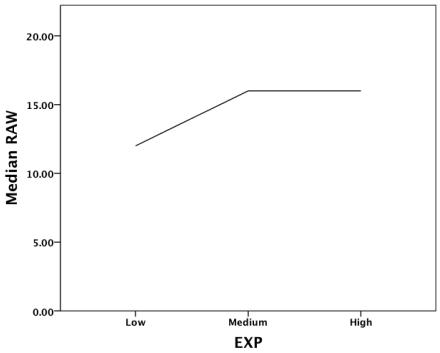


(c) AVG_F

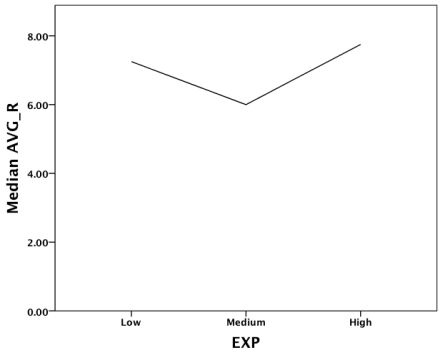


(d) AVG_I

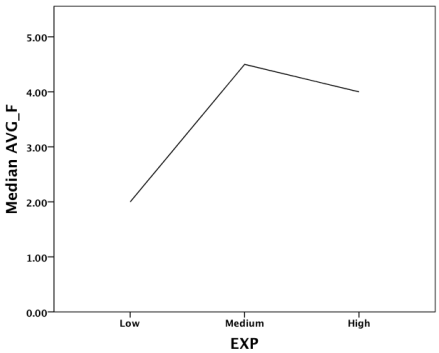
Fig. 28: Ideas vs. EXP (Unfiltered)



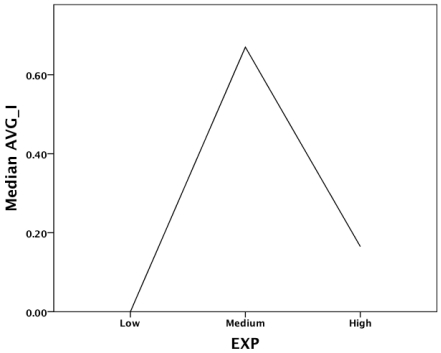
(a) RAW



(b) AVG_R



(c) AVG_F



(d) AVG_I

Fig. 29: Ideas vs. EXP (Filtered)

Each of the subsections of this section describes 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 [49] is conducted to compare all possible pairs of means of the dependent variable to show which means are significantly different from each other.
- 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 [11] is conducted to compare all possible pairs of medians of the dependent variable to show which medians are significantly different from each other.

Based on this plan, each subsection gives the following in short order with no 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 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.

However, as with the interpretations of the plots in Figures 10 through 29, the careful, precise natural language text for carrying out the plan is mind numbing. Therefore, tables summarizing what the text says are provided in the last subsection of this section, Section 12.12.

12.1 One-Way ANOVA on MIX

Table 15 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 16 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.

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

Table 15: Results of the Levene Test for MIX (Unfiltered)

Table 17 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 18 shows the results of the ANOVA test of the fil-

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

Table 16: Results of the Levene Test for MIX (Filtered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^{2e}	*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 17: Results of the One-Way ANOVA of the Effect of MIX (Unfiltered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*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

Table 18: Results of the One-Way ANOVA of the Effect of MIX (Filtered)

tered NRAW and NF plotted against MIX. The test shows no significant effect of the team's MIX on any of these variables.

Table 19 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 20 shows the results of the Kruskal-Wallis test of the effect of 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.

Dependent Variable	*p*
AVG_I	.555

Table 19: Results of the Kruskal-Wallis Test of the Effect of MIX (Unfiltered)

Dependent Variable	*p*
AVG_R	.697
AVG_I	.264

Table 20: Results of the Kruskal-Wallis Test of the Effect of MIX (Filtered)

12.2 One-Way ANOVA on CR

Table 21 shows that the Levene test result of the unfiltered dependent variables plotted against CR is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these unfiltered variables. Table 22 shows that the Levene test result of the filtered dependent variables plotted against CR 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.

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	.986	2	37	.383
NR	2.111	2	37	.136
NF	.824	2	37	.446
NI	2.289	2	37	.116

Table 21: Results of the Levene Test for CR (Unfiltered)

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	.636	2	33	.536
NR	4.463	2	35	.019
NF	2.432	2	36	.102
NI	.601	2	33	.554

Table 22: Results of the Levene Test for CR (Filtered)

Table 23 shows the results of the ANOVA test of the unfiltered dependent variables plotted against CR. The test shows no significant effect of the team's CR on any of these variables. Table 24 shows the results of the ANOVA test of the filtered NRAW, NF, and NI plotted against CR. The test shows no significant effect of the team's CR on any of these variables.

Table 25 shows the results of the Kruskal-Wallis test of the effect of a team's CR on the unfiltered AVG_I generated by the team. The test shows no significant effect of the team's CR on this variable. Table 26 shows the results of the Kruskal-Wallis test of the effect of a team's CR on the filtered AVG_R and AVG_I generated by the team. The test shows no significant effect of the team's CR on any of these variables.

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	1.342	2	.671	.692	.507	.036	.158
NR	2.058	2	1.029	1.091	.346	.056	.227
NF	1.831	2	.915	.980	.385	.050	.207
NI	3.089	2	1.544	1.980	.152	.097	.383

Table 23: Results of the One-Way ANOVA of the Effect of CR (Unfiltered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	2.414	3	.805	1.092	.366	.090	.268
NF	3.286	3	1.095	1.386	.263	.104	.336
NI	4.209	3	1.403	2.471	.079	.183	.561

Table 24: Results of the One-Way ANOVA of the Effect of CR (Filtered)

Dependent Variable	*p*
AVG_I	.102

Table 25: Results of the Kruskal-Wallis Test of the Effect of CR (Unfiltered)

Dependent Variable	*p*
AVG_R	.380
AVG_I	.060

Table 26: Results of the Kruskal-Wallis Test of the Effect of CR (Filtered)

12.3 One-Way ANOVA on REXP

Table 27 shows that the Levene test result of the unfiltered dependent variables plotted against REXP is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these unfiltered variables. Table 28 shows that the Levene test result of the filtered dependent variables plotted against REXP is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these filtered variables. Table 29 shows the results of the ANOVA test of the unfiltered dependent variables plotted against REXP. The test shows no significant effect of the team's REXP on NRAW, NF, and NI, but shows a significant effect of the team's REXP on NR. Table 30 shows the results of the ANOVA test of the filtered dependent variables plotted against REXP. The test shows no significant effect of the team's REXP on any of these variables.

Table 31 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team's REXP on the unfiltered NR generated by the team. The test shows

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	.167	3	36	.918
NR	.210	3	36	.888
NF	1.208	3	36	.321
NI	1.850	3	36	.156

Table 27: Results of the Levene Test for REXP (Unfiltered)

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	1.759	3	32	.175
NR	.662	3	34	.581
NF	1.568	3	35	.215
NI	2.095	3	32	.120

Table 28: Results of the Levene Test for REXP (Filtered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	2.601	3	.867	.903	.449	.070	.228
NR	7.769	3	2.590	3.195	.035	.210	.689
NF	2.778	3	.926	.992	.408	.076	.247
NI	1.040	3	.347	.404	.751	.033	.122

Table 29: Results of the One-Way ANOVA of the Effect of REXP (Unfiltered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	1.732	4	.433	.554	.698	.065	.166
NR	5.473	4	1.368	1.956	.124	.187	.528
NF	1.673	4	.418	.487	.745	.053	.151
NI	.691	4	.173	.248	.909	.030	.097

Table 30: Results of the One-Way ANOVA of the Effect of REXP (Filtered)

that the difference between the means of the NR of the teams is significant when REXP = “Medium” is paired with REXP = “High”.

Sample 1	*Sample 2*	*Mean Difference*	*Std. Error*	*p*
None	Low	.236	.52	.969
	Medium	.935	.52	.291
	High	-.159	.52	.99
Low	Medium	.699	.368	.245
	High	-.395	.368	.708
Medium	High	-1.094	.368	.026

Table 31: Results of the Tukey HSD Pairwise Comparison Test of the Effect of REXP on NR (Unfiltered)

Tables 32 and 33 show the results of Kruskal-Wallis tests of the effect of a team's REXP on the unfiltered and filtered AVG_I generated by the team, respectively. The tests indicate no significant effect of the team's REXP on any of these variables.

Dependent Variable	***p***
AVG_I	.782

Table 32: Results of the Kruskal-Wallis Test of the Effect of REXP (Unfiltered)

Dependent Variable	***p***
AVG_I	.948

Table 33: Results of the Kruskal-Wallis Test of the Effect of REXP (Filtered)

12.4 One-Way ANOVA on IREXP

Table 34 shows that the Levene test result of the unfiltered dependent variables plotted against IREXP is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these unfiltered variables. Table 35 shows that the Levene test result of the filtered dependent variables plotted against IREXP is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these filtered variables.

Dependent Variable	***Levene Statistic***	***df1***	***df2***	***p***
NRAW	.401	3	36	.753
NR	.441	3	36	.725
NF	.793	3	36	.506
NI	1.469	3	36	.239

Table 34: Results of the Levene Test for IREXP (Unfiltered)

Dependent Variable	***Levene Statistic***	***df1***	***df2***	***p***
NRAW	.604	3	32	.617
NR	1.325	3	34	.282
NF	.857	3	35	.473
NI	1.108	3	32	.360

Table 35: Results of the Levene Test for IREXP (Filtered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	.878	3	.293	.290	.832	.024	.100
NR	4.455	3	1.485	1.645	.196	.121	.394
NF	1.845	3	.615	.641	.594	.051	.171
NI	1.688	3	.563	.669	.576	.053	.177

Table 36: Results of the One-Way ANOVA of the Effect of IREXP (Unfiltered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	1.937	4	.484	.625	.648	.072	.183
NR	1.641	4	.410	.505	.732	.056	.155
NF	2.775	4	.694	.838	.510	.087	.240
NI	5.290	4	1.323	2.397	.071	.231	.622

Table 37: Results of the One-Way ANOVA of the Effect of IREXP (Filtered)

Table 36 shows the results of the ANOVA test of the unfiltered dependent variables plotted against IREXP. The test shows no significant effect of the team's IREXP on any of these variables. Table 37 shows the results of the ANOVA test of the filtered dependent variables plotted against IREXP. The test shows no significant effect of the team's IREXP on any of these variables.

Tables 38 and 39 show the results of Kruskal-Wallis tests of the effect of a team's IREXP on the unfiltered and filtered AVG_I generated by the team, respectively. The tests indicate no significant effect of the team's IREXP on any of these variables.

Dependent Variable	*p*
AVG_I	.449

Table 38: Results of the Kruskal-Wallis Test of the Effect of IREXP (Unfiltered)

Dependent Variable	*p*
AVG_I	.060

Table 39: Results of the Kruskal-Wallis Test of the Effect of IREXP (Filtered)

12.5 One-Way ANOVA on IEXP

Table 40 shows that the Levene test result of the unfiltered dependent variables plotted against IEXP is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these unfiltered variables. Table 41 shows that

the Levene test result of the filtered dependent variables plotted against IEXP is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these filtered variables.

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	1.030	3	36	.391
NR	.525	3	36	.668
NF	.906	3	36	.448
NI	.435	3	36	.729

Table 40: Results of the Levene Test for IEXP (Unfiltered)

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	.802	3	32	.502
NR	.657	3	34	.584
NF	.678	3	35	.571
NI	.188	3	32	.904

Table 41: Results of the Levene Test for IEXP (Filtered)

Table 42 shows the results of the ANOVA test of the unfiltered dependent variables plotted against IEXP. The test shows no significant effect of the team's IEXP on any of these variables. Table 43 shows the results of the ANOVA test of the filtered dependent variables plotted against IEXP. The test shows no significant effect of the team's IEXP on any of these variables.

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	2.539	3	.846	.879	.461	.068	.222
NR	1.921	3	.640	.658	.583	.052	.174
NF	6.726	3	2.242	2.721	.059	.185	.611
NI	1.760	3	.587	.699	.559	.055	.183

Table 42: Results of the One-Way ANOVA of the Effect of IEXP (Unfiltered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	2.961	4	.74	.996	.424	.111	.278
NR	1.039	4	.26	.313	.867	.036	.111
NF	6.592	4	1.648	2.294	.079	.208	.606
NI	2.186	4	.546	.842	.509	.095	.238

Table 43: Results of the One-Way ANOVA of the Effect of IEXP (Filtered)

Tables 44 and 45 show the results of Kruskal-Wallis tests of the effect of a team's IEXP on the unfiltered and filtered AVG_I generated by the team, respectively. The tests indicate no significant effect of the team's IEXP on any of these variables.

Dependent Variable	***p***
AVG_I	.564

Table 44: Results of the Kruskal-Wallis Test of the Effect of IEXP (Unfiltered)

Dependent Variable	***p***
AVG_I	.504

Table 45: Results of the Kruskal-Wallis Test of the Effect of IEXP (Filtered)

12.6 One-Way ANOVA on NCS

Table 46 shows that the Levene test result of the unfiltered dependent variables plotted against NCS is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these unfiltered variables. Table 47 shows that the Levene test result of the filtered dependent variables plotted against NCS is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these filtered variables.

Dependent Variable	***Levene Statistic***	***df1***	***df2***	***p***
NRAW	.065	3	36	.978
NR	.499	3	36	.685
NF	1.053	3	36	.381
NI	1.433	3	36	.249

Table 46: Results of the Levene Test for NCS (Unfiltered)

Dependent Variable	***Levene Statistic***	***df1***	***df2***	***p***
NRAW	.385	3	32	.765
NR	1.294	3	34	.292
NF	.646	3	35	.591
NI	.235	3	32	.871

Table 47: Results of the Levene Test for NCS (Filtered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	2.976	3	.992	1.044	.385	.080	.259
NR	2.818	3	.939	.991	.408	.076	.247
NF	5.230	3	1.743	2.015	.129	.144	.474
NI	5.615	3	1.872	2.558	.070	.176	.582

Table 48: Results of the One-Way ANOVA of the Effect of NCS (Unfiltered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	5.738	4	1.434	2.186	.093	.215	.577
NR	1.581	4	.395	.486	.746	.054	.151
NF	7.833	4	1.958	2.867	.037	.247	.717
NI	7.607	4	1.902	3.968	.010	.332	.858

Table 49: Results of the One-Way ANOVA of the Effect of NCS (Filtered)

Table 48 shows the results of the ANOVA test of the unfiltered dependent variables plotted against NCS. The test shows no significant effect of the team's NCS on any of these variables. Table 49 shows the results of the ANOVA test of the filtered dependent variables plotted against NCS. The test shows no significant effect of the team's NCS on each of NRAW and NR, but shows a significant effect of the team's NCS on each of NF and NI.

Table 50 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team's NCS on the filtered NF generated by the team. The test shows that the difference between the means of the NF of the teams is rather significant when NCS = 0 is paired with NCS = 3 and when NCS = 1 is paired with NCS = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Std. Error*	*p*
0	1	-.138	.460	.990
	2	-.970	.477	.195
	3	-1.027	.385	.053
1	2	-.832	.460	.286
	3	-.889	.363	.086
2	3	-.0572	.385	.999

Table 50: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NCS on NF (Filtered)

Table 51 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team's NCS on the filtered NI generated by the team. The test shows that the difference between the means of the NI of the teams is rather significant when NCS = 0 is paired with NCS = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Std. Error*	*p*
0	1	-.0238	.406	1.00
	2	-.817	.438	.263
	3	-.976	.348	.040
1	2	-.793	.405	.226
	3	-.952	.306	.019
2	3	-.159	.348	.968

Table 51: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NCS on NI (Filtered)

Table 52 shows the results of the Kruskal-Wallis test of the effect of a team's NCS on the unfiltered AVG_I generated by the team. The test shows no significant effect of the team's NCS on this variable. Table 53 shows the results of the Kruskal-Wallis test of the effect of a team's NCS on the filtered AVG_I generated by the team. The test shows a significant effect of the team's NCS on this variable.

Dependent Variable	*p*
AVG_I	.052

Table 52: Results of the Kruskal-Wallis Test of the Effect of NCS (Unfiltered)

Dependent Variable	*p*
AVG_I	.010

Table 53: Results of the Kruskal-Wallis Test of the Effect of NCS (Filtered)

Table 54 shows the results of the Dunn-Bonferroni Pairwise Comparison Test of the effect of a team's NCS 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 NCS = 0 is paired with NCS = 3.

Sample 1	*Sample 2*	*Test Statistic*	*Std. Error*	*Std. Test Statistic*	*p*[*]
3	2	3.292	5.038	.653	1.000
	1	10.375	5.884	1.763	.467
	0	12.325	4.243	2.905	.022
2	1	7.083	6.794	1.043	1.000
	0	9.033	5.435	1.662	.579
1	0	1.950	6.227	.313	1.000

[*] Adjusted by Bonferroni correction method.

Table 54: Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of NCS on AVG_I (Filtered)

12.7 One-Way ANOVA on NSE

Table 55 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 56 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.

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 55: Results of the Levene Test for NSE (Unfiltered)

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

Table 56: Results of the Levene Test for NSE (Filtered)

Table 57 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.

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* 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 57: Results of the One-Way ANOVA of the Effect of NSE (Unfiltered)

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

Table 58: Results of the One-Way ANOVA of the Effect of NSE (Filtered)

Table 58 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 59 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.

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 59: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NSE on NF (Unfiltered)

Table 60 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.

Sample 1	*Sample 2*	*Mean Difference*	*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

Table 60: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NSE on NF (Filtered)

Table 61 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 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.

Sample 1	*Sample 2*	*Mean Difference*	*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 61: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NSE on NI (Filtered)

Table 62 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 63 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.

Dependent Variable	*p*
AVG_I	.069

Table 62: Results of the Kruskal-Wallis Test of the Effect of NSE (Unfiltered)

Dependent Variable	*p*
AVG_R	.538
AVG_I	.005

Table 63: Results of the Kruskal-Wallis Test of the Effect of NSE (Filtered)

Table 64 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.

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

Table 64: Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of NSE on AVG_I (Filtered)

12.8 One-Way ANOVA on NGRAD

Table 65 shows that the Levene test result of the unfiltered dependent variables plotted against NGRAD is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these unfiltered variables. Table 66 shows that the Levene test result of the filtered dependent variables plotted against NGRAD is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these filtered variables.

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	.257	3	36	.856
NR	1.468	3	36	.240
NF	2.678	3	36	.062
NI	.604	3	36	.617

Table 65: Results of the Levene Test for NGRAD (Unfiltered)

Table 67 shows the results of the ANOVA test of the unfiltered dependent variables plotted against NGRAD. The test shows no significant effect of the team's NGRAD on any of these variables. Table 68 shows the results of the ANOVA test

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	.508	3	32	.680
NR	2.148	3	34	.112
NF	1.826	3	35	.160
NI	.401	3	32	.753

Table 66: Results of the Levene Test for NGRAD (Filtered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	4.311	3	1.437	1.574	.213	.116	.379
NR	.573	3	.191	.189	.903	.016	.082
NF	6.614	3	2.205	2.666	.062	.182	.602
NI	4.190	3	1.397	1.811	.163	.131	.431

Table 67: Results of the One-Way ANOVA of the Effect of NGRAD (Unfiltered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	6.726	4	1.682	2.689	.049	.252	.679
NR	.643	4	.161	.191	.941	.022	.086
NF	8.191	4	2.048	3.044	.03	.258	.747
NI	7.206	4	1.802	3.663	.014	.314	.825

Table 68: Results of the One-Way ANOVA of the Effect of NGRAD (Filtered)

of the filtered dependent variables plotted against NGRAD. The test shows no significant effect of the team's NGRAD on NR, but shows a significant effect on each of NRAW, NF, and NI.

Table 69 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team's NGRAD on the filtered NRAW generated by the team. The test shows that the difference between the means of the NRAW of the teams is rather significant when NGRAD = 0 is paired with NGRAD = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Std. Error*	*p*
0	1	.136	.468	.991
	2	.665	.408	.378
	3	.894	.319	.040
1	2	.528	.510	.730
	3	.758	.442	.333
2	3	.229	.379	.929

Table 69: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NGRAD on NRAW (Filtered)

Table 70 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team's NGRAD on the filtered NF generated by the team. The test shows that the difference between the means of the NF of the teams is rather significant when NGRAD = 0 is paired with NGRAD = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Std. Error*	*p*
0	1	.0831	.47889	.998
	2	.956	.397	.094
	3	.957	.317	.023
1	2	.873	.514	.340
	3	.874	.456	.239
2	3	.001	.368	1.000

Table 70: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NGRAD on NF (Filtered)

Table 71 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team's NGRAD on the filtered NI generated by the team. The test shows that the difference between the means of the NI of the teams is rather significant when NGRAD = 0 is paired with NGRAD = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Std. Error*	*p*
0	1	.078	.457	.998
	2	.576	.339	.341
	3	.969	.278	.008
1	2	.498	.483	.733
	3	.891	.443	.206
2	3	.393	.321	.617

Table 71: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NGRAD on NI (Filtered)

Table 72 shows the results of the Kruskal-Wallis test of the effect of a team's NGRAD on the unfiltered AVG_I generated by the team. The test shows no significant effect of the team's NGRAD on this variable. Table 73 shows the results of the Kruskal-Wallis test of the effect of a team's NGRAD on the filtered AVG_I generated by the team. The test shows a significant effect of the team's NGRAD on this variable.

Dependent Variable	*p*
AVG_I	.119

Table 72: Results of the Kruskal-Wallis Test of the Effect of NGRAD (Unfiltered)

Dependent Variable	*p*
AVG_I	.018

Table 73: Results of the Kruskal-Wallis Test of the Effect of NGRAD (Filtered)

Table 74 shows the results of the Dunn-Bonferroni Pairwise Comparison Test of the effect of a team's NGRAD 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 NGRAD = 0 is paired with NGRAD = 3.

Sample 1	*Sample 2*	*Test Statistic*	*Std. Error*	*Std. Test Statistic*	*p*
3	2	5.481	4.696	1.167	1.000
	1	12.433	6.489	1.916	.332
	0	11.994	4.073	2.945	.019
2	0	6.513	4.960	1.313	1.000
	1	6.952	7.080	.982	1.000
1	0	-.439	6.682	-.066	1.000

Table 74: Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of NGRAD on AVG_I (Filtered)

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

Table 75 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 76 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.

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 75: Results of the Levene Test for MIX, EXP, and EDU (Unfiltered)

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

Table 76: Results of the Levene Test for MIX, EXP, and EDU (Filtered)

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 77 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 78 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.

*12.9.1 MIX * EXP * EDU (Unfiltered)*

Figure 30 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.

Source	*Dependent Variable*	*Sum of Squares*	*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 77: Results of the Three-Way ANOVA of the Effect of MIX, EXP, and EDU (Unfiltered)

An issue with the sub-plots of Figure 30 is that there are not enough data points to show the interactions between all values of the affecting independent variables. Also, comparing Figure 30(a) with Figure 30(b) and Figure 30(c) with Figure 30(d), the correlations seem to be contradictory for EXP = “Low” and EXP = “High”. All in all, the plots do not show anything interesting.

One possible explanation for the interactions shown in Figure 30 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.

*12.9.2 EXP * EDU (Filtered)*

Figure 31 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

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	.039	.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	.

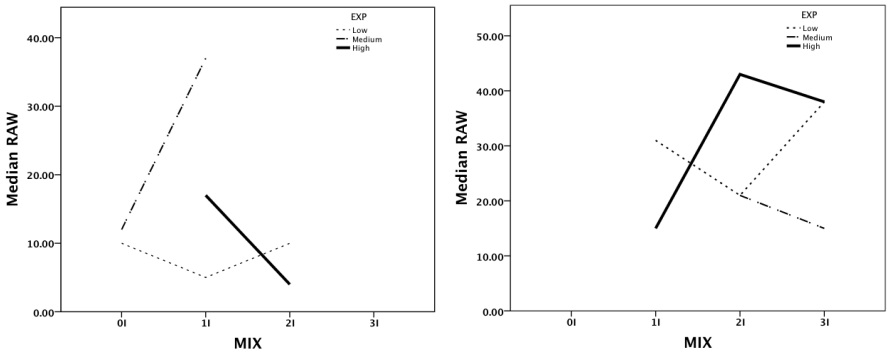
[a] When the number of data points needed to calculate the effect of a variable or interactions of some variables is not enough, SPSS outputs a value of 0 for sum of squares and degrees of freedom and "." for the other fields.

Table 78: Results of the Three-Way ANOVA of the Effect of MIX, EXP, and EDU (Filtered)

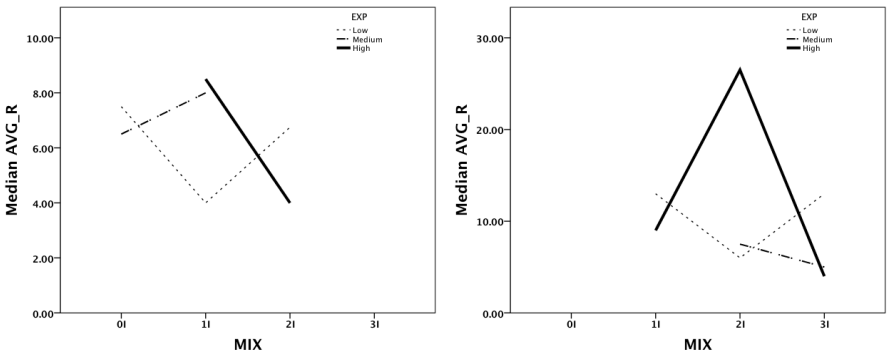
generated by teams with EDU = "Low" is positively correlated with the teams' EXP. On the other hand, the medians of the filtered RAW generated by teams' with EDU = "High" is negatively correlated with the teams' EXP.

12.10 One-Way ANOVA on EDU

Table 79 shows that the Levene test result of the unfiltered dependent variables plotted against EDU is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these unfiltered variables. Table 80 shows that the Levene test result of the filtered dependent variables plotted against EDU is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these filtered variables.



(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. 30: Ideas vs. MIX * EXP * EDU (Unfiltered)

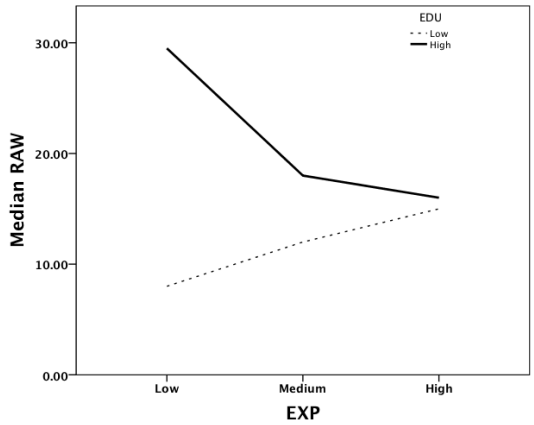


Fig. 31: RAW vs. EXP * EDU (Filtered)

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	.004	1	38	.951
NR	1.053	1	38	.311
NF	1.213	1	38	.278
NI	1.422	1	38	.240

Table 79: Results of the Levene Test for EDU (Unfiltered)

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	.445	1	34	.509
NR	2.004	1	36	.166
NF	.606	1	37	.441
NI	.026	1	34	.872

Table 80: Results of the Levene Test for EDU (Filtered)

Table 81 shows the results of the ANOVA test of the unfiltered dependent variables plotted against EDU. The test shows no significant effect of the team's EDU on NR but shows a significant effect of the team's EDU on each of NRAW, NF, and NI. Table 82 shows the results of the ANOVA test of the filtered dependent variables plotted against EDU. The test shows no significant effect of the team's EDU on NR but shows a significant effect of the team's EDU on each of NRAW, NF, and NI. Since EDU has only two values, no Tukey HSD Pairwise Comparison Test was performed, as it would return the same results as the one-way ANOVA.

Table 83 shows the results of the Kruskal-Wallis test of the effect of a team's EDU on the unfiltered AVG_I generated by the team. The test indicates a significant effect of the team's EDU on this variable. Table 84 shows the results of the Kruskal-Wallis

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	3.944	1	3.944	4.509	.040	.106	.544
NR	.620	1	.620	.648	.426	.017	.123
NF	10.621	1	10.621	15.665	.000	.292	.971
NI	4.828	1	4.828	6.763	.013	.151	.717

Table 81: Results of the One-Way ANOVA of the Effect of EDU (Unfiltered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	4.178	1	4.178	6.610	.015	.163	.705
NR	1.106	2	.553	.707	.500	.038	.160
NF	13.305	2	6.652	13.354	.000	.419	.996
NI	8.551	2	4.275	10.100	.000	.373	.977

Table 82: Results of the One-Way ANOVA of the Effect of EDU (Filtered)

test of the effect of a team's EDU on the filtered AVG_I generated by the team. The test indicates a significant effect of the team's EDU on this variable.

Dependent Variable	*p*
AVG_I	.008

Table 83: Results of the Kruskal-Wallis Test of the Effect of EDU (Unfiltered)

Dependent Variable	*p*
AVG_I	.000

Table 84: Results of the Kruskal-Wallis Test of the Effect of EDU (Filtered)

Since EDU has only two values, no Dunn-Bonferroni Pairwise Comparison Test was performed, as it would return the same results as the one-way ANOVA.

12.11 One-Way ANOVA on EXP

Table 85 shows that the Levene test result of the unfiltered dependent variables plotted against EXP is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these unfiltered variables. Table 86 shows that the Levene test result of the filtered dependent variables plotted against EXP is not significant for each of the four dependent variables. Thus, an ANOVA is applicable to each of these filtered variables.

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	.192	2	37	.826
NR	.892	2	37	.419
NF	.052	2	37	.949
NI	.274	2	37	.762

Table 85: Results of the Levene Test for EXP (Unfiltered)

Table 87 shows the results of the ANOVA test of the unfiltered dependent variables plotted against EXP. The test shows no significant effect of the team's EXP on each of NRAW, NR, and NF but shows a significant effect of the team's EXP on NI. Table 88 shows the results of the ANOVA test of the filtered dependent variables plotted against EXP. The test shows no significant effect of the team's EXP on each of NRAW, NR, and NF but shows a significant effect of the team's EXP on NI.

Table 89 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team's EXP on the unfiltered NI generated by the team. The test shows

Dependent Variable	*Levene Statistic*	*df1*	*df2*	*p*
NRAW	.177	2	33	.838
NR	.414	2	35	.664
NF	.250	2	36	.780
NI	.346	2	33	.710

Table 86: Results of the Levene Test for EXP (Filtered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	.618	2	.309	.312	.734	.017	.096
NR	.867	2	.433	.444	.645	.023	.117
NF	2.319	2	1.160	1.259	.296	.064	.257
NI	6.830	2	3.415	5.029	.012	.214	.783

Table 87: Results of the One-Way ANOVA of the Effect of EXP (Unfiltered)

Dependent Variable	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*	*Partial* η^2	*Observed Power*
NRAW	1.470	3	.490	.640	.595	.055	.169
NR	1.492	3	.497	.627	.602	.051	.167
NF	1.678	3	.559	.670	.576	.053	.177
NI	6.632	3	2.211	4.472	.010	.289	.837

Table 88: Results of the One-Way ANOVA of the Effect of EXP (Filtered)

that the difference is rather significant between the means of the NI of the teams with EXP = “Low” and EXP = “Medium”.

Sample 1	*Sample 2*	*Mean Difference*	*Std. Error*	*p*
Low	Medium	-.875	.297	.015
	High	-.111	.352	.947
Medium	High	.764	.340	.076

Table 89: Results of the Tukey HSD Pairwise Comparison Test of the Effect of EXP on NI (Unfiltered)

Table 90 shows the results of the Tukey HSD Pairwise Comparison Test of the effect of a team’s EXP on the filtered NI generated by the team. The test shows that the difference between the means of the NI of the teams is rather significant when EXP = “Low” is paired with EXP = “Medium” and when EXP = “Medium” is paired with EXP = “High”.

Sample 1	*Sample 2*	*Mean Difference*	*Std. Error*	*p*
Low	Medium	-.854	.266	.008
	High	-.068	.316	.975
Medium	High	.787	.308	.040

Table 90: Results of the Tukey HSD Pairwise Comparison Test of the Effect of EXP on NI (Filtered)

Table 91 shows the results of the Kruskal-Wallis test of the effect of a team's EXP on the unfiltered AVG_I generated by the team. The test shows a significant effect of the team's EXP on this variable. Table 92 shows the results of the Kruskal-Wallis test of the effect of a team's EXP on the filtered AVG_I generated by the team. The test shows a significant effect of the team's EXP on this variable.

Dependent Variable	*p*
AVG_I	.019

Table 91: Results of the Kruskal-Wallis Test of the Effect of EXP (Unfiltered)

Dependent Variable	*p*
AVG_I	.013

Table 92: Results of the Kruskal-Wallis Test of the Effect of EXP (Filtered)

Table 93 shows the results of the Dunn-Bonferroni Pairwise Comparison Test of the effect of a team's EXP on the unfiltered 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 EXP = "Low" is paired with EXP = "Medium". Table 94 shows the results of the Dunn-Bonferroni Pairwise Comparison Test of the effect of a team's EXP 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 EXP = "Low" is paired with EXP = "Medium".

Sample 1 - Sample 2	*Test Statistic*	*Std. Error*	*Std. Test Statistic*	*p*
Low-High	-1.667	4.899	-.340	1.00
Low-Medium	-10.882	4.139	-2.629	.026
Medium-High	9.216	4.727	1.950	.154

Table 93: Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of EXP on AVG_I (Unfiltered)

Sample 1 - Sample 2	*Test Statistic*	*Std. Error*	*Std. Test Statistic*	*p*
Low-High	-1.03	4.61	-.22	1.00
Low-Medium	-10.62	3.89	-2.73	.019
Medium-High	9.60	4.49	2.14	.098

Table 94: Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of EXP on AVG_I (Filtered)

12.12 Summary of Statistical Analyses

Tables 97 and 98, at the end of the paper, summarize the statistical results. Table 97 summarizes the one-way ANOVAs and Table 98 summarizes the three-way ANOVA. Table 98 should actually be a section of Table 97, but the tables require different headers; so it's easier to make Table 98 a separate table, while marking the place in Table 97 in which 98's data would appear.

Table 97 is divided into two 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 II. A section of this table is the 10 rows lying between two consecutive double horizontal lines. There is one section per independent variable. 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 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 a section is about the independent variable, IV, that is displayed in the section's vertical middle 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 in either unfiltered or filtered as indicated by the value in the the subsection's vertical middle in the column headed by "Filt'd?" in the lower header row. For the independent variable, IV, and the dependent variables of a particular filtration, $UorF$, of a subsection:
  - The first row row of the subsection gives in the columns headed by "LeveneT#", "ANOVA_T#", and "K-W_T#" in the upper header row, the numbers of the tables in which results can be found of the Levene test, the ANOVA, and the Kruskal-Wallis test for IV and the dependent variables of filtration $UorF$.
  - 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 98 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 is the 10 rows lying beneath the double horizontal lines. 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, (IV_1, IV_2, IV_3), of independent variables that is displayed stacked vertically in the section's vertical middle in the column headed by "IV". For the triple, (IV_1, IV_2, IV_3), of independent variables of the section:

- The five rows of either subsection of the section for (IV_1, IV_2, IV_3) is about the relationship between (IV_1, IV_2, IV_3) and four dependent variables, which are in either unfiltered or filtered as indicated by the value in the the subsection's vertical middle in the column headed by "Filt'd?" in the lower header row. For the triple of independent variables, (IV_1, IV_2, IV_3), and the dependent variables of a particular filtration, $UorF$ of a subsection:
  - The first row row of the subsection gives in the columns headed by "LeveneT#" 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 (IV_1, IV_2, IV_3) and the normalized dependent variables of filtration $UorF$.
  - Each of the four other rows of the subsections is about the relationship between (IV_1, IV_2, IV_3) 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 97 because of the various tests done, only the ANOVA, which needs normalized variables, works in the three-way mode.

13 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. However, 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.

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 [52]. The first author's PhD thesis [36] addresses all known 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.

13.1 Threats to Conclusion Validity

Conclusion validity addresses whether the conclusions about the hypotheses follow from the results of the experiment [17]. 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 [36] 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 evaluations of ideas, at least two and in some cases, three, evaluators were used. Moreover, statistical tests were used to show that there was high agreement among the evaluators.

13.2 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 evaluation procedure each time.

Nevertheless, because there were two distinct collections of runs in two experiments E1 and E2, and each experiment had its own evaluation, there is a chance that the evaluations of ideas in the two experiments might be different from what they would be if there had been only *one* evaluation of *all* the ideas at once. Section 8 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).

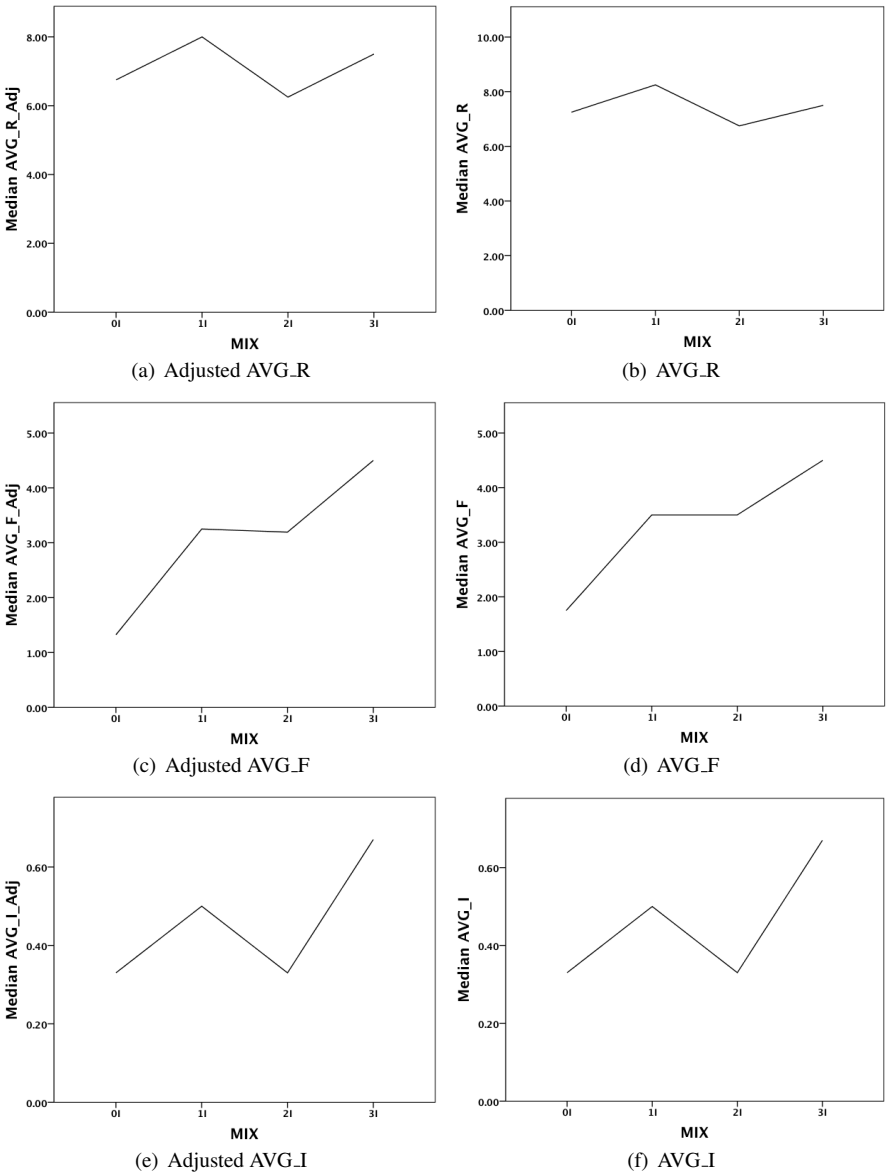
Recall the plots that were examined to make the initial observations in Section 11. These plots can be redone with the adjusted data. When the plots of adjusted data are compared with the corresponding plots of the original data, it is clear that none of the correlations observed in Section 11 have changed to the point that the conclusions drawn in Section 14 would have to be changed. Figures 32 through 47 show each of the plots from Figures 10 through 25 to the right of the corresponding plot of the adjusted data. Specifically, these plots show that the 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.

The new plots that are based on the adjusted data for teams in E2 show that the preliminary conclusions are unchanged. Therefore, it is unlikely that the more detailed analysis would show any 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.

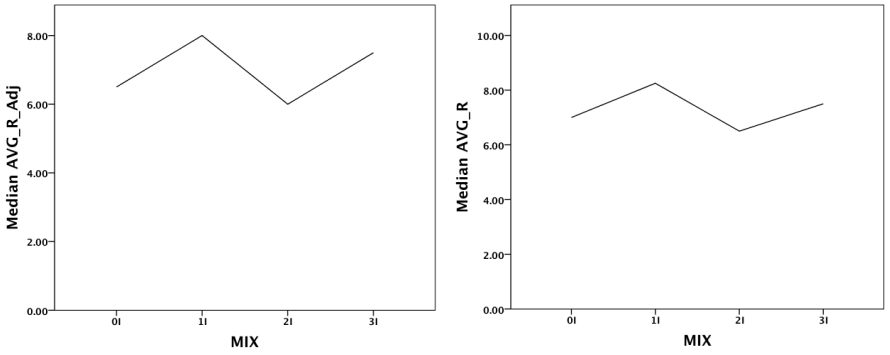


(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

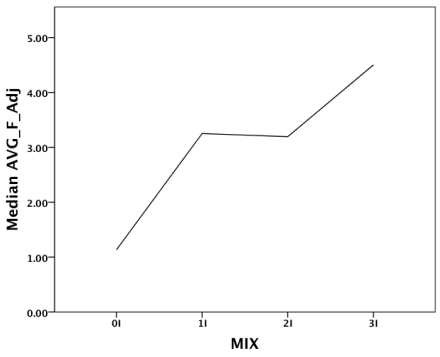
(e) Adjusted AVG_I (f) AVG_I

Fig. 32: Adjusted Ideas vs. MIX – Ideas vs. MIX (Unfiltered)

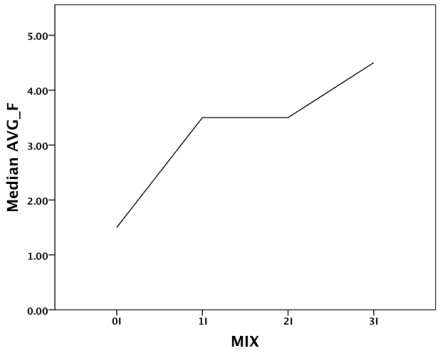


(a) Adjusted AVG_R

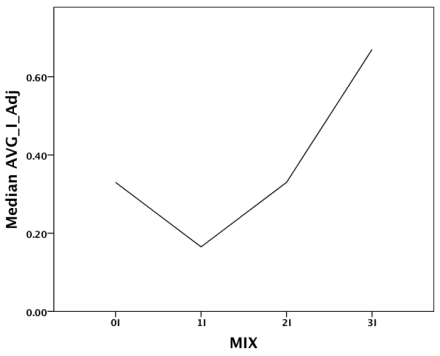
(b) AVG_R



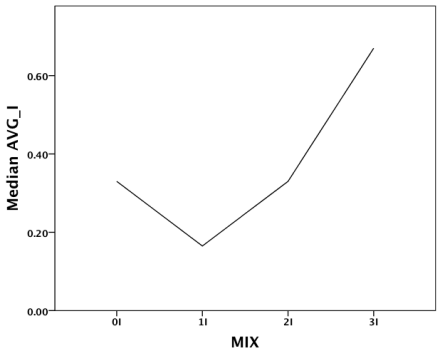
(c) Adjusted AVG_F



(d) AVG_F



(e) Adjusted AVG_I

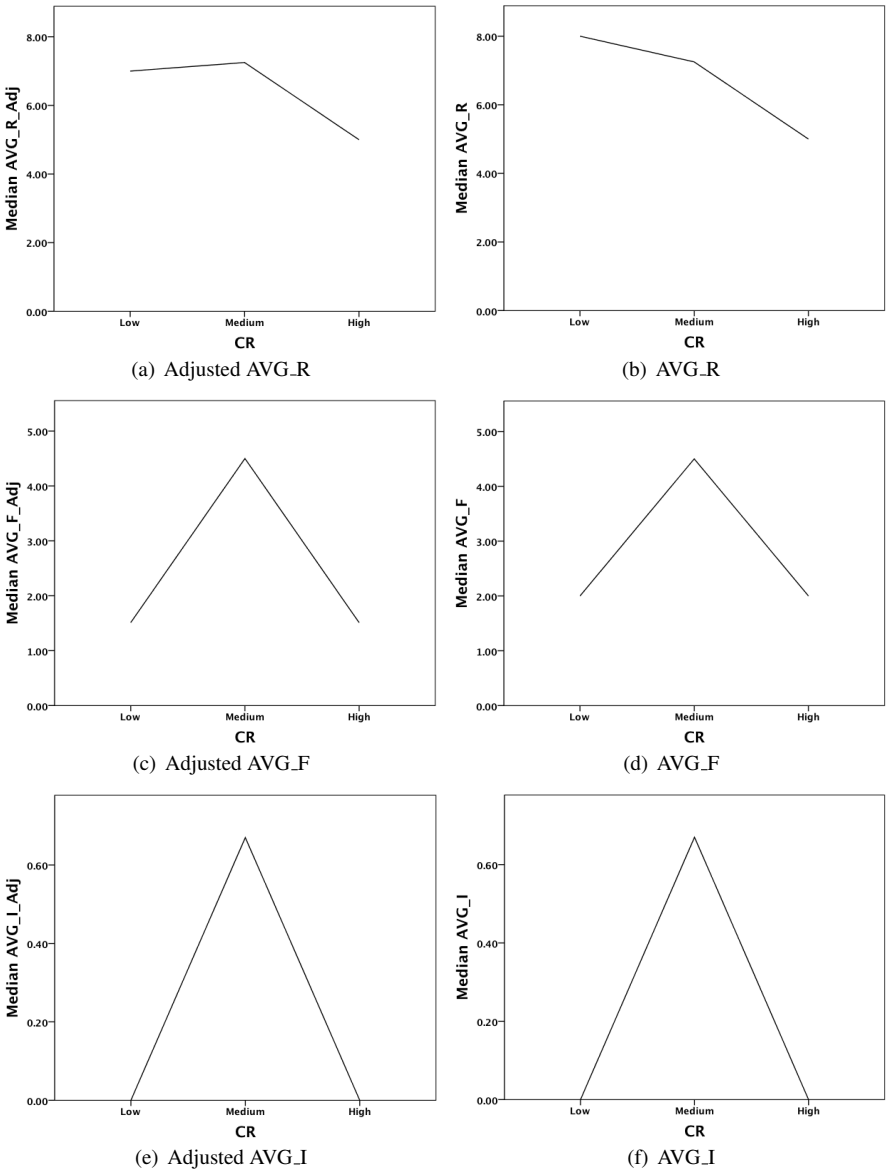


(f) AVG_I

Fig. 33: Adjusted Ideas vs. MIX – Ideas vs. MIX (Filtered)

13.3 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



(a) Adjusted AVG_R (b) AVG_R

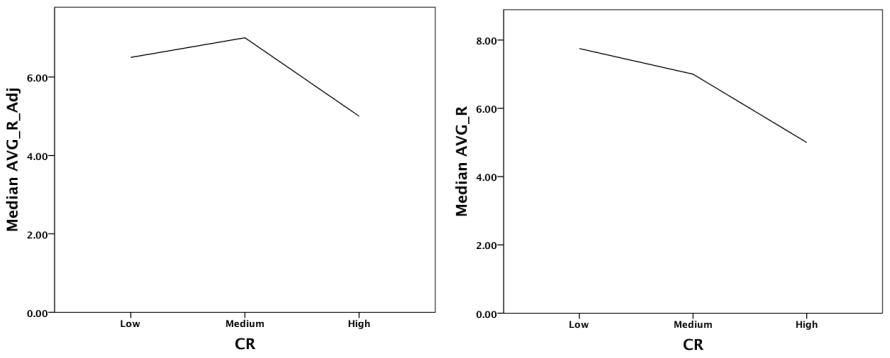
(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

Fig. 34: Adjusted Ideas vs. CR – Ideas vs. CR (Unfiltered)

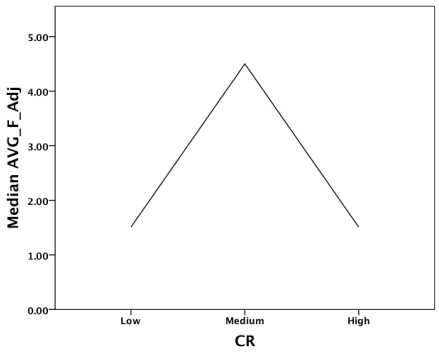
results imply what they are intended to imply. The construct validity threats present in these experiments are

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,

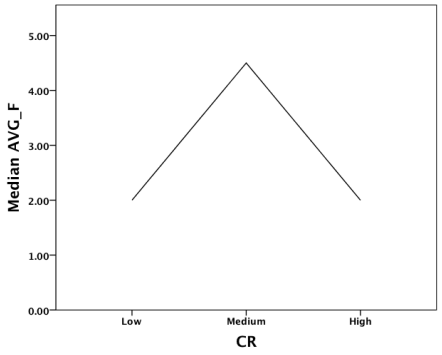


(a) Adjusted AVG_R

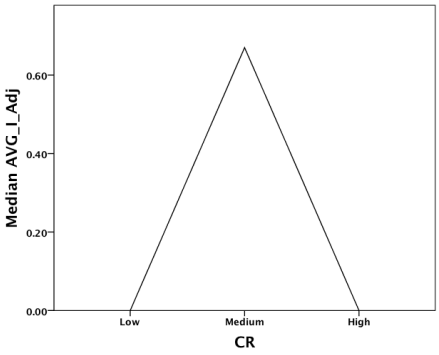
(b) AVG_R



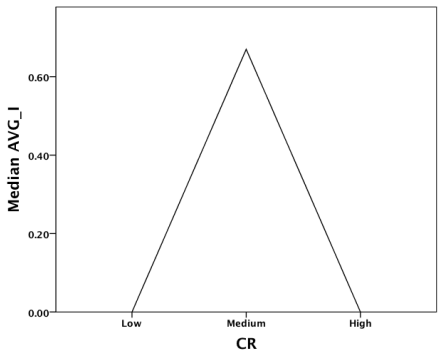
(c) Adjusted AVG_F



(d) AVG_F



(e) Adjusted AVG_I

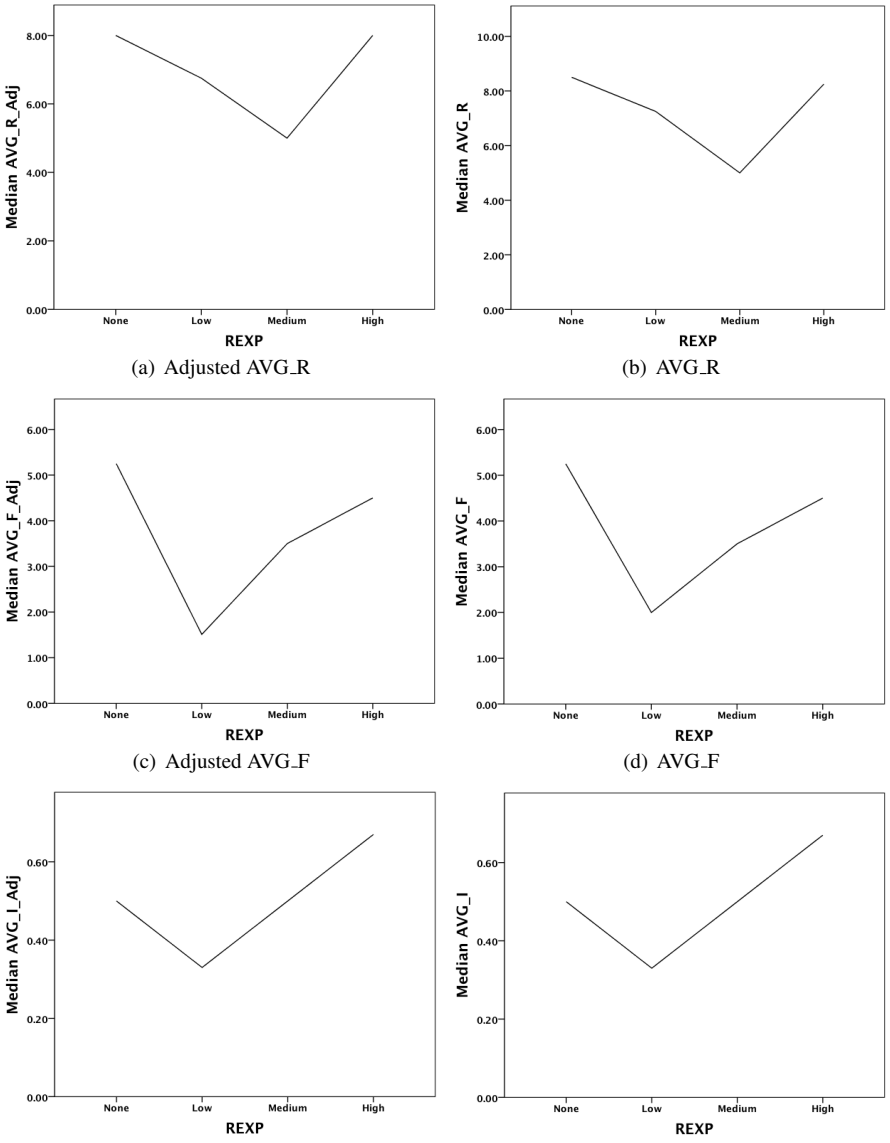


(f) AVG_I

Fig. 35: Adjusted Ideas vs. CR – Ideas vs. CR (Filtered)

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

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



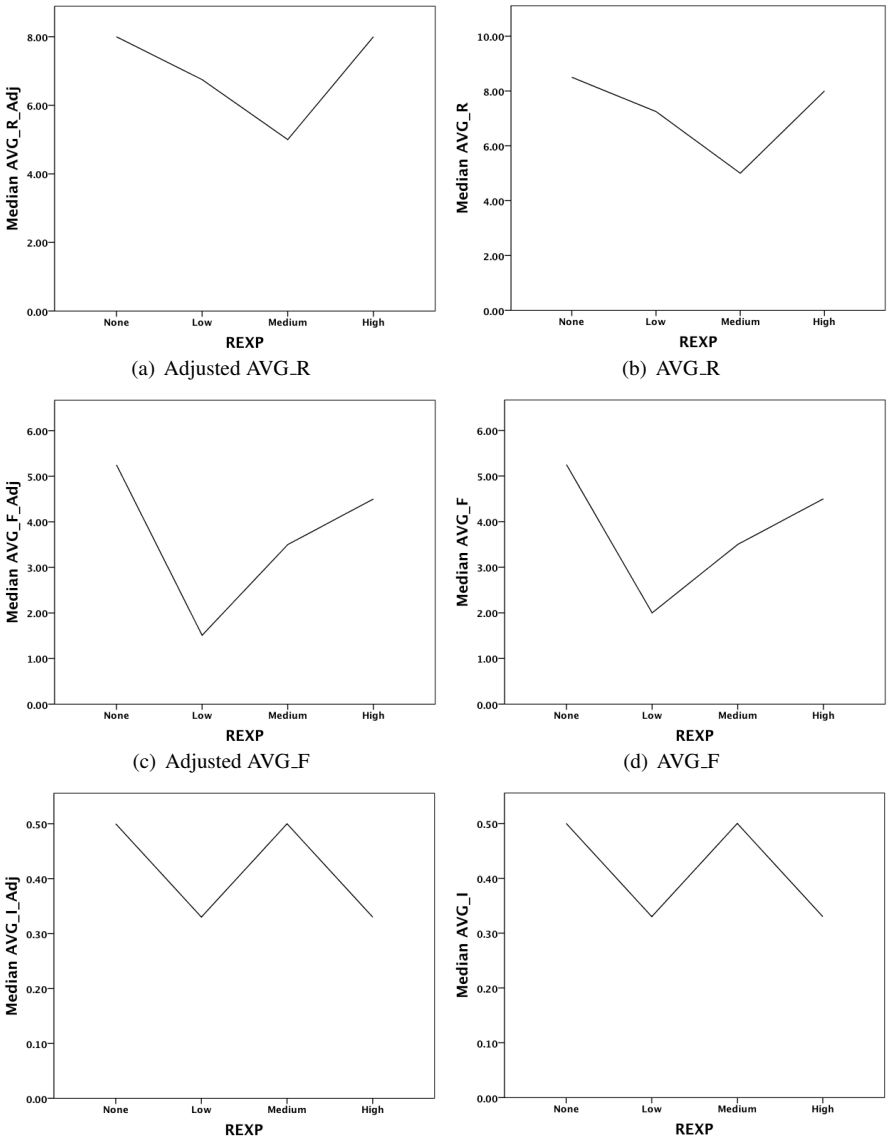
(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

Fig. 36: Adjusted Ideas vs. REXP – Ideas vs. REXP (Unfiltered)

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.



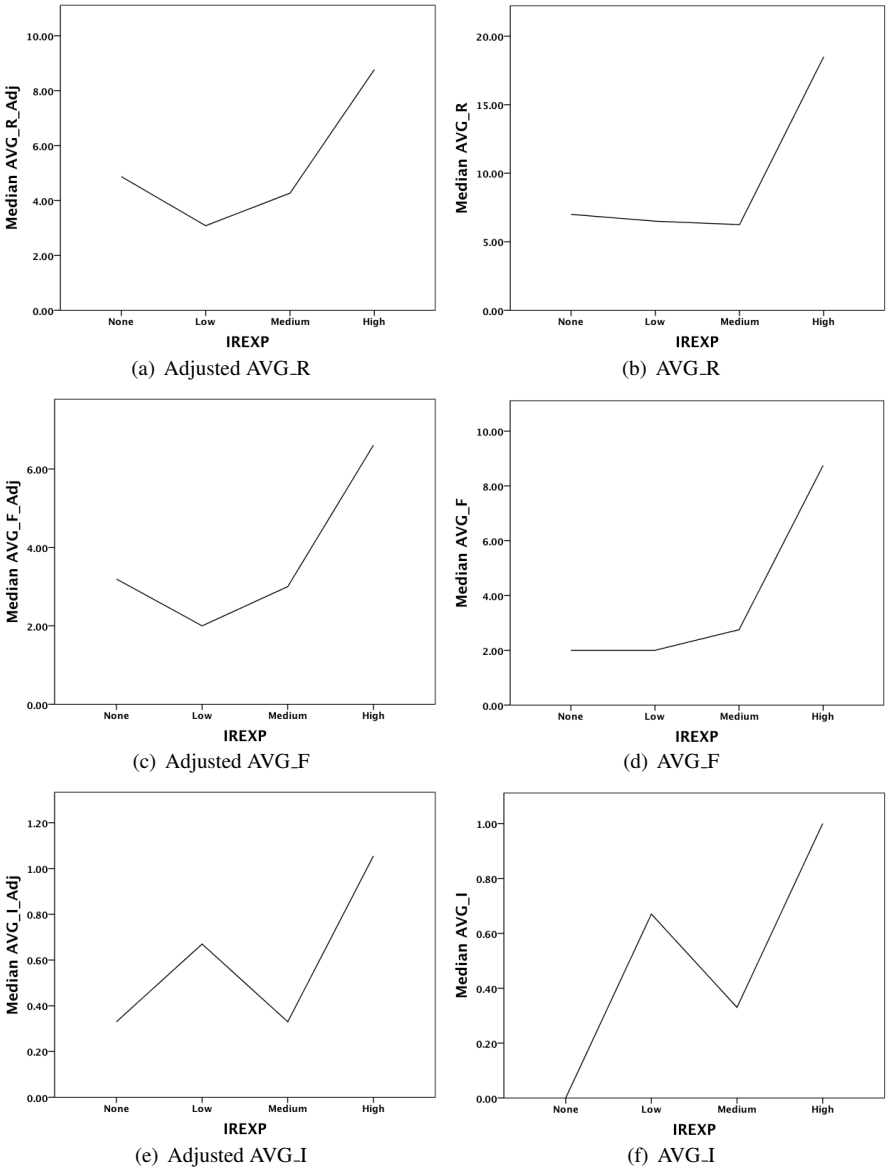
(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

Fig. 37: Adjusted Ideas vs. REXP – Ideas vs. REXP (Filtered)

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 requirements engineering projects.



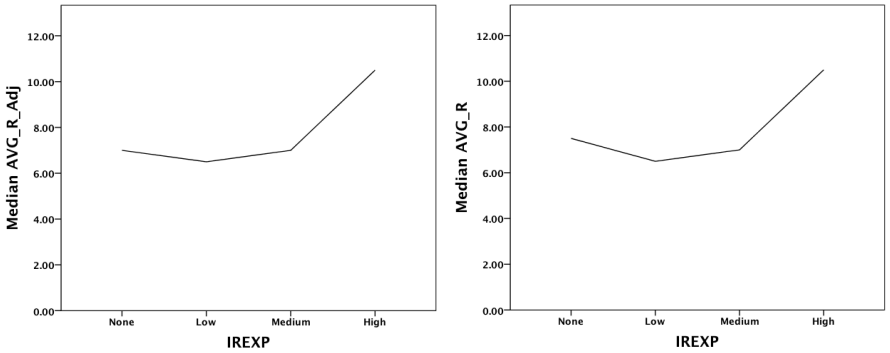
(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

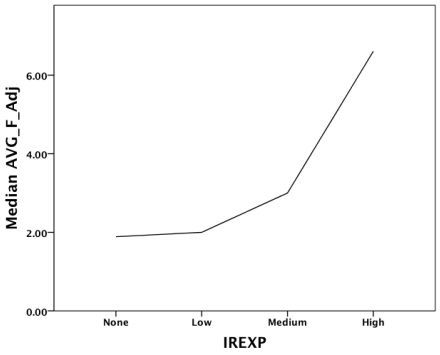
Fig. 38: Adjusted Ideas vs. IREXP – Ideas vs. IREXP (Unfiltered)

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.

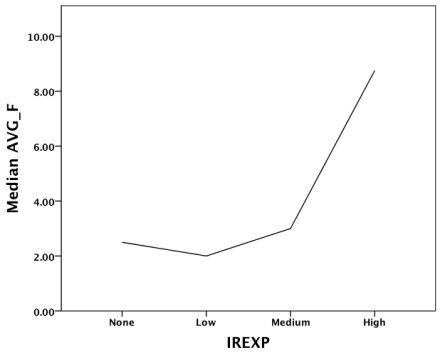


(a) Adjusted AVG_R

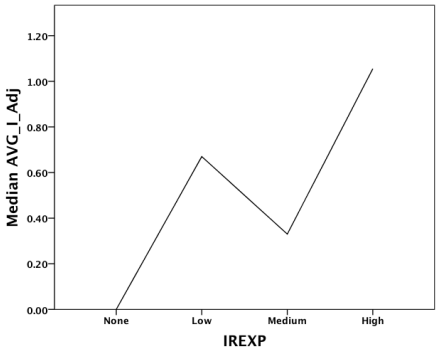
(b) AVG_R



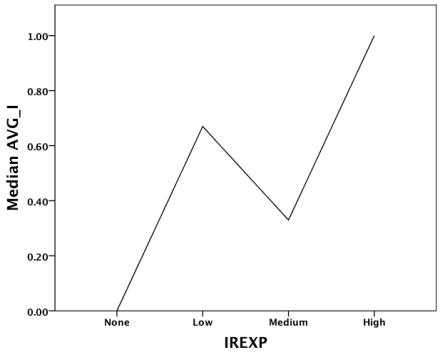
(c) Adjusted AVG_F



(d) AVG_F



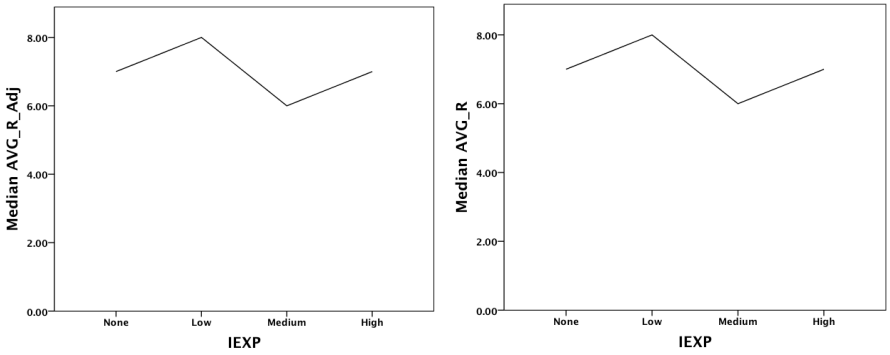
(e) Adjusted AVG_I



(f) AVG_I

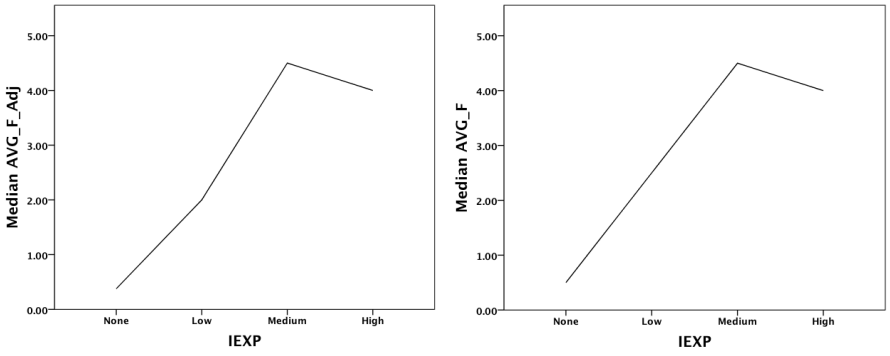
Fig. 39: Adjusted Ideas vs. IREXP – Ideas vs. IREXP (Filtered)

5. Our method of having the experimenters evaluate generated requirement ideas, described in Section 7.3 ensures that no evaluator knew from which team any idea came and thus that each evaluator could focus on applying his or her expertise to evaluate all ideas accurately and uniformly.



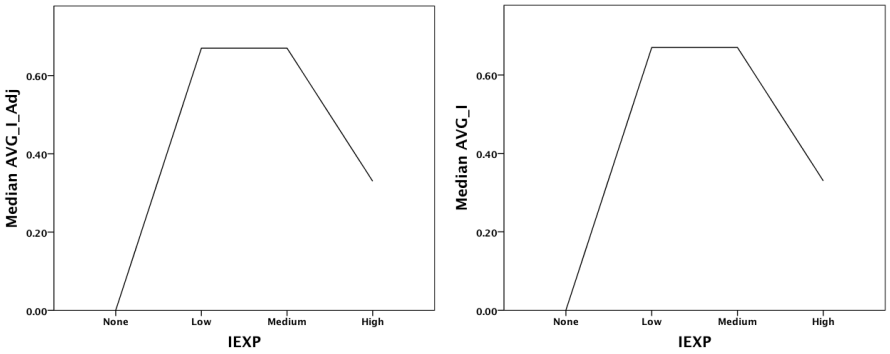
(a) Adjusted AVG_R

(b) AVG_R



(c) Adjusted AVG_F

(d) AVG_F



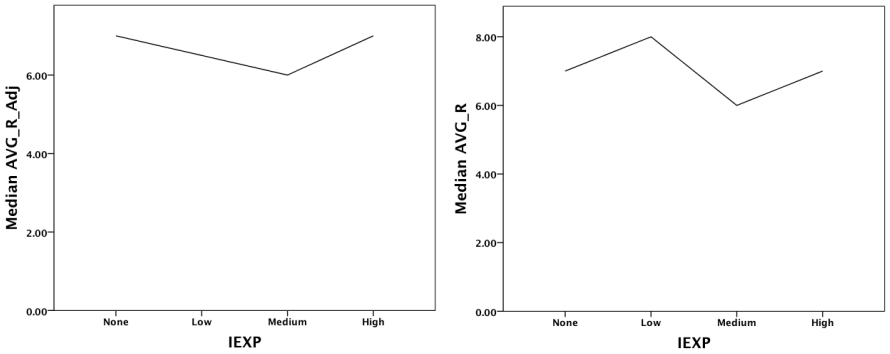
(e) Adjusted AVG_I

(f) AVG_I

Fig. 40: Adjusted Ideas vs. IEXP – Ideas vs. IEXP (Unfiltered)

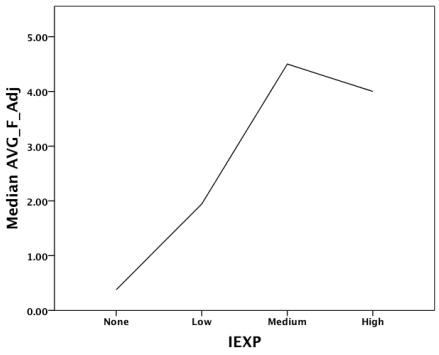
13.4 Threats to External Validity

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

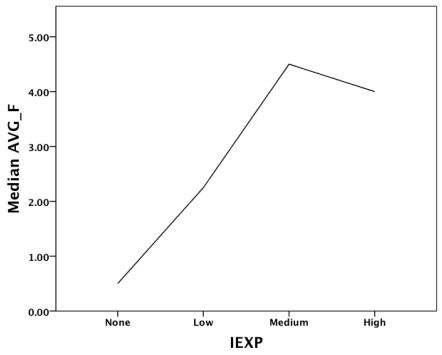


(a) Adjusted AVG_R

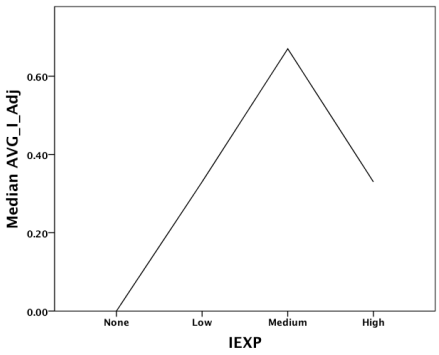
(b) AVG_R



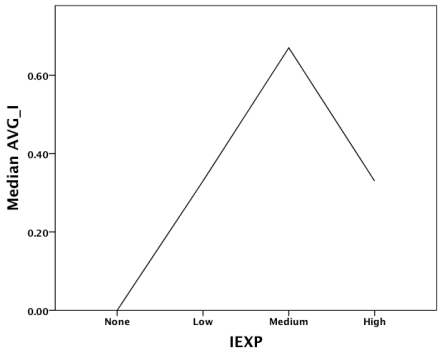
(c) Adjusted AVG_F



(d) AVG_F



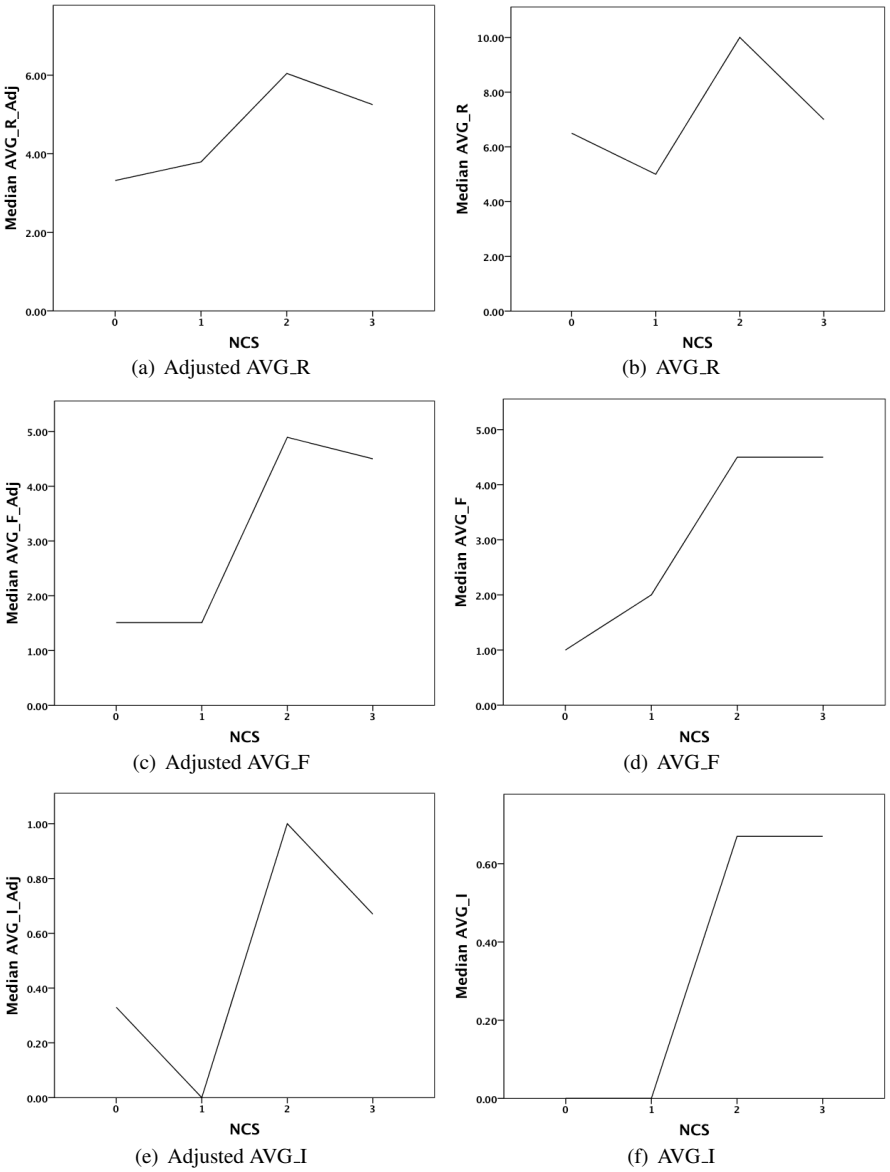
(e) Adjusted AVG_I



(f) AVG_I

Fig. 41: Adjusted Ideas vs. IEXP – Ideas vs. IEXP (Filtered)

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



(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

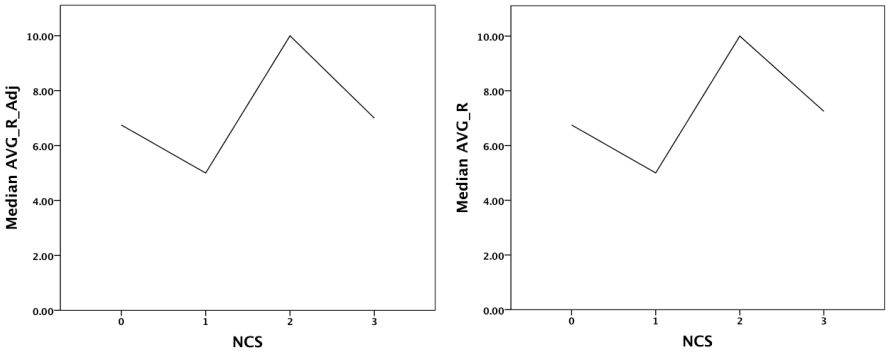
(e) Adjusted AVG_I (f) AVG_I

Fig. 42: Adjusted Ideas vs. NCS – Ideas vs. NCS (Unfiltered)

3. the use of the medium-sized application of a BDWP as the application about which to generate requirement ideas.

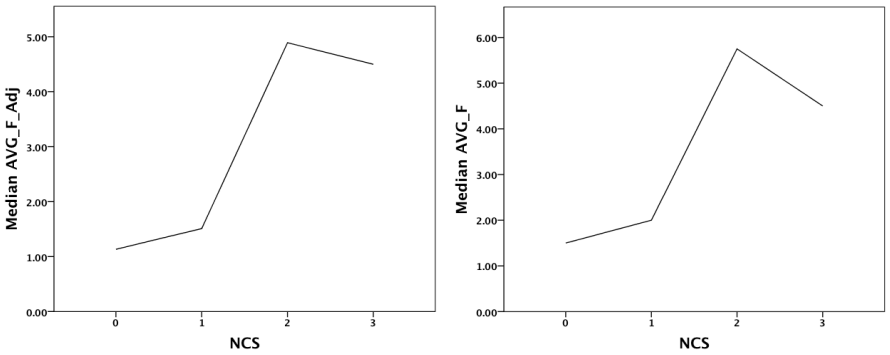
These threats require more attention than most of the threats to internal validity.

1. The goal of most empirical studies in software engineering is to draw conclusions valid for practitioners. However, convincing companies to allow their prac-



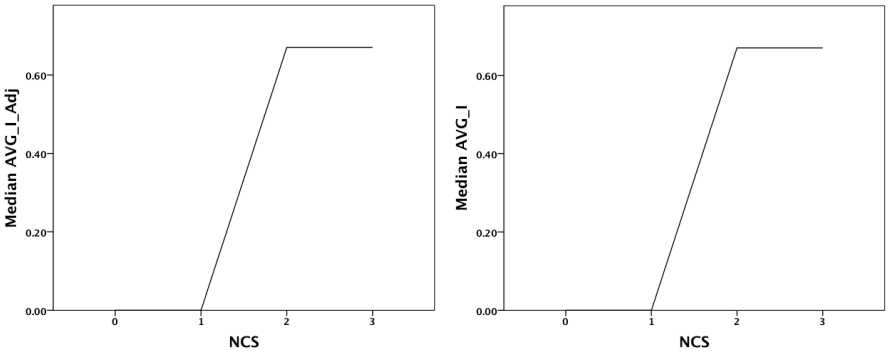
(a) Adjusted AVG_R

(b) AVG_R



(c) Adjusted AVG_F

(d) AVG_F

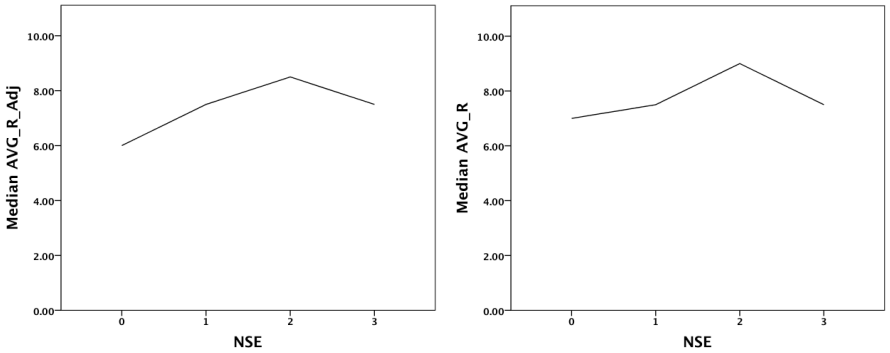


(e) Adjusted AVG_I

(f) AVG_I

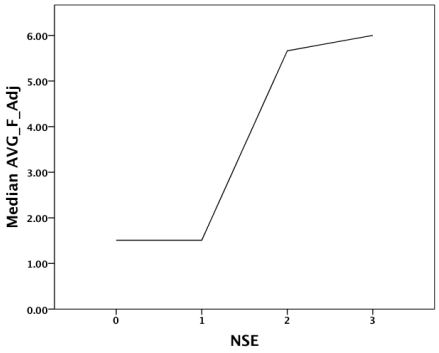
Fig. 43: Adjusted Ideas vs. NCS – Ideas vs. NCS (Filtered)

titioner 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. [28], conducted some experiments using both stu-

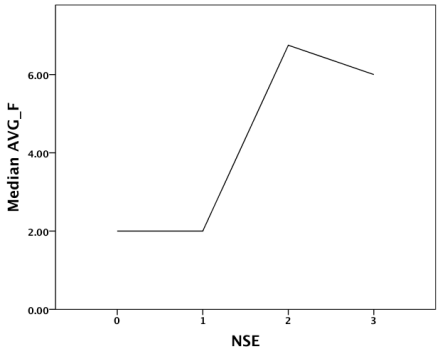


(a) Adjusted AVG_R

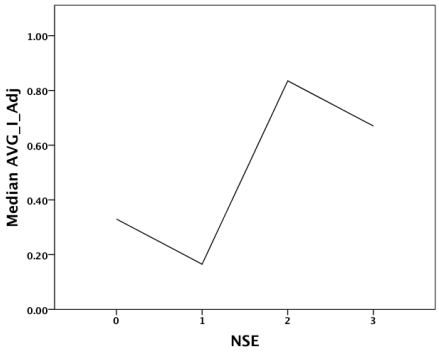
(b) AVG_R



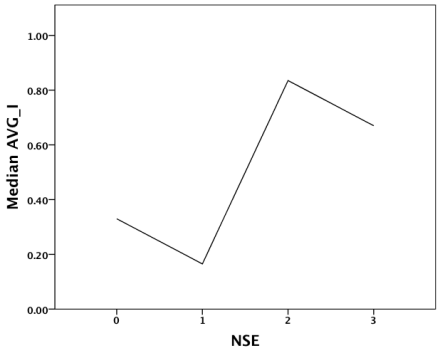
(c) Adjusted AVG_F



(d) AVG_F



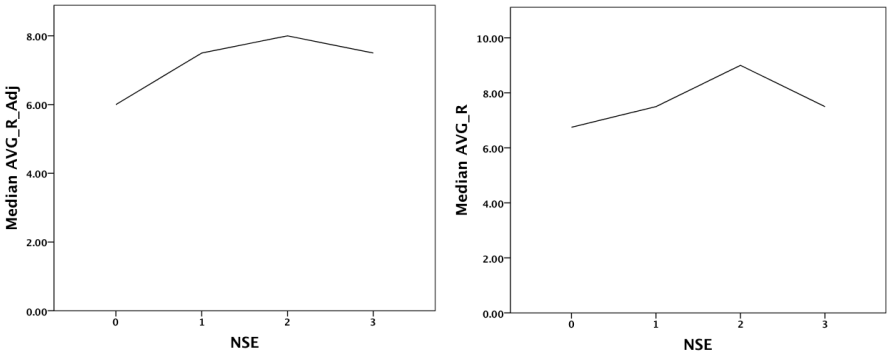
(e) Adjusted AVG_I



(f) AVG_I

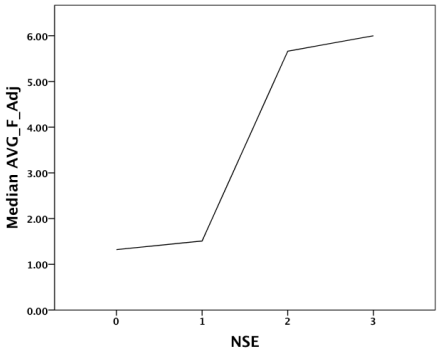
Fig. 44: Adjusted Ideas vs. NSE – Ideas vs. NSE (Unfiltered)

dents 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 software engineering. Note that the purpose of their experiments was to identify the factors affecting the lead time of software development projects.

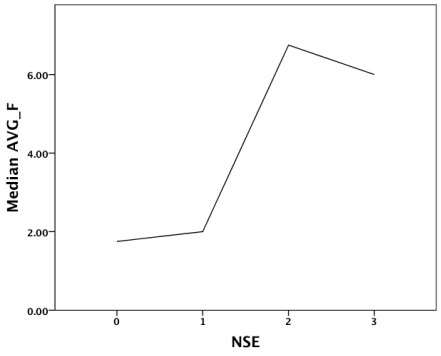


(a) Adjusted AVG_R

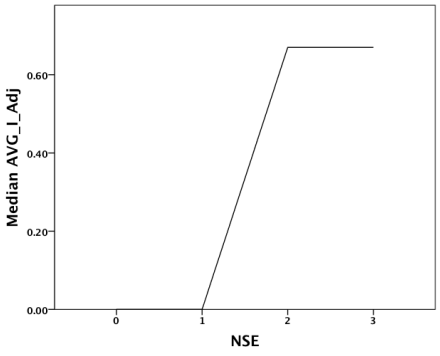
(b) AVG_R



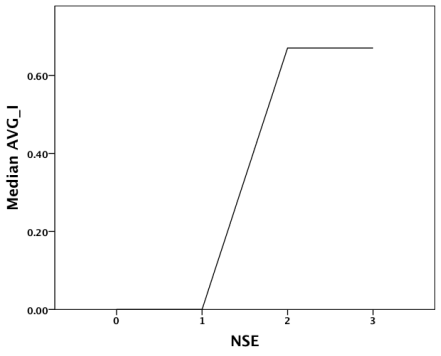
(c) Adjusted AVG_F



(d) AVG_F



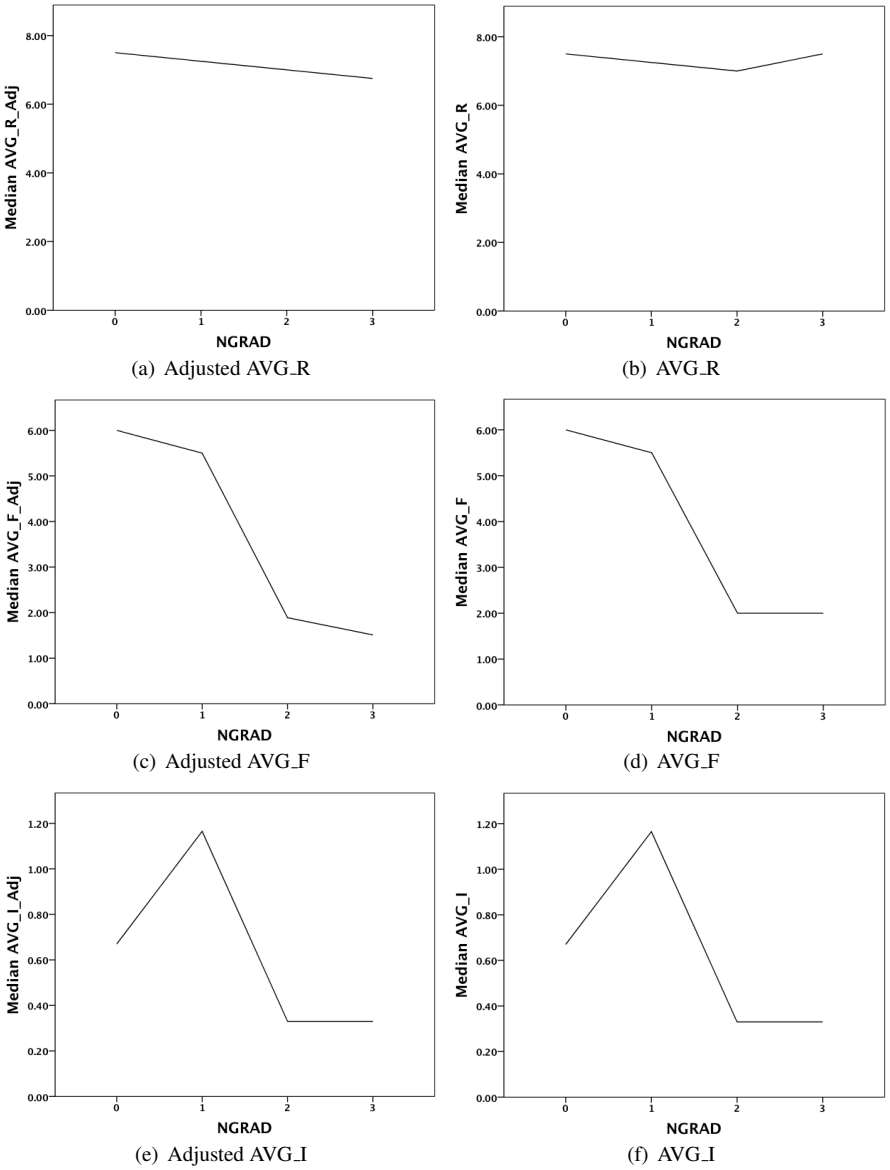
(e) Adjusted AVG_I



(f) AVG_I

Fig. 45: Adjusted Ideas vs. NSE – Ideas vs. NSE (Filtered)

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 UW includes courses that cover software requirements and specification. Moreover, almost all University of Waterloo 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



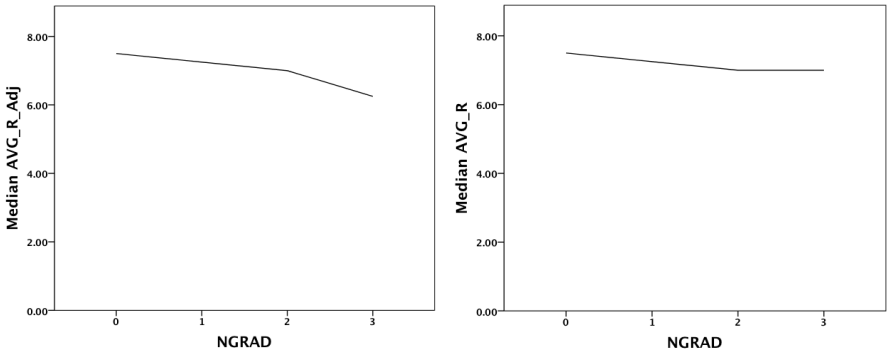
(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

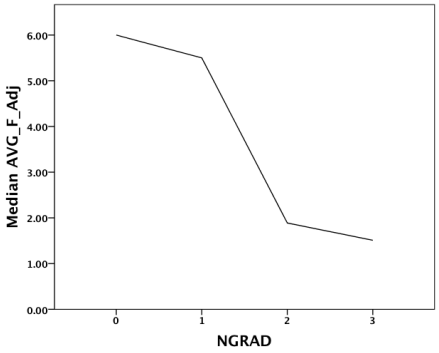
Fig. 46: Adjusted Ideas vs. NGRAD – Ideas vs. NGRAD (Unfiltered)

these students includes software development. Finally, 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. The purpose of the REXP, IREXP, and IEXP independent variables is to measure the extent to which these assumptions about the student participants is correct.

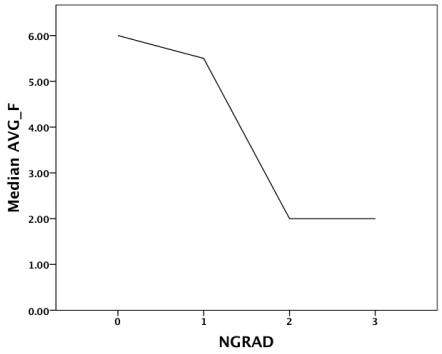


(a) Adjusted AVG_R

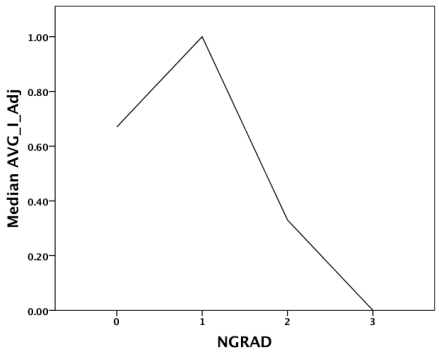
(b) AVG_R



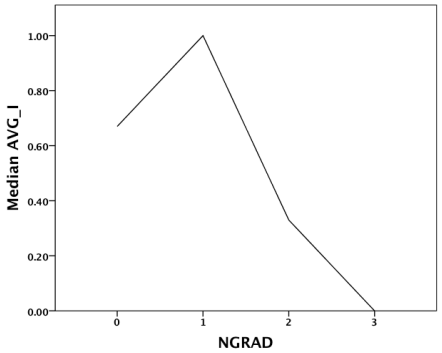
(c) Adjusted AVG_F



(d) AVG_F



(e) Adjusted AVG_I



(f) AVG_I

Fig. 47: Adjusted Ideas vs. NGRAD – Ideas vs. NGRAD (Filtered)

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 12.7

and 12.12 and as discussed in Section 15, the results *were* effected. 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 [51], 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 [38].

14 Conclusions

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. Initial observations were drawn from plots of the data.
3. Statistical analyses were performed next on the eight original independent variables plus the two factors identified by the factor analysis.

Table 95 summarizes the initial observations of Section 11 and the statistical analysis results of Section 12.

Recall that a team's effectiveness in requirement idea generation is measured by the number of requirement ideas of all kinds that the team generated.

MIX: In general, teams with at least one DI were more effective than teams with no DIs.

CR: Also, teams with a medium level of CR were more effective than the others. Therefore, it appears that an average level of creativity is required for a team to be effective. Left open is the question of why more creativity does not necessarily lead to more effectiveness.

REXP: For REXP, teams with no REXP were at least as effective as teams with some REXP. A possible explanation for this phenomenon is that the teams totally naive to RE were generating ideas more freely without being constrained by standard RE practices.

IEXP: Unlike for REXP, teams with more IREXP were more effective than the others. A team's IEXP was positively correlated with the effectiveness of a team. However, the effectiveness of the teams with a high level of IEXP is slightly less than that of the teams with a medium level of IEXP.

NCS: Considering educational background, teams with NCS of at least 2 were generally most effective. Also CS knowledge is sort of domain knowledge, but it is different from problem domain knowledge.

Independent Variable	*Initial Observations*	*Statistical Analysis*
MIX	is partially positively correlated with the number of generated ideas.	has no significant effect on any dependent variable.
CR	is partially negatively correlated with the number of generated ideas.	has no significant effect on any dependent variable.
REXP	is not correlated with the number of generated ideas.	has a significant effect on only one unfiltered dependent variable, NR, but has no statistically significant effect on the other dependent variables.
IREXP	is partially positively correlated with the number of generated ideas.	has no significant effect on any dependent variable.
IEXP	is partially positively correlated with the number of generated ideas.	has no significant effect on any dependent variable.
NCS	is partially positively correlated with the number of generated ideas.	has a significant effect on two filtered dependent variables, NF and NI, but has no statistically significant effect on the other dependent variables.
NSE	is partially positively correlated with the number of generated ideas.	has a significant effect on one unfiltered dependent variable, NF, and two filtered dependent variables, NF and NI, but has no significant effect on the other dependent variables.
NGRAD	is partially negatively correlated with the number of generated ideas.	has a significant effect on three filtered dependent variables, NRAW, NF, and NI, but has no statistically significant effect on the other dependent variables.
EDU	is partially positively correlated with the number of generated ideas.	has a significant effect on three dependent variables, NRAW, NF and NI, in both their filtered and unfiltered versions.
EXP	is partially positively correlated with the number of generated ideas.	has a significant effect on only one dependent variable, NI, in both its filtered and unfiltered versions.

Table 95: Summary of the Initial Observations and Statistical Analysis Results

NSE: Similar to with NCS, teams with NSE of at least 2 were generally most effective. The same explanation made about NCS makes sense here as well. Also SE knowledge is a sort of domain knowledge, but it is different from problem domain knowledge.

The results of the initial observations and statistical analysis on the full set of data for forty teams are taken into account to confirm or disprove the hypotheses:

H_{MIX}: The initial observations revealed that the effectiveness of a team is affected by the team's MIX. The statistical analysis showed that this variable is statistically significant only in conjunction with EXP and EDU. Therefore H_{MIX_1} is weakly rejected and H_{MIX_0} is weakly accepted.

H_{CR}: The initial observations revealed that the effectiveness of a team is positively affected by the team's CR. The statistical analysis did not show any significant effect of this variable 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, NSE and NCS. The initial observations revealed that the effectiveness of a team is positively affected by the team's NCS and NSE. The statistical analysis showed that the effect of NCS and

NSE is statistically significant on most dependent variables. Therefore, H_{EDU_1} is strongly accepted and H_{EDU_0} is strongly rejected.

H_{NGRAD}: The initial observations revealed that the effectiveness of a team is negatively affected by the team's NGRAD. The statistical analysis showed that the effect of this variable is statistically significant on most dependent variables. Therefore, H_{NGRAD_1} is strongly accepted and H_{NGRAD_0} is rejected.

H_{EXP}: A team's EXP incorporates three separate variables, REXP, IREXP, and IEXP. The initial observations revealed that the effectiveness of a team is positively affected by the team's IEXP and IREXP, but is negatively affected by the team's REXP. The statistical analysis did not show any significant effect of IEXP and IREXP on any dependent variable and REXP showed a small effect on only one dependent variable. Therefore, H_{EXP_1} is rejected and H_{EXP_0} is accepted.

15 Comparing results of E1 and E1+E2

In E1, each of the participants was a CS or SE student. The results reported in the conference paper by the same authors [37] suggest that those RE teams with a mix of domain familiarities are more effective than teams composed of only one domain familiarity. 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.

Although the initial observations of the results of the combined E1+E2 data are not very different from those of E1, 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, the 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 and the initial graphical analysis of the combined E1+E2 data showed some support for accepting the main hypothesis, the statistical analysis of the combined E1+E2 data did not provide *any* support for accepting this hypothesis.

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 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 technnology fields, of the participants affected the results. Certainly, the results of Section 12.9 say that the educational backgrounds of the members of a team affects the number of ideas generated by the team. The reality is that main hypothesis carries an assumption that all analysts involved in idea generation *are competent in their CS-or-SE-related professions*. So, in having to to use participants from outside CS and SE, we may 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.

16 Future Work

There are many activities other than requirement idea generation that could benefit from domain ignorance. One such RE activity is requirements specification inspection. Requirements specification inspection is basically brainstorming for signs of defects in the inspected requirements specification.

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 will potentially benefit from domain ignorance. The literature shows that a few of the disciplines that benefit from domain knowledge are cross-functional communication [13], data mining [2, 31], and exploratory software testing [29]. 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 [22]. 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.

As for any empirical study, more data points will improve the results of the controlled experiment described in this study. 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 15, 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,
2. Domain Generalist (DG): those who have only a general picture of the domain or have some knowledge of a similar domain that can make analogies with the domain under study,
3. Domain Novice (DN): those who have a limited knowledge of the domain by being exposed to the domain without becoming a DE, e.g. iPhone users vs iPhone application programmers, and
4. Domain Ignorant (DI): those who have no domain knowledge 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* DI was thoroughly ignorant of the domain and *every* DA was was a user of a BDWP, there were no DNs, and only Berry, having implemented a BDWP, was a DE.

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.

Also, conducting the experiment on different problem domains is beneficial in order to extend the external validity of the experiment. 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 much less significance than with controlled experiments.

Finally, while this work focused on RE, the findings might be applicable to the broader domain of SE.

Acknowledgements 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

1. Al-Rawas, A., Easterbrook, S.: 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 (1996)
2. Anand, S.S., Bell, D.A., Hughes, J.G.: The role of domain knowledge in data mining. In: Proceedings of the Fourth International Conference on Information and Knowledge Management (CIKM), pp. 37–43 (1995)
3. Apfelbaum, E.P., Phillips, K.W., Richeson, J.A.: Rethinking the baseline in diversity research: Should we be explaining the effects of homogeneity? Perspectives on Psychological Science **9**(3), 235–244 (2014)
4. Basili, V.R., Caldiera, G., Rombach, D.H.: The Goal Question Metric Approach. In: J.J. Marciniak (ed.) Encyclopedia of Software Engineering, vol. I. John Wiley & Sons (1994)
5. Berenbach, B., Paulish, D.J., Kazmeier, J., Rudorfer, A.: Software & Systems Requirements Engineering: In Practice. McGraw-Hill, New York, NY USA (2009)
6. Berry, D.M.: The importance of ignorance in requirements engineering. Journal of Systems and Software **28**(2), 179–184 (1995)
7. Berry, D.M.: The importance of ignorance in requirements engineering: An earlier sighting and a revisitation. Journal of Systems and Software **60**(1), 83–85 (2002)
8. Blom, G.: Statistical estimates and transformed beta-variables. The Incorporated Statistician **10**(1), 53–55 (1960)
9. Brooks, F.P.: The Mythical Man-Month: Essays on Software Engineering, 20th Anniversary Edition. Addison-Wesley Professional, Boston, MA, USA (1995)
10. Carver, J.C., Nagappan, N., Page, A.: The impact of educational background on the effectiveness of requirements inspections: An empirical study. Software Engineering, IEEE Transactions on **34**(6), 800–812 (2008)
11. Corp., I.: Post hoc comparisons for the Kruskal-Wallis test (2013). URL `http://www-01.ibm.com/support/docview.wss?uid=swg21477370`. [Online; accessed 11-Sep-2013]
12. Dagenais, B., Ossher, H., Bellamy, R.K.E., Robillard, M.P., de Vries, J.P.: Moving into a new software project landscape. In: Proceedings of the International Conference on Software Engineering (ICSE), Volume 1, pp. 275–284 (2010)
13. Damian, D., Helms, R., Kwan, I., Marczak, S., Koelewijn, B.: 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 (2013)
14. Dieste, O., Juristo, N., Shull, F.: Understanding the customer: What do we know about requirements elicitation? IEEE Software **25**(2), 11–13 (2008)
15. Dunbar, K.: How scientists build models invivo science as a window on the science mind. In: L. Magnani, N. Nersessian, P. Thagard (eds.) Model-Based Reasoning in Scientific Discovery, pp. 85–99. Kluwer Academic/Plenum Publishers, New York, NY, USA (1999)
16. Dybå, T., Kampenes, V.B., Sjøberg, D.I.K.: A systematic review of statistical power in software engineering experiments. Information & Software Technology **48**(8), 745–755 (2006)
17. Feldt, R., Magazinius, A.: 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 (2010)
18. Ferrari, R., Madhavji, N.H.: 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) (2007)
19. Finkelstein, A.: Requirements engineering: a review and research agenda. In: Proceedings of the First Asia-Pacific Software Engineering Conference, pp. 10–19 (1994)
20. Firestein, S.: Ignorance (Course) (2013). URL `http://bioweb.biology.columbia.edu/firestein/?page_id=36`
21. Fischer, G.: Symmetry of igorance, social creativity, and meta-design. In: Proceedings of the 3rd Conference on Creativity & Cognition (C&C), pp. 116–123 (1999)
22. Frappaolo, C.: Implicit knowledge. Knowledge Management Research and Practice **6**(1), 23–25 (2008)
23. Glass, G.V., Peckham, P.D., Sanders, J.R.: Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of Educational Research **42**(3), 237–288 (1972)

24. Grace-Martin, K.: Outliers: To drop or not to drop (2013). URL http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/
25. Hadar, I., Soffer, P., Kenzi, K.: The role of domain knowledge in requirements elicitation via interviews: an exploratory study. Requirements Engineering Journal **19**(2), 143–149 (2014)
26. Hanebutte, N., Taylor, C.S., Dumke, R.R.: Techniques of successful application of factor analysis in software measurement. Empirical Software Engineering **8**(1), 43–57 (2003)
27. Hinton, P.R., McMurray, I., Brownlow, C.: SPSS Explained. Routledge, East Sussex, UK (2004)
28. Höst, M., Regnell, B., Wohlin, C.: Using students as subjects - a comparative study of students and professionals in lead-time impact assessment. Empirical Software Engineering **5**(3), 201–214 (2000)
29. Itkonen, J., Mantyla, M.V., Lassenius, C.: The role of the tester's knowledge in exploratory software testing. IEEE Transactions on Software Engineering **39**(5), 707–724 (2013)
30. Jarke, M., Jr., J.A.B., Rolland, C., Sutcliffe, A.G., Vassiliou, Y.: Theories underlying requirements engineering: an overview of NATURE at Genesis. In: Proceedings of the IEEE International Symposium on Requirements Engineering (RE), pp. 19–31 (1993)
31. Kopanas, I., Avouris, N.M., Daskalaki, S.: The role of domain knowledge in a large scale data mining project. In: C.D.S. Ioannis P. Vlahavas (ed.) Methods and Applications of Artificial Intelligence, *Lecture Notes in Computer Science*, vol. 2308, pp. 288–299. Springer, Berlin, Germany (2002)
32. Kristensson, P., Gustafsson, A., Archer, T.: Harnessing the creative potential among users. Journal of Product Innovation Management **21**(1), 4–14 (2004)
33. Lehrer, J.: Accept defeat: The neuroscience of screwing up (2009). URL http://www.wired.com/2009/12/fail_accept_defeat. [Online; accessed 6-May-2014]
34. Mehrotra, G.: Role of domain ignorance in software development. Master's thesis, University of Waterloo, Waterloo (2011). URL http://se.uwaterloo.ca/~dberry/FTP_SITE/students.theses/gaurav.mehrotra/gauravMehrotraThesis.pdf
35. Naur, P., Randell, B.: Software Engineering: Report of a conference sponsored by the NATO Science Committee. Scientific Affairs Division, NATO, Brussels, Belgium (1969)
36. Niknafs, A.: The impact of domain knowledge on the effectiveness of requirements engineering activities. Ph.D. thesis, University of Waterloo, Waterloo (2014). URL https://uwspace.uwaterloo.ca/handle/10012/8470
37. Niknafs, A., Berry, D.M.: 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 (2012)
38. Osborn, A.: Applied Imagination. Charles Scribner's, New York, NY, USA (1953)
39. Pascal, B., Krailsheimer, A.J.: Pensees: Translated with an Introduction by A.J. Krailsheimer. Penguin, London, UK (1968)
40. Rose, P., Kumar, M., Ajmeri, N., Agrawal, M., Sivakumar, V., Ghaisas, S.: 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 (2009)
41. Sharp, H.: The role of domain knowledge in software design. Behaviour and Information Technology **10**(5), 383–401 (1991)
42. Taylor, C.W., Williams, F.E.: Instructional Media and Creativity: The Proceedings of the Sixth Utah Creativity Research Conference. Distributed by ERIC Clearinghouse, Washington, D.C., USA (1965). URL http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED010651,http://nla.gov.au/nla.cat-vn5184417
43. Technologies, B.: Prophet 5: Statistical and sequence analysis software (2013). URL http://www.basic.northwestern.edu/biotools/prophet.html
44. Thagard, P.: Collaborative knowledge. Noûs **31**(2), 242–261 (1997)
45. University of Utah, P.D.: Psystats (2013). URL http://psystats.wikispaces.com/
46. Warner, R.M.: Applied Statistics: From Bivariate Through Multivariate Techniques: From Bivariate Through Multivariate Techniques. Sage Publications, Thousand Oaks, CA, USA (2012)
47. Wikipedia: Kurtosis — Wikipedia, the free encyclopedia (2013). URL http://en.wikipedia.org/wiki/Kurtosis. [Online; accessed 22-Aug-2013]
48. Wikipedia: Skewness — Wikipedia, the free encyclopedia (2013). URL http://en.wikipedia.org/wiki/Skewness. [Online; accessed 22-Aug-2013]
49. Wikipedia: Tukey's range test — Wikipedia, the free encyclopedia (2013). URL http://en.wikipedia.org/wiki/Tukey's_range_test. [Online; accessed 1-Sept-2013]
50. Wikipedia: Analysis of variance — Wikipedia, the free encyclopedia (2014). URL http://en.wikipedia.org/wiki/Analysis_of_variance. [Online; accessed 7-Feb-2014]

51. Wikipedia: Textedit— Wikipedia, the free encyclopedia (2014). URL `http://en.wikipedia.org/wiki/TextEdit`. [Online; accessed 2-May-2014]
52. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in software engineering: an introduction. Kluwer Academic Publishers, Norwell, MA, USA (2000)

Tables 96, 97, and 98 are below:

IV	DV	Fig#	Filt'd?	Corr?	Diff?	WhenMax?	H{IV}1	H{IV}0	E1match?
MIX	RAW	10(a)	U	+	No	3I	supp'd	not supp'd	Yes
		11(a)	F	+		3I			
	AVG_R	10(b)	U	Not	No	1I			
		11(b)	F	Not		1I			
	AVG_F	10(c)	U	+	No	3I			
		11(c)	F	+		3I			
	AVG_I	10(d)	U	Not	Yes	3I			
		11(d)	F	part+		3I			
CR	RAW	12(a)	U	part–	Yes	Med	not supp'd	supp'd	Yes
		13(a)	F	Not		Med			
	AVG_R	12(b)	U	–	No	Low			
		13(b)	F	–		Low			
	AVG_F	12(c)	U	Not	No	Med			
		13(c)	F	Not		Med			
	AVG_I	12(d)	U	Not	No	Med			
		13(d)	F	Not		Med			
REXP	RAW	14(a)	U	Not	No	None	not supp'd	supp'd	Yes
		15(a)	F	Not		None			
	AVG_R	14(b)	U	Not	No	None			
		15(b)	F	Not		None			
	AVG_F	14(c)	U	Not	No	None			
		15(c)	F	Not		None			
	AVG_I	14(d)	U	part+	Yes	High			
		15(d)	F	Not		None & Med			
IREXP	RAW	16(a)	U	+	Slight	High	supp'd	not supp'd	N.A.
		17(a)	F	part+		Med			
	AVG_R	16(b)	U	part+	No	High			
		17(b)	F	part+		High			
	AVG_F	16(c)	U	part+	No	High			
		17(c)	F	part+		High			
	AVG_I	16(d)	U	part+	No	High			
		17(d)	F	part+		High			
IEXP	RAW	18(a)	U	part+	Slight	Med	supp'd	not supp'd	Yes
		19(a)	F	part+		Med			
	AVG_R	18(b)	U	Not	No	Low			
		19(b)	F	Not		Low			
	AVG_F	18(c)	U	part+	No	Med			
		19(c)	F	part+		Med			
	AVG_I	18(d)	U	part+	Slight	Low & Med			
		19(d)	F	part+		Med			
NCS	RAW	20(a)	U	+	Slight	3	supp'd	not supp'd	N.A.
		21(a)	F	part+		3			
	AVG_R	20(b)	U	Not	No	2			
		21(b)	F	Not		2			
	AVG_F	20(c)	U	part+	No	2 & 3			
		21(c)	F	part+		2			
	AVG_I	20(d)	U	part+	No	2 & 3			
		21(d)	F	part+		2 & 3			

Table 96: Summary of Initial Analysis: Part I

IV	DV	Fig#	Filt'd?	Corr?	Diff?	WhenMax?	H{IV}1	H{IV}0	E1match?
NSE	RAW	22(a)	U	+	No	3	supp'd	not supp'd	N.A.
		23(a)	F	+		3			
	AVG_R	22(b)	U	Not	No	2			
		23(b)	F	Not		2			
	AVG_F	22(c)	U	part+	No	2			
		23(c)	F	part+		2			
	AVG_I	22(d)	U	part+	No	2			
		23(d)	F	part+		2 & 3			
NGRAD	RAW	24(a)	U	–	Slight	0	not supp'd	supp'd	N.A.
		25(a)	F	part–		0 & 1			
	AVG_R	24(b)	U	Not	No	0 & 3			
		25(b)	F	Not		0			
	AVG_F	24(c)	U	–	No	0			
		25(c)	F	–		0			
	AVG_I	24(d)	U	part–	No	1			
		25(d)	F	part–		1			
EDU	RAW	26(a)	U	+	No	High	supp'd	not supp'd	N.A.
		27(a)	F	+		High			
	AVG_R	26(b)	U	+	No	High			
		27(b)	F	+		High			
	AVG_F	26(c)	U	+	No	High			
		27(c)	F	+		High			
	AVG_I	26(d)	U	+	No	High			
		27(d)	F	+		High			
EXP	RAW	28(a)	U	part+	No	Med	not supp'd	supp'd	N.A.
		29(a)	F	part+		Med & High			
	AVG_R	28(b)	U	Not	No	High			
		29(b)	F	Not		High			
	AVG_F	28(c)	U	part+	No	Med			
		29(c)	F	part+		Med			
	AVG_I	28(d)	U	Not	No	Med			
		29(d)	F	Not		Med			

Legend

section	A section of this table is the 8 rows lying between two consecutive double horizontal lines.
subsection	A subsection of this table is the 2 rows lying between two consecutive single horizontal lines, which are not the full width of the table.
IV	the independent variable that is considered in the current section
DV	the dependent variable that is considered in the current subsection
Fig#	The plot giving the results for the current row is found in the figure whose number is given.
Filt'd?	Is the DV in the current row filtered (denoted "F") or unfiltered (denoted "U")?
Corr?	According to the plot of the current row, is the DV in the current row correlated to the IV in the current section, and if so, how?
supp'd	supported
not supp'd	not supported
+	positively correlated
part+	partially positively correlated
–	negatively correlated
part–	partially negatively correlated
Not	not correlated
Diff?	According to the plots of the two rows of the current subsection, does removing the outliers (filtering) make a real difference?
WhenMax?	For which value of the IV of the current section is the value of the DV variable of the current subsection at its maximum?
H{IV}1	According to all of the plots of the current section, is the alternative hypothesis for the independent variable of the current section supported or unsupported?
H{IV}0	According to all of the plots of the current section, is the null hypothesis for the independent variable of the current section supported or unsupported?
E1match?	Do the conclusions for H{IV}1 and H{IV}0 for the current section match those of E1?
N.A.	not applicable, because E1 did not test any hypotheses for IV

Table 96: Summary of Initial Analysis: Part II

IV	LeveneT#			ANOVA_T#					K-W_T#		
	Filt'd?	DV	App'le?	SigEff?	TukeyT#	WhenSig?	DV	Need?	SigEff?	DB_T#	WhenSig?
MIX	15			17					19		
	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	–	–
	16			18					20		
	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	21			23					25		
	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	–	–
	22			24					26		
	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	27			29					32		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	Yes	31	Med:High	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
	28			30					33		
	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	34			36					38		
	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	–	–
	35			37					39		
	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	–	–
IEXP	40			42					44		
	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	–	–
	41			43					45		
	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 97: Summary of Statistical Analysis: One-Way ANOVAs: Part I

IV	LeveneT#			ANOVA_T#					K-W_T#		
	Filt'd?	DV	App'le?	SigEff?	TukeyT#	WhenSig?	DV	Need?	SigEff?	DB_T#	WhenSig?
NCS	46			48					52		
	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	–	–
	47			49					53		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	Yes	50	0:3	AVG_F	No	–	–	–
		NI	Yes	Yes	51	0:3&1:3	AVG_I	Yes	Yes	54	0:3
NSE	55			57					62		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	Yes	59	0:2&0:3	AVG_F	No	–	–	–
		NI	Yes	No	–	–	AVG_I	Yes	No	–	–
	56			58					63		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	No	–	–	–	AVG_R	Yes	No	–	–
		NF	Yes	Yes	60	0:2&0:3& 1:3	AVG_F	No	–	–	–
		NI	Yes	Yes	61	0:2&0:3	AVG_I	Yes	Yes	64	0:3
NGRAD	65			67					72		
	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	–	–
	66			68					73		
	F	NRAW	Yes	Yes	69	0:3	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	Yes	70	0:3	AVG_F	No	–	–	–
		NI	Yes	Yes	71	0:3	AVG_I	Yes	Yes	74	0:3
The 3-way ANOVA data would come here, but they require different headings. See Table 98 for these data.											
EDU	79			81					83		
	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	*	–
	80			82					84		
	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	85			87					91		
	U	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	Yes	89	Low:Med	AVG_I	Yes	Yes	93	Low:Med
	86			88					92		
	F	NRAW	Yes	No	–	–	RAW	No	–	–	–
		NR	Yes	No	–	–	AVG_R	No	–	–	–
		NF	Yes	No	–	–	AVG_F	No	–	–	–
		NI	Yes	Yes	90	Low:Med& Med:High	AVG_I	Yes	Yes	94	Low:Med

Legend

section	A section of this table is the 10 rows lying between two consecutive double horizontal lines.
subsection	A subsection of this table is 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
LeveneT#	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?
TukeyT#	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.

Table 97: Summary of Statistical Analysis: One-Way ANOVAs: Part II

The postponed 3-way ANOVA data come here.

IV	LeveneT#			ANOVA_T#						
	Filt'd?	DV	App'le?	ALLsig?	MIX*EXPsig?	MIX*EDUsig?	EXP*EDUsig?	MIXsig?	EXPsig?	EDUsig?
MIX	75			77						
	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
EXP	76			78						
EDU	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 lying between two consecutive double horizontal lines.
subsection	A subsection of this table is 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
LeveneT#	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?
"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.

Table 98: Summary of Statistical Analysis: Three-Way ANOVA