The Impact of Domain Knowledge on the Effectiveness of Requirements Engineering Activities

by

Ali Niknafs

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2014

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

I acknowledge that my supervisor, Dr. Daniel Berry, has co-authored the papers included in this thesis:

1. Ali Niknafs and Daniel M. Berry. The impact of domain knowledge on the effectiveness of requirements idea generation during requirements elicitation. In 20th IEEE International Requirements Engineering Conference (RE), pages 181–190, Chicago, IL, USA, 2012

   Contributed in part to Chapter 5.

2. Ali Niknafs and Daniel M. Berry. An industrial case study of the impact of domain ignorance on the effectiveness of requirements idea generation during requirements elicitation. In 21st IEEE International Requirements Engineering Conference (RE), pages 279–283, Rio de Janeiro, Brazil, 2013

   Contributed in part to Chapter 7.

Abstract

[Context] One of the factors that seems to influence an individual's effectiveness in requirements engineering activities is his or her knowledge of the problem being solved, i.e., domain knowledge. While in-depth domain knowledge enables a requirements engineer to understand the problem easier, he or she can fall for tacit assumptions of the domain and might overlook issues that are obvious to domain experts and thus remain unmentioned.

[Objective] The purpose of this thesis is to investigate the impact of domain knowledge on different requirements engineering activities. The main research question this thesis attempts to answer 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] This thesis presents two controlled experiments and an industrial case study to test a number of hypotheses. The main hypothesis states that a requirements engineering team for a computer-based system in a particular domain, consisting of a mix of requirements analysts that are ignorant of the domain and requirements analysts that are aware of the domain, is more effective at requirement idea generation than a team consisting of only requirements analysts that are aware of the domain.

[Results] The results of the controlled experiments, although not conclusive, provided some support for the positive effect of the mix on effectiveness of a requirements engineering team. The results also showed a significant effect of other independent variables, especially educational background. The data of the case study corroborated the results of the controlled experiments.

[Conclusion] The main conclusion that can be drawn from the findings of this thesis is that the presence in a requirements engineering team of a domain ignorant with a computer science or software engineering background improves the effectiveness of the team.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Daniel Berry, whose guidance, knowledge, patience and engagement throughout the whole process brought this work into existence.

I would also like to thank my committee members, Dr. Paul Thagard, Dr. Richard Trefler, Dr. Krzysztof Czarnecki, and Dr. Didar Zowghi, for letting me have an enjoyable defence.

I am deeply thankful to the participants in my experiments who have willingly shared their time. Of course, this work would not have been possible without their participation. Also, I like to thank the company and its employees who participated in one of my studies.

My special thanks goes to Sahar, for her support, encouragement, and patience, both by keeping me motivated and standing by me through the ups and downs.

I would like to thank my parents whom I am forever indebted to, for all their unconditional love, unceasing prayers, and constant encouragement in all my pursuits.

This work was supported in part by a Canadian NSERC–Scotia Bank Industrial Research Chair (NSERC-IRCPJ365473-05).

To my beloved mom, dad, and sister

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANOVA	Analysis of variance
AVG_F	Average number of feasible ideas
AVG_I	Average number of innovative ideas
AVG_R	Average number of relevant ideas
BDWP	Bidirectional word processor
CBS	Computer-based system
CR	Average creativity score level
DA	Domain aware
DE	Domain expert
DGU	Desired generated unit
DI	Domain ignorant
E1	Controlled experiment 1
E2	Controlled experiment 2
EXP	Experience - Aggregate of REXP, IREXP, and IEXP
EDU	Education - Aggregate of NSE and NCS
GT	Generative task
IEXP	Average industrial experience
IREXP	Average industrial RE experience
MIX	Mix of domain familiarities
NCS	Number of participants with CS background
NF	Normalized AVG_F
NGRAD	Number of graduate student participants
NI	Normalized AVG_I
NR	Normalized AVG_R
NSE	Number of participants studying SE
NRAW	Normalized RAW
RAW	Raw number of ideas
RE	Requirements engineering
REXP	Average RE experience
RQ	Research question

1 Introduction

A key step of any software development is deciding precisely what to build [12]. 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 [25]. Therefore, the more emphasis that is given to RE, the better the chances are of obtaining high quality software.

Of the three Ps, process, product, and people, in software engineering, people have been least scrutinized. As shown in Figure 1.1, Boehm observed that software development personnel are the most powerful factor in software productivity [11]. While empirical evidence supports the importance of personnel skills in software development, there are fewer similar studies in RE [25]. The qualifications of the individuals involved in the RE process highly affects the effectiveness of the whole process, but it appears that many managers' RE staffing decisions[1] are made by instinct on the basis of personal experience and anecdotal evidence rather than on the basis of scientific studies.

Agile processes, in particular, are favoring the role of people over process: Individuals and interactions over processes and tools, or as the Lean development method puts it: "find good people and let them do their own job" [70].

[1] This is not to say that the decisions are necessarily bad. Instinct based on experience is often correct.

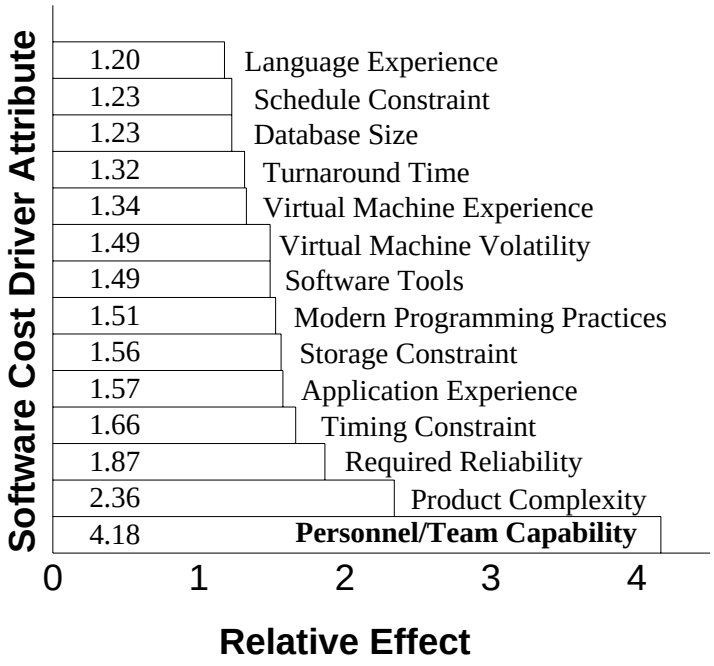


Figure 1.1: Factors Influencing Software Productivity [11]

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, 37, 51].

However, deep knowledge of the problem domain seems to lead to falling into the tacit assumption tarpit [8]. Lack of domain knowledge might, in fact, have some benefits in RE activities. One such benefit has been observed by Berry [8], namely the abilities of a domain ignorant to state his[2] ideas independent of any domain assumptions and to ask revealing questions that could lead to exposing issues that domain experts have overlooked. Domain ignorance is a good tool to surface the tacit assumptions of domain experts [27]. This surfacing can lead to the necessary shared understanding of the topics of the tacit assumptions.

Creativity plays an important role in generating requirement ideas since one of the obstacles in RE is the fact that even the customer cannot explain the exact set of requirements [31]. Brain-

[2]Although a person could be a man or woman, I have assumed any nonspecific person is a man throughout this document.

storming, for instance, has proved useful in encouraging requirements analysts to become more creative in generating requirement ideas.

Creativity basically means generating novel ideas, but creative ideas do not need to be generated from scratch. An idea that is well understood in some domain will seem new and therefore creative to people working in a different domain [57]. Therefore, ideas can be borrowed from domains different from the domain under study and they will still look creative. Also including people with different domain perspectives can help discover alternatives [27]. The more complex a problem is, the more different perspectives would be beneficial. Many innovative companies utilize some methods to foster creativity in their companies. For instance, 3M's famous 15% time rule, requires employees to put 15% of their time to work on completely new ideas. For a long time, Google followed a similar practice in which employees were allowed to put 20% of their time working on new ideas. Gmail was one of the products that came from Google's 20% program. As Lehrer [43] puts it, the success of this practice is due to the fact that innovations come when people face new areas that they are not experienced before. Lehrer puts it even better by saying that "anyone can stay creative as long as he works to maintain the perspective of the outsider". To foster creativity, he suggests borrowing ideas from other areas by immigrating employees to new areas of expertise.

1.1 Background and Related Work

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

Most software engineering 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 thesis clearly distinguishes knowledge from experience.

Berry 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 [8]. As he noted later [9], one of the earliest observations of the impact of ignorance comes from Burkinshaw's statement during the second NATO conference on software engineering in 1969 [46]:

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., 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 [18].

Kenzi et al., 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 [39, 32]. They did not explore the specific effect of an analyst's prior domain knowledge.

Ferrari et al., [24] 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., conducted a controlled experiment having two types of participants, those who have studied computer science as their university major and those who have studied something else. They observed that the general knowledge of computer science 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 [13].

RE is similar to business analysis because many of the processes involved in the two areas, such as the elicitation of information from end users, are identical [52]. Barker [4] observed about business analysis:

. . . and we shouldn't forget, that if we just stick to the same domain and the same group of insiders, innovation will be much harder to come by. . . . People will usually come up with the same ideas that they have at their last domains.

In an experiment conducted on software design, Sharp [56] 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.

During a talk given by Berry, Dim pointed out Zen's don't-know mind. Apparently, Zen advises its followers to throw all opinions and assumptions away prior to studying a concept. In other words, become ignorant in the problem domain at hand [54].

Mehrotra [44] 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. The categorization is represented in Table 1.1. Later, he showed, by mining histories reported by Dagenais et al., [16] 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.

Schooenderwoert [55] reports a case study on an agile development project for which the development team was formed of newbies lacking the technical experience needed for the project. They compared the results of this newbies' team with the industry data obtained from a number of available resources. The comparison showed that the newbies' team performed their agile software development better than an experienced team doing the same job.

Kristensson et al., [41] studied idea generation for a problem in the mobile technology domain using three types of participants: 1) advanced users who were computer science students, 2) ordinary users who were non-computer science 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. On the other hand, the ideas of professionals and advanced users were more feasible than those of ordinary users. If we consider the ordinary users to be domain ignorants due to their lack of knowledge about the problem, and the professionals and advanced users to be domain experts, we can say that even if domain ignorants do not produce feasible ideas, they can bring in valuable ideas that domain experts cannot think of.

Stuart Firestein teaches a course called *Ignorance* at the University of Columbia [26]. 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 [50]:

Activities Helped by Domain Ignorance	*Activities Not Affected by Domain Ignorance*	*Activities Hindered by Domain Ignorance*
Eliciting requirements	Learning processes/practices/technologies used	Designing and specifying software architecture
Analyzing requirements	Source/version control tasks	Reviewing software architecture
Identifying project risks	Coding simple features	Specifying requirements
Creating high-level software design	Other code oriented tasks	Validating requirements
User interface design	Automating test cases	Reusing and managing requirements
Developing black box test cases	Reviewing trace information	Managing builds of a software
Analyzing defects to find common trends	Attending courses/trainings	deployment planning
Identifying security risks	Attending formal project meetings	Risk planning/monitoring and control
Writing user manuals and release notes	Attending code/project walkthroughs	Creating low-level software design
Inspecting/Reviewing design documents	Compiling project code	Identifying design and implementation rationale
Inspecting/Reviewing user manuals	Installing and configuring development environment	Fixing bugs
Inspecting/Reviewing test plans		
Inspecting/Reviewing requirements document		
Reading product documentation		

Table 1.1: Relationship Between Software Development Activities and Domain Knowledge (Adapted from [44])

> "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 studied how scientists study things in practice [19]. 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 [42]. 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 [60, 19]. This situation is called also *distributed reasoning* [19]. 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 the individuals' judgements. On the other hand, diversity in a group led to an increase in the individuals' objectivity. Therefore, the authors suggested to further study the potential effects of diversity in a team.

Finally, during a talk on the subject of this thesis given by Berry, an attendee pointed out that General Electric, Healthcare has practiced the staffing of teams with domain ignorants for some time now.

1.2 Context

The context of the research described in this study is the RE activities 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 RE activities 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 studies described herein strived to make the distinction between domain ignorance and domain awareness sharp.

1.3 Research Questions and Hypothesis

Following the Goal-Question-Metric template [5], 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?

The major RQ can be decomposed into two specific RQs:

RQ_1 *Does a team consisting of a mix of DIs and DAs perform an RE activity more effectively than a team consisting of only DAs?*

Answering this RQ properly requires particularizing the question to several activities in RE. Two of these activities are: 1) requirement idea generation during requirements elicitation and 2) inspection of a requirements specification.

RQ_2 *Do factors other than a team's mix of DIs and DAs impact the effectiveness of the team's performing an RE activity?*

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.

The main hypothesis coming from the RQs is:

A team consisting of a mix of DIs and DAs is more effective in an RE activity 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 an RE activity.

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 an RE activity.

1.4 Desired Contributions

It is hoped that the results of this study will help RE managers in forming more effective teams for doing 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 any RE activity,
- providing some RE activities 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.

1.5 Structure of the Thesis

The rest of this thesis is structured as follows:

Chapter 2 describes the general design of the controlled experiments.

Chapter 3 describes the pilot studies conducted in order to improve the general experiment design.

Chapter 4 describes the refined design of the controlled experiments based on the lessons learned from the pilot studies.

Chapter 5 describes the first controlled experiment (E1) to test the hypotheses.

Chapter 6 describes the second controlled experiment (E2) to test the hypotheses.

Chapter 7 describes a case study conducted to corroborate the results of the experiments in an industrial context.

Chapter 8 draws the conclusions of this study and proposes future work.

2

General Experiment Design

This chapter explains the general design of a controlled experiment [71] aiming to answer the RQs. Section 2.1 lists the RE activities assumed in the experiment design. Section 2.2 describes the different variables of the experiment. Section 2.3 reexpresses the RQs and the hypotheses in terms of the variables. Section 2.4 describes the participants and the composition of teams. Section 2.5 lays out the general procedure of the experiment. Section 2.6 indicates how the results of the experiment are gathered and prepared for statistical analysis. Finally, Section 2.7 lists each threat to the validity of any experiment and the mitigation applied for this experiment.

This general experiment design is followed and particularized in the experiments described in Chapters 5 and 6.

2.1 RE Activities

The focus of the experiments was on RE tasks that generate things, such as requirement ideas. Thus, each RE task that might be the subject of an experiment is called a *generative task* (GT). Example GTs are requirements elicitation and requirements document inspection. The unit generated by a GT is called a *desired generated unit* (DGU). For the two example GTs, the DGUs are requirement ideas and defects in a requirements document. There are other GTs that could

be evaluated. For any knowledge-intensive activity, domain ignorance might be beneficial. Thus, possible GTs include architecture design, requirements prioritizing, writing test cases, and SRS writing.

Effectiveness in performing a GT is measured by counting the number of DGUs generated during the performance of the GT. The more DGUs generated, by a person or team, the more effective he or it is.

The general design of the controlled experiments done for this research is expressed in terms of GTs that generate DGUs.

2.2 Variables

There are several independent and dependent variables about each team performing a GT for a CBS in a domain.

The independent variables about a team performing a GT are:

- *GT*: the GT used as the team's task,
- *Mix of Domain Familiarities*: the team's mix of DIs and DAs, according to each team member's stated familiarity of the domain, and
- *Creativity Level*: the creativity level of the team's members, if the task requires creativity.

The dependent variables about a team performing a GT are:

- *DGU*: the DGU to be generated during the GT, and
- *Effectiveness*: the number of DGUs generated by the team.

2.3 Restated RQs and Hypotheses

This section restates the RQs and hypotheses given in Chapter 1 in terms of GTs, DGUs, and the variables of Section 2.2. Recall that the major question considered by this study is:

RQ *How does one form the most effective team, consisting of some mix of DIs and DAs, for a GT involving knowledge about the domain of the CBS whose DGUs are being generated by the team?*

RQ can be refined into two specific research questions, RQ_1 and RQ_2 that use the new terms.

RQ_1 *Does a team consisting of a mix of DIs and DAs performing a GT generate more DGUs than a team consisting of only DAs?*

RQ_2 *Do factors other than a team's mix of DIs and DAs impact the number of DGUs generated by the team's performing a GT?*

The main hypothesis coming from RQ_1 and RQ_2 is:

> A team consisting of a mix of DIs and DAs generates more DGUs while performing a GT than does 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 number of DGUs generated while the team is performing a GT.

The corresponding non-directed alternative hypothesis is:

> The mix of DIs and DAs in a team has an effect on the number of DGUs generated while the team is performing a GT.

2.4 Participants and Composition of Teams

Participants are drawn from academia and industry. Participants are assigned to teams to participate in the experiment.

To be able to adequately exercise the mix-of-domain-familiarities independent variable, I needed to build as many teams as possible of each mix of DIs and DAs. A team of fewer than three members is hardly a team in the sense of having more communication than between one pair of persons. However, the more members a team has, the many more participants are needed to achieve enough teams of each mix for statistical validity and power, and the much more money is needed to pay the participants. To gain the maximum number of teams from any number of available participants, while having enough members per team that at least one domain familiarity would have a majority in any team, I decided to have three-member teams. Three members per team allows four types of teams:

- *3I*: a team consisting of three DIs and zero DAs,
- *2I*: a team consisting of two DIs and one DAs,
- *1I*: a team consisting of one DIs and two DAs, and
- *0I*: a team consisting of zero DIs and three DAs.

Note that the name of the type of any team is an indication of the number of DIs it has.

2.5 Procedure

The steps of the experiment are depicted in Figure 2.1. Each participant is asked to fill out a questionnaire about his education level, field of study, prior industrial experience, experience in RE, and familiarity with the problem domain under study.

To find out about each participant's familiarity with the domain under study there are two options: 1) ask him to indicate his knowledge of the domain, or 2) ask him to take a placement test. This test would ask a few questions about the domain. Those who score low on the test are considered as DIs, and those who score high are considered as DAs. Although, people have different levels of domain knowledge, in the main setting of the experiment, only DAs and DIs become the main participants, since I focus on the mere presence of domain knowledge. Thus, it pays to find a domain that sharply divides the population into entirely DAs and DIs.

All participants are asked to attend a tutorial session about the domain under study in order to simulate the likely situation in a real-life GT session. In a real-life GT, all participants have at least been exposed to the problem that the CBS is to solve, having read an RFP (Request For Proposal), having attended a walk-through about the CBS, etc. Most participants are relatively ignorant about the CBS's domain, never having used a similar CBS, but a few, perhaps from the client's organization, are DAs. In the normal such situation, the DAs and DIs will all have attended the same walk-throughs.

Apart from achieving a team's configuration, the assignment of participants to teams is such that the teams are balanced, to the extent possible, by the average creativity of their members. Therefore, each participant is given a creativity test adapted from the one developed by Frank Williams [58]. All teams do the same GT with the same CBS and the same artifact for the same amount of time.

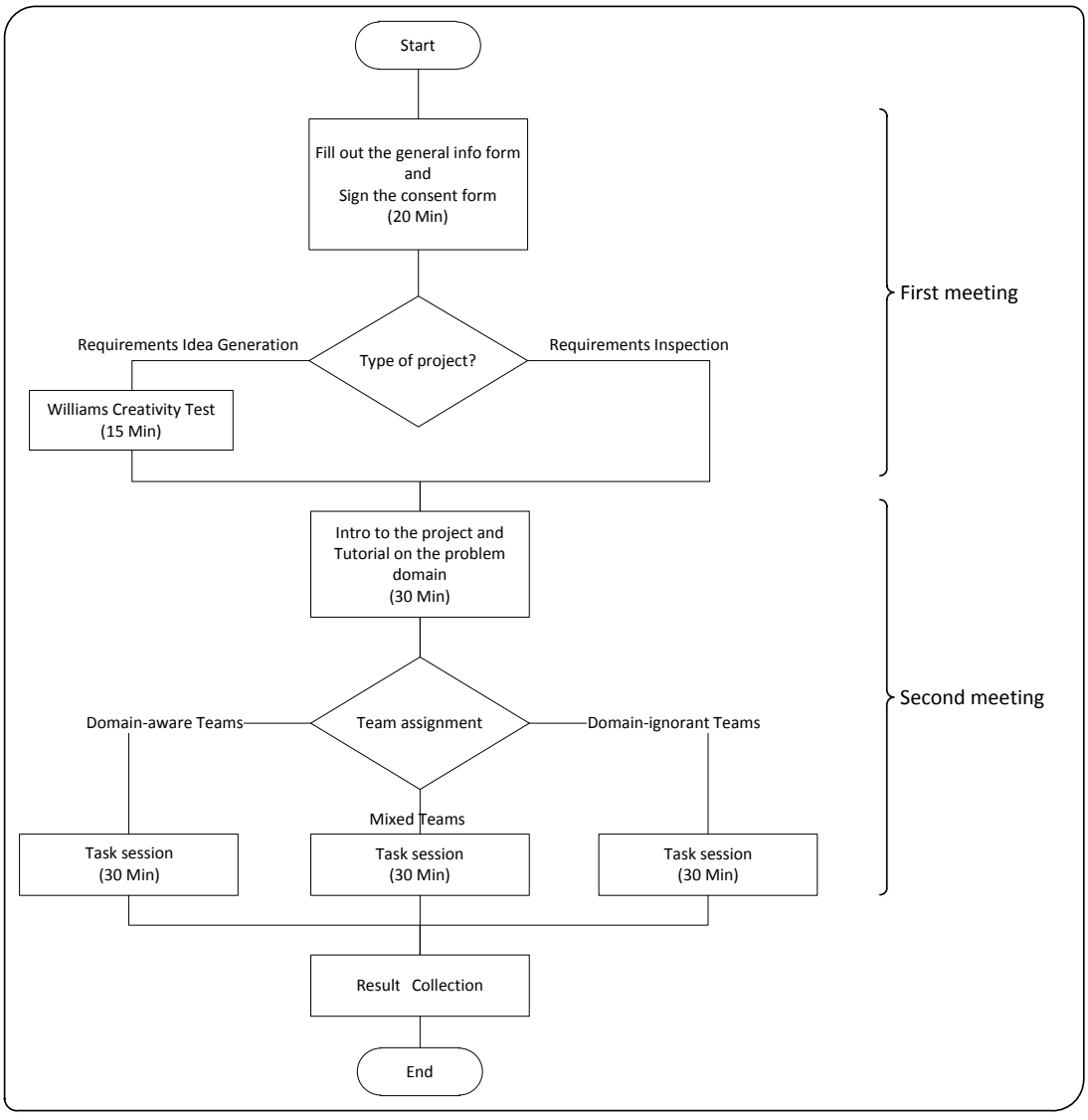


Figure 2.1: Steps of the Controlled Experiment

2.6 Analysis of the Results

It is necessary to apply both quantitative and qualitative metrics to data obtained from the experiments. The raw number of DGUs is considered as quantity. The quantitative measure makes sense for experiments on either of the two mentioned GTs, since these GTs have in common that the standard advice given to people about to begin the task is to go for quantity not quality. However, it is not enough to evaluate a team's performance based solely on the number of DGUs it has found. Therefore, it is necessary to also consider the quality of the findings of each team. The number of high quality DGUs for a team is considered as a measure of the quality of the team's DGUs. Each GT has its own definition of quality.

In general, experts in the domain of the CBS are chosen to be DGU evaluators. To eliminate any bias in classifying an idea that might arise from the evaluators' knowing the domain familiarity mix of the team from which the idea came, a list of all ideas generated by all teams would be created and sorted using the first letters of each idea. Each evaluator classifies the ideas in the full list. Once all evaluations are done, the correlations between the classifiers' classifications are computed. Then, each evaluators' classifications of each idea can be transferred to the ideas' occurrences in the individual team lists. Finally, suitable statistical methods, such as t-test and ANOVA, will be performed on the evaluated ideas.

2.7 Threats to Validity

As in any other empirical research, this study is trying to provide practical results to be 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 valuable results. In the case of real-world projects there are usually a lot of constraints imposed by managers, and experimenters are not free to modify different aspects of the project.

If it is not possible to run a controlled experiment on a real-world project, it might be possible to run a case study. A problem with every case study is the huge dependence of the results to the context of the study. Therefore, the results of case studies have low external validity due to this lack of generalizability. However, repeating a case study would increase the chance of higher external validity.

A more feasible study is a controlled experiment on students as participants with realistically sized artifacts. A problem with this type of study is the population under study. The target population for the experiment is practitioners. Therefore, results of experiments done with student

participants may not be generalizable to the target population. However, the University of Waterloo students being used as participants have some software development experience since most of them are co-op students with a few years of industrial experience. The artifacts are chosen to be as close as possible to real-world artifacts.

There are four main types of validity of the experiments that are subject to threats: conclusion, internal, construct and external [71]. The following subsections address all known threats in the study and explain the planned mitigations.

2.7.1 Threats to Conclusion Validity

Conclusion validity refers to the relation between the results and the hypotheses , i.e., the relation between the treatment used and the results of the experiment [23]:

- *Low statistical power* [1] : If the power of the results of the statistical test is low, the results are not reliable. A small sample size is a cause for low statistical power.
  - In this study, a post-hoc power analysis is performed to detect the minimum sample size required to achieve the standard minimum power value of 0.8.
- *Violated assumptions of statistical tests*: If the pre-conditions of any statistical test are not met, the results of the test are not valid.
  - In this study, for each statistical test, its pre-conditions are precisely studied so that the most valid test is chosen. Also, a statistician is consulted prior to applying the statistical analysis.
- *Fishing and the error rate*: If the results are affected by the researchers' bias toward the outcome that they seek, the results are not reliable.
  - In this study, the purpose is to confirm or refute the hypotheses. The best is done to determine outcomes totally from the data in an unbiased and fair manner.
- *Reliability of measures*: If subjective measures are used in an experiment the results might be affected by human judgement. Objective measures are obviously more reliable than subjective ones.

[1] The threat names used here are the most common names used in the literature.

  - In this study, the number of DGUs is used as an objective measure. The quality of DGUs, on the other hand, is necessarily subjective. For subjective measures, persons other than or in addition to myself are used to do any subjective tasks, e.g. evaluations.

- *Reliability of treatment*[2] *implementation*: If the procedure is applied differently for each treatment, the results comparing the treatments are not reliable.
  - In this study, the same procedure is applied for each treatment.

- *Random irrelevancies in experimental setting*: If external elements disturb the experiment sessions, the results might be affected.
  - In this study, every effort is made to prevent any interruptions during the experiment.

- *Random heterogeneity of participants*: If the participants differ a lot, then other, personal, factors might affect the results more than the factor being studied in the experiment.
  - In this study, every effort is made to use a homogeneous group of participants, e.g. undergraduate students. However, that decreases the external validity since students are not the target participants of this experiment, and it reduces the generalizability of the results.

2.7.2 Threats to Internal Validity

Internal validity refers to how well confounding factors within the experiment design are controlled so that the outcome of the experiment shows the causal relationship between the treatment and outcome.

- *History*: If events outside of an experiment affect participants' behavior, especially in case of large scale events, such as natural disasters [69], the results might not be reliable.
  - In this study, as far as I know, no major event happened that could have possibly affected the participants.

[2]Treatment is the statistical term used to refer to each combination of factor levels whose effect is being studied. In this study, four treatments are being compared based on different combinations of teams.

- *Maturation*: If the experiment session is long, the passage of time might affect the participants' performance. One common aspect of this threat is that the participants will learn how to do the experiment's task during the experiment, thus affecting their performance during the task.
  - In this study, the duration of the experiment is kept short. Participants are more or less familiar with both GTs and since the duration of the tasks are short, about 30 minutes, there is a little chance of learning to do the task any better in this short period of time.
- *Repeated testing (Learning effect)*: If any participant is used in more than one experiment session, he might respond differently, because he has learned from the first time.
  - In this study, no participant is used more than once for the experiment.
- *Instrument change*: If the instruments of the experiment, e.g. RFPs, specifications, test cases, user manuals, etc. change between different experiment sessions, the results might be affected.
  - In this study, the same instruments are used in every experiment session.
- *Regression toward the mean (test unreliability)*: If the participants are asked to take a pre-test to be classified to different experimental teams, their responses might not reflect the attribute(s) they have been tested for. Regression toward the mean occurs when participants are classified on the basis of extreme scores, i.e., one far away from the mean [69]. The more reliable the measures are, the more reliable the results are.
  - In this study, participants are asked to take a pre-test, the Williams Test. Since the test is a standard test being used in the schools of the US, I assumed the measures to be well defined.
- *Selection*: If a participant has joined voluntarily, he might behave differently from involuntary participants [71].
  - In this study, all participants join voluntarily.
- *Mortality*: If a participant drops out, his data might be meaningless.
  - In this study, if a participant drops out during the course of an experiment, the data from his team are ignored.

- *Interaction with selection*: There are three types of interactions that might affect internal validity [14]:
    1. *Selection-maturation*: If the participants learn at different rates, the results might be affected [38].
        - In this study, the participants are more or less knowledgeable with the GTs. There is not much they can learn during a half-hour session.
    2. *Selection-history*: If participants have different experiences, their performance and therefore the results might be affected.
        - In this study, experience is considered as an independent variable. Also, factor analysis is used to check the independence of the independent variables.
    3. *Selection-instrumentation*: If the experiment instruments are different for different participants, the results are not reliable.
        - In this study, the same instruments are provided to each team.
- *Ambiguous temporal precedence*: If it is not clear which variable occurred first, a conclusion cannot be made from the results.
    - In this study, there is no temporal precedence between the variables, neither in the independent variables nor in the dependent variables.
- *Diffusion or imitation of treatments*: If one participant observes other participants' progress and begins to adjust his behaviour based on what he has observed, then the results might be affected.
    - In this study, each team's experiment is run separately. Alternatively, if the teams can be kept far enough apart that they cannot hear each other, multiple experiment sessions can be held at the same time.
- *Compensatory equalization of treatments*: If a participant knows about how he differs from other participants, his performance might be affected.
    - In this study, the experiment conducted with each team is totally separate from the experiments conducted on other teams, no teams knows how it differs from other teams, and no participant knows how he differs from his teammates.
- *Compensatory rivalry/Resentful demoralization*: If two participants have different treatments, the one who has the less desirable treatment might act differently from his norm.
    - In this study, the participating teams are kept uninformed of the different treatments.

2.7.3 Threats to Construct Validity

Construct validity refers to the relation between the results obtained from the experiment and the theory behind the experiment.

- *Inadequate preoperational explication of constructs*: If the theory is not clear enough or if clear measures are not defined, it is difficult to analyze the results.
  - In this study, both quantitative and qualitative measures are defined to compare the effectiveness of different teams.
- *Mono-operation bias*: If the number of independent variables or participants are very low, then the theory might not be tested well by the experiment.
  - In this study, the best is done to find all relevant independent variables and to have enough participants.
- *Mono-method bias*: If only one measurement method is used, the risk of unreliable results will be increased.
  - In this study, multiple measurement methods are used, both quantitative and qualitative.
- *Confounding constructs and levels of constructs*: If only the presence or absence of some independent variable is considered when actually its level affects the results, the results are not reliable. For example, the simple presence or absence of industrial experience might not be enough detail about the industrial experience independent variable for good results, but the level of industrial experience might better explain any strangeness in the results.
  - In this study, most independent variables are defined as having more than just "present or absent" as levels.
- *Interaction of different treatments*: If participants are shared between different treatments, then it is not possible to say whether their work has been affected by a single treatment or a combination of treatments.
  - In this study, no participant is shared between different treatments.
- *Interaction of testing and treatment*: If a measurement method is applied during the experiment, it might affect the participants' performance. E.g. if the number of errors made by participants are counted during the experiment, then the participants will be more cautious about the errors they make.

  - In this study, there is no measurement method applied during the course of the experiment and the participants are not aware of the measurement methods. Although, they might guess that the number of DGUs would be a measure since they are told that for their GT, "quantity is more important than quality". However, this advice is normal for the GT.

- *Restricted generalizability across constructs*: If the settings of the experiment affect one dependent variable positively while affecting other variables negatively, the results of such a study are hard to generalize.
  - In this study, the settings are kept as simple as possible. Thus, there is little chance that they affect any dependent variable, let alone positively or negatively.

- *Hypothesis guessing*: If a participant can guess the hypothesis of the experiment, he might alter his behavior either positively or negatively toward the hypothesis.
  - In this study, the hypothesis is not given out. Although, it might be guessed.

- *Evaluation apprehension (Hawthorne effect)*: If a participant knows that he is being evaluated, he might modify his behaviour from his norm.
  - In this study, since some participants are drawn from an RE course, they were assured that their performance during the experiment and the outcomes they produce neither affect their course grades. All participants are also told that the remuneration they receive at the end of the experiment is not affected by their performance.

- *Experimenter expectancies*: If the results are not what the experimenter is looking for, the data obtained from the experiment might be affected by the experimenter's bias toward the result he is expecting.
  - In this study, for quantitative measures, multiple DGU evaluators review the data separately. Blind review is also applied so that no evaluator knows which team has produced which data.

2.7.4 Threats to External Validity

External validity refers to the generalizability of the results of an experiment to contexts other than the experiment's context. The origin of external validity threats come from the interactions of treatments with people involved in the experiment, place, and time of the experiment.

- *Interaction of selection and treatment (Population validity)*: If the experiment is applied on only a specific type of participants, the results are hardly generalizable to other categories of people. Sometimes participant population does not represent the goal population of the experiment. The goal of most empirical studies in software engineering is to draw conclusions valid for practitioners. However, since convincing companies to provide experimenters with the company's practitioners is barely feasible, these kinds of studies are usually performed on students. 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, Host et al., [35] conducted some experiments using both students and professionals as participants and showed that the student participants did perform as well as the professional participants with no major difference, although they emphasize that their student participants possessed a good knowledge of software engineering. They did this study to identify the factors affecting the lead time of software development projects. Their results can be generalized based on the selection of participants chosen and the familiarity of the participants with the performed task.

  - In this study, the goal population is practitioners involved in knowledge-intensive activities. Due to the low availability of practitioners, students are used as participants. However, most undergraduate students participating in the experiments are co-op students, who have had a few years of software development experience.

- *Interaction of setting and treatment (ecological validity)*: If the settings of the experiment do not represent a real industrial setting, the results are not generalizable to other settings.

  - In this study, the experiment is done on a real world problem that needs to be solved. Moreover, the size of the GT performed in the experiment is realistic, in that the time allowed for the GT is the same as that which is used in an industrial setting.

- *Interaction of history and treatment*: If the settings of the experiment are tied to a specific period of time, the results are not generalizable to other periods of time.

  - In this study, the settings are not tied to a specific period of time.

3
Pilot Studies

Controlled experiments are probably the most effective method by which to validate a hypothesis. Therefore, 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 were conducted in order to identify issues in the experiment and improve the design. Although the purpose of pilot studies is to clarify aspects of the experiment and therefore the results will not be of any importance, the whole process was carried out in these studies so that defects in the design of the experiment could be identified. Section 3.1 describes the first pilot study, while Section 3.2 describes the second pilot study.

3.1 Pilot Study I

The first study was performed on students taking an advanced graduate course on requirements engineering. The first difficulty was deciding on a CBS and its domain for which there would be enough DIs and DAs. The way the domain for this pilot was selected was to ask each participant to indicate his knowledge of 34 different domains in a Likert-scaled questionnaire (See Appendix A). The plan was to choose the domain whose numbers of DIs and DAs were the closest to each other. The CBS chosen for the first pilot was a requirements tracing tool. A features list of

Team	*Distribution*		*Number of generated requirements*	*Creativity Score*
	DAs	DIs		
0I	3	0	28	68
1I	2	1	36	76
2I	1	2	30	81
3I	0	3	33	63

Table 3.1: Results of Pilot Study I

a requirements traceability tool was prepared based on a few existing tools' descriptions. The experiment was conducted on four teams of three.

The results of this pilot study are presented in Table 3.1. The all-DA, *0I*, team surprisingly produced the least number of requirements, while the all-DI, *3I*, team generated a rather large number of requirements; but the best result obtained by the *1I* team containing two DAs and only one DI. The results do support the hypothesis, but of course, there are not enough data to reliably draw any conclusions.

Apart from my understanding of the issues involved in the experiment, I also gathered feedback from participants after the experiment to see what they thought of the experiment procedure.

The lessons learned from this pilot study are:

- *Finding a suitable problem*: It is not easy to find a problem of which one half of the participants are knowledgeable and the other half are not. So, it is necessary to pay significant attention to find a problem that well divides the participants.
- *Not enough time for lively discussion*: The participants complained about the allowed time. Apparently the task was too difficult for an hour-long session. So, it is necessary to simplify the task to be doable in the allotted time.
- *Too detailed features list*: There were also concerns about the features list, which was given to participants to be used as a starting point to extract requirements out of. It appears that the list was so detailed that some features were actually considered as requirements by different teams. So, it is necessary to provide higher-level features in the feature list.
- *Other factors need to be considered in analysis of the results*: This pilot showed also that to guard against misidentification of the cause of the observed effects, it was necessary to consider factors other than domain knowledge that might influence the effectiveness of

requirement idea generation. These other factors include RE experience, industrial experience, educational background, and creativity. In particular, creativity is widely assumed to play an important role in any idea generation task, such as brainstorming.

- *Evaluating results*: Each team used a different layout, making it difficult to compare the results. Giving a standard layout would make the evaluation easier. For this study, I considered only the number of requirements generated as the effectiveness measure. It became clear that also the quality of the requirements generated should have been considered.

3.2 Pilot Study II

Pilot study II consisted of a controlled experiment on the requirements inspection of a user manual, which can serve as a requirements specification [7]. The user manual was intended for an online admissions system used in the School of Computer Science at the University of Waterloo. Only six subjects were available. So, three teams of two were formed. They were asked to study the user manual and report as many requirements' defects as they could find in it. The participants were given a description of the characteristics of a good requirements specification to use as a baseline [6].

Participants who had used the system to apply to the university were considered as DAs and undergraduate students who were not familiar with the system were considered as DIs. Teams were made of all possible combinations: a team of two DAs, a team of two DIs, and a team of one DA and one DI. The results, shown in Table 3.2, appeared to be almost identical for the mixed and all-DA teams, while the all-DI team did poorly in respect to the number of defects detected. The results were triply useless. They did not say anything with respect to the hypothesis, the team sizes were too small, and there were too few teams. Nevertheless, lessons were learned.

Team	***Distribution***		***Number of requirement defects found***
	DAs	DIs	
Team 1	1	1	53
Team 2	2	0	58
Team 3	0	2	33

Table 3.2: Results of Pilot Study II

The lessons learned from this pilot study are:

- *Quality of the results should also be assessed*: The quantity of the defects found by each team is not enough. Some defects were not actually defects in requirements, while some had conflicts or overlaps with other defects. So, it is necessary to define quality measures for the DGUs.
- *Small team size*: This pilot study confirmed the supposition that teams of size less than three are really not teams in their interactions. A team of one has only one interaction path. A team of three has three interaction paths that must be juggled by the team members. So, a team's size must be at least three.
- *DIs were not real DIs*: I assumed that students who have not used the graduate admission system are DIs. However, I learned that undergraduate students had applied to the university through a similar undergraduate admission system. So, it is necessary to find a problem domain that clearly and completely partitions the population of participants.

3.3 Conclusion

The lessons learned from the pilot studies described in this chapter are used to refine the general experiment design described in Chapter 2. The refined experiment design is described in Chapter 4.

4

Refined Experiment Design

The general experiment design presented in Chapter 2 was applied in the two pilot studies presented in Chapter 3. The lessons learned from the pilot studies are applied to the general design to get a refined version which is applied in the controlled experiments. Section 4.1 introduces the selected domain for the experiments. Section 4.2 details the evaluation procedure of gathered data. Section 4.3 lists the refined independent and dependent variables. Section 4.4 reworks the hypotheses to reflect the refined variables. Section 4.5 refines the procedure of the experiment. Section 4.6 discusses the statistical analysis methods to be applied on the results.

4.1 Domain Selection

The main lesson learned from the pilot studies was that finding a suitable CBS with a suitable domain to use in experiments was critical. Domains in computer science or university administration were too familiar to the participant population of university students that are competent in computer science. For such domains, it is hard to build teams with DIs. It was clear that I needed a domain outside computer science, 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 computer science 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 I 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 me 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, Berry and I 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.

4.2 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, I considered also the quality of their generated ideas. Based on the characteristics of a good requirement in the IEEE 830 Standard [6], I 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.

I decided to use myself and Berry, as experts in the BDWP domain, as idea evaluators. To eliminate any bias in classifying an idea that might arise from the evaluator's knowing the domain

familiarity mix of the team from which the idea came, I 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.

4.3 Variables

Based on the lessons learned from the pilot studies described in Chapter 3, the variables of the experiment are extended.

4.3.1 Independent Variables about a Team

Table 4.1 shows the complete sets of independent and dependent variables. Two new independent variables, REXP and IEXP, have been added to the independent variables about a team. For the sake of statistical analyses, each of the independent variables are converted into a numerical value:

- *Mix of Domain Familiarities (MIX)*: a team's value is the number of DIs it has, which is the number in the name of its type, *0I*, *1I*, *2I*, or *3I*.
- *Creativity (CR)*: the teams' mean Williams test scores were transformed linearly into three values, 1, 2, or 3.
- *Industrial Experience (IEXP)*: a team's mean number of years of industrial experience which is divided into three values: 1) none or less than a year, converted to none, 2) one to two years, and 3) more than two years.
- *RE Experience (REXP)*: a team's presence of a member with RE experience which is divided to two values: 1) some RE experience, and 2) no RE experience.

4.3.2 Dependent Variables about a Team

The dependent variables about a team's performing a GT are refined based on the classification of the requirement ideas described in Section 4.2:

Name	*Independent Variable about a Team*	*Values*
MIX	Mix of domain familiarities	0, 1, 2, 3
CR	Average creativity score level	Low, Medium, High
REXP	Presence of RE experience	None, Some
IEXP	Average industrial experience	None, 1–2 years, More than 2 years
Name	***Dependent Variable about a Team***	***Values***
RAW	Raw number of ideas	Numeric
AVG_R	Average number of relevant ideas	Numeric
AVG_F	Average number of feasible ideas	Numeric
AVG_I	Average number of innovative ideas	Numeric

Table 4.1: Variables of the First Experiment

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

With these specific variables, effectiveness of a team in requirement idea generation is measured by counting the number of each type of ideas generated during the performance of the GT.

4.4 Hypotheses

To test the main hypothesis described in Section 2.3, the following sub-hypotheses, labeled H_{X1}, and corresponding null sub-hypotheses, labeled H_{X0}, are defined:

$\boldsymbol{H}_{MIX_1}$: The effectiveness of a team in requirement idea generation is affected by the team's mix of domain familiarities.

$\boldsymbol{H}_{MIX_0}$: The effectiveness of a team in requirement idea generation is not affected by the team's mix of domain familiarities.

$\boldsymbol{H}_{CR_1}$: The effectiveness of a team in requirement idea generation is affected by the team's creativity level.

$\boldsymbol{H}_{CR_0}$: The effectiveness of a team in requirement idea generation is not affected by the team's creativity level.

$\boldsymbol{H}_{REXP_1}$: The effectiveness of a team in requirement idea generation is affected by the team's RE experience.

$\boldsymbol{H}_{REXP_0}$: The effectiveness of a team in requirement idea generation is not affected by the team's RE experience.

$\boldsymbol{H}_{IEXP_1}$: The effectiveness of a team in requirement idea generation is affected by the team's industrial experience.

$\boldsymbol{H}_{IEXP_0}$: The effectiveness of a team in requirement idea generation is not affected by the team's industrial experience.

4.5 Procedure

As described in Figure 4.1, the experiment is divided into two parts. In the first part, each participant is asked to fill out a questionnaire about his education level, RE experience, industrial experience, and familiarity with the bidirectional word processing domain. Each is asked also to take the Williams creativity test [58] to detect the presence of significant differences in personal creativity. The gathered creativity scores will be used to balance the teams based on creativity. The information gathered in the participants' first parts allows forming teams. Each team has one particular needed mix of DIs and DAs, and each is invited to attend a second part.

As described in Section 2.5, in the second part, each team attends a one-half hour lecture about reading bidirectional text. The lecture is 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 describes possible ways of storing and displaying bidirectional text in existing word processors.

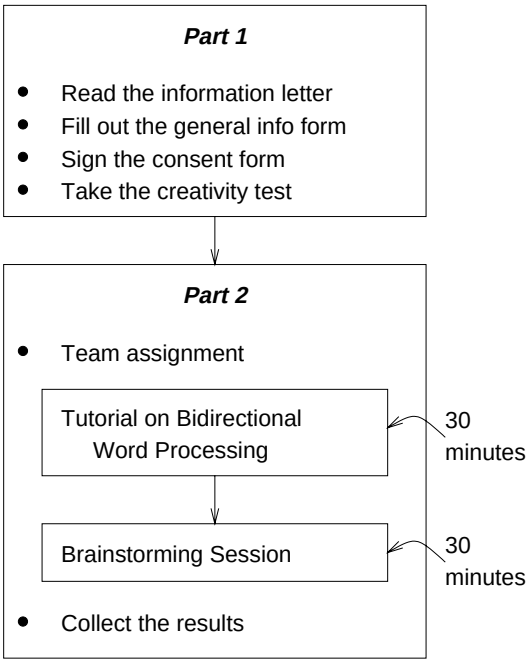


Figure 4.1: Steps of the Controlled Experiment (Refined)

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

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

4.6 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, I 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 [20]. 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) [68]. 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 ANOVA are not met, a non-parametric substitute for 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.

5

Controlled Experiment I

As discussed earlier, the main research vehicle of this study is controlled experiments to compare the effectiveness of different mixes of DI and DA participants in teams performing requirement idea generation for a BDWP. This chapter explains the first of two controlled experiments [47], E1, conducted using the refined experiment design presented in Chapter 4. Section 5.1 describes the participants used in E1. Section 5.2 describes the data gathered from the participating teams. Section 5.3 determines the outliers of E1. Section 5.4 details the statistical analysis conducted on the gathered data. Finally, Section 5.5 discusses the results of the statistical analysis and concludes the chapter.

5.1 Participants

A majority of the participants for the first experiment were recruited from a fourth-year undergraduate course titled "Software Requirements and Specification" that introduces students to RE. Thus, these participants have some RE experience. Participants were recruited also from outside the course, but nevertheless in computer science or a related discipline.

5.2 Gathered Data

Table 5.1 shows data about the independent variables of the teams. There are unequal numbers of teams of the different mixes, affecting the statistical results.

Mix of Teams	***Number of Teams***	***Creativity***	***RE Experience***	***Industrial Experience***
		Mean	Mean	Mean
3I	9	69.11	0.89	3.06
2I	4	71.75	0.75	3.33
1I	3	70.67	1.00	1.33
0I	3	71.33	1.00	2.00

Table 5.1: Data about the Teams

As planned in Section 4.2, after all the idea generation sessions were done, one list of all ideas generated by all teams was created, and each of the two domain-expert classifiers classified all ideas in this list. When classifications of the ideas were finished, the correlations between the classifiers' classifications were computed using the Pearson test[1]. As is shown in Table 5.2, the classifications of the two classifiers have a strong correlation ($p = 0.000$).

Even though the correlations between the classifiers' classifications are high, it is not clear how any team's ideas should be classified. Taking the average of classifications is not well defined. Instead, I decided to use as a team's measure of the ideas of any type the number of ideas it generated of that type. Therefore, it is reasonable to use as a team's number of ideas of any type, the average of the numbers of ideas of that type as determined by the different classifiers. For instance, if one classifier classified n of Team X's ideas as relevant, and the other classifier classified m of Team X's ideas as relevant, the number of relevant ideas of Team X used in the analysis is $(n+m)/2$.

	Relevant Ideas	***Feasible Ideas***	***Innovative Ideas***
Pearson Correlation	0.975	0.956	0.892
Significance	0.000	0.000	0.000

Table 5.2: Correlation between the Classifiers' Classifications of Ideas

The classified data about the teams' ideas are summarized in Table 5.3.

[1] Statistical analyses are performed using SPSS.

Type of Teams	*No. of Raw Ideas*		*No. of Relevant Ideas*		*No. of Feasible Ideas*		*No. of Innovative Ideas*	
	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
3I	25.33	15.66	7.94	5.14	6.56	4.30	1.11	1.39
2I	23.50	7.14	7.00	1.58	6.25	1.32	0.88	0.63
1I	27.00	12.49	8.67	4.04	4.33	2.25	1.17	2.02
0I	22.33	11.68	6.83	3.88	3.33	2.02	1.00	1.73

Table 5.3: Data about the Ideas Generated by the Teams

5.3 Outliers

Boxplots were used to graphically expose any outliers[2]. As is shown in Figure 5.1, the only outlier is the team with RAW = 91, that generated 91 raw ideas. An examination of its ideas shows an extremely large number of irrelevant ideas. They had taken "quantity over quality" too literally. Notice that there were no outliers in AVG_R, AVG_F, and AVG_I. Therefore, I removed only this team from consideration in the analysis.

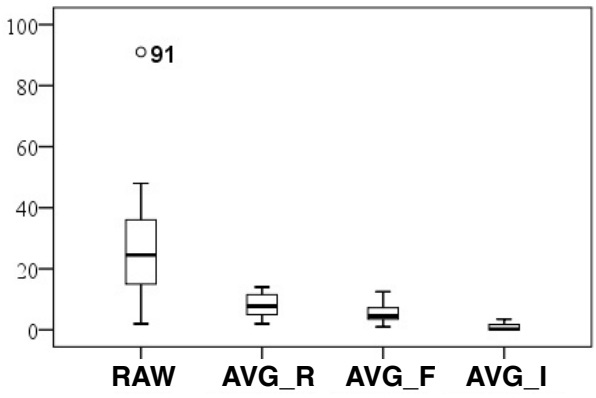


Figure 5.1: Numbers of Ideas Generated for each Classification

[2]Unusually small or large values in the dataset

Dependent Variable	***Shapiro-Wilk***		
	Statistic	df	p
RAW	.961	19	.589
AVG_R	.912	19	.081
AVG_F	.921	19	.119
AVG_I	.777	19	.001

Table 5.4: Test of Normality of the Dependent Variables

Dependent Variable:	*Levene Statistic*	*df1*	*df2*	*p*
RAW	2.862	14	4	.160
AVG_R	1.678	14	4	.328
AVG_F	2.441	14	4	.201

Table 5.5: Levene Test of Equality of Variances

5.4 Statistical Analysis

As discussed in Section 4.6, the preconditions of ANOVA are checked:

1. *Dependent variables are normally distributed* : As is shown in Table 5.4, the Shapiro-Wilk test showed that RAW, AVG_R, and AVG_F are all normally distributed. However, AVG_I is not normally distributed. I tried several methods of transforming AVG_I but none yielded a normal distribution.

2. *Homogeneity of variances*: The assumption of homogeneity of variances is met in the data since, as shown in Table 5.5, Levene test results are not significant ($p > 0.05$).

3. *All observations are independent*: As discussed in Section 4.6 the observations about the teams are totally independent of each other.

An ANOVA was applied to the dependent variables that meet the prerequisites of the ANOVA. These variables are RAW, AVG_R, and AVG_F. The test results are described in the following subsection.

Independent Variable	*RAW*				*AVG_R*				*AVG_F*			
	F	*p*	f^2	*P*	*F*	*p*	f^2	*P*	*F*	*p*	f^2	*P*
MIX	.165	.915	.011	.068	8.675	.032	.319	.816	13.486	.015	.449	.941
CR	.921	.469	.048	.146	3.918	.114	.159	.459	.984	.449	.051	.153
IEXP	.563	.609	.031	.107	10.089	.027	.331	.833	4.381	.098	.173	.499
REXP	.145	.722	.008	.063	.173	.699	.009	.65	.035	.861	.002	.53

* F is the value of the F-test, p is the probability of the F-test, f^2 is Cohen's effect size, P is the post-hoc power.

Table 5.6: ANOVA Summary

5.4.1 ANOVA Results

Unless explicitly stated otherwise, each p-value in the following results is given under the assumption that $\alpha = 0.05$. That is, if $p < 0.05$, for a result, the result is significant.

Impact of MIX:

Figure 5.2(a) shows that the means of the AVG_R generated by teams is partially positively correlated with the teams' MIX. The ANOVA test shows a significant effect ($p = 0.032$) of MIX on the means of the AVG_R generated by teams.

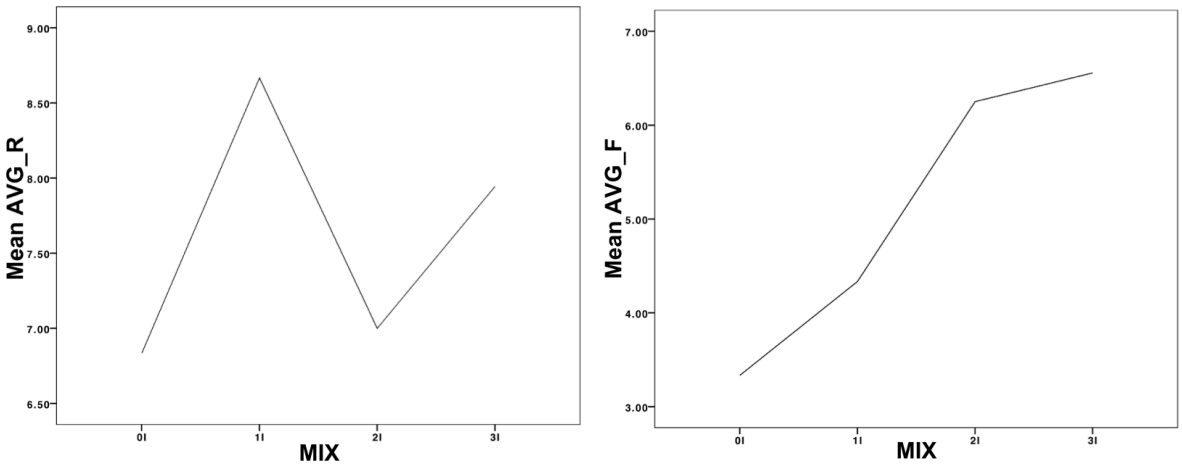
Figure 5.2(b) shows that the means of the AVG_F generated by teams is positively correlated with the teams' MIX. The ANOVA test shows a significant effect ($p = 0.015$) of MIX on the means of the AVG_F generated by teams.

Impact of IEXP:

Figure 5.3(a) shows that the means of the AVG_R generated by teams is negatively correlated with the teams' IEXP. The ANOVA test shows a significant effect ($p = 0.027$) of IEXP on the means of the AVG_R generated by teams.

Figure 5.3(b) shows that the means of the AVG_F generated by teams is not correlated with the teams' IEXP. The ANOVA test shows no significant effect ($p = 0.098$) of IEXP on the means of the AVG_F generated by teams.

Common sense says that there should be a positive correlation between the quality of requirement ideas generated by a team and the IEXP of the team. Nevertheless, the results



(a) AVG_R (b) AVG_F

Figure 5.2: Mean Number of Ideas vs. MIX

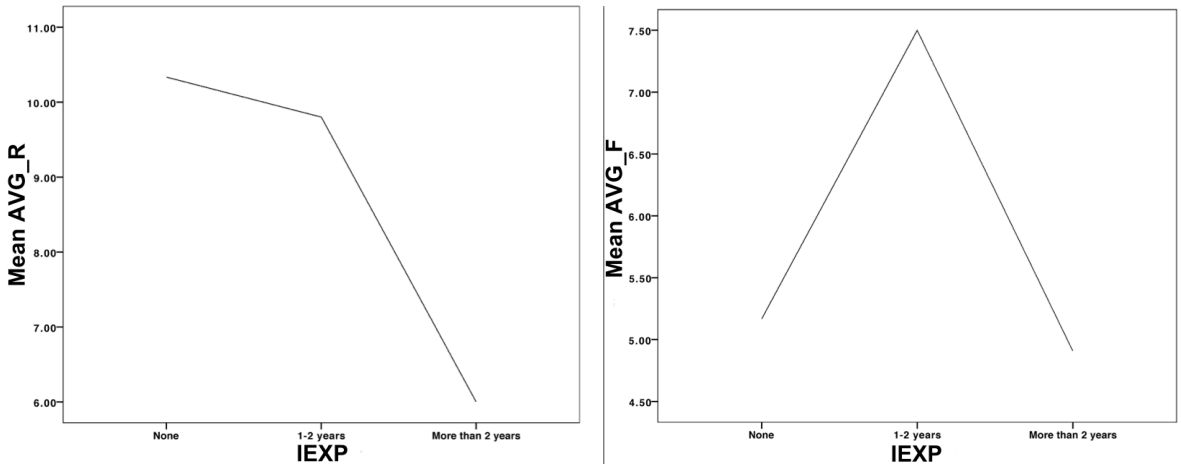
Interaction	*Dependent Variable*	*Mean Square*	*F*	*p*
	RAW	127.841	.451	.538
CR-MIX	AVG_R	37.278	14.460	.019
	AVG_F	37.278	8.212	.013
	RAW	114.019	.403	.560
CR-IEXP	AVG_R	134.082	52.007	.002
	AVG_F	106.736	52.146	.002

Table 5.7: Interactions: CR-MIX And CR-IEXP

show that for the means of the AVG_R generated by teams, there is no such correlation. As is shown in Figure 5.3(a), each team with an IEXP value of 1–2 years performed a bit worse than any team with an IEXP value of "none". Team performance in requirement idea generation drops dramatically for the teams with an IEXP value of more than 2 years.

Other Factors:

Figure 5.4(a) shows the interaction between the means of the AVG_R generated by teams for the four values of MIX, and the three values of CR. For the *0I* and *2I* teams, it appears that the higher the value of CR a team has, the lower is the AVG_R generated by the team. On the other hand, *3I* teams with a higher value of CR had a higher



(a) AVG_R (b) AVG_F

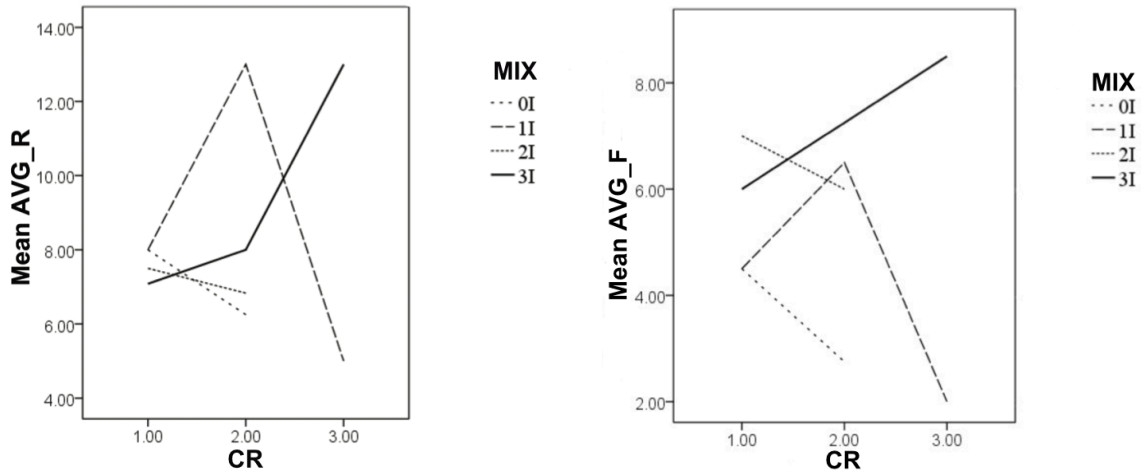
Figure 5.3: Mean Number of Ideas vs. IEXP

AVG_R generated by the team. The *1I* teams do not show any consistent behavior.

Figure 5.4(b) shows the interaction between the means of the AVG_F generated by teams for the four values of MIX and the three values of CR. *3I* teams with a higher CR had a higher AVG_F, while *0I* and *2I* teams with a lower CR had a higher AVG_F. The *1I* teams do not show any consistent behavior.

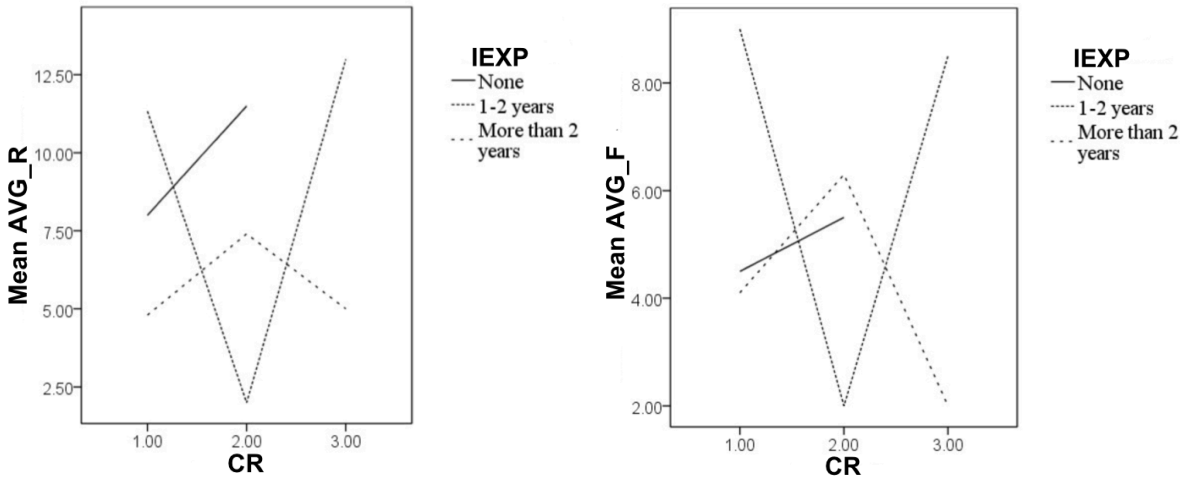
Figures 5.5(a) and 5.5(b) show the interaction between the means of the AVG_R and of the AVG_F generated by teams for the four values of MIX and the three values of CR, respectively for the teams. The figures show nothing very interesting except that, as with no IEXP, an increase in the CR for a team increases the AVG_R and the AVG_F generated by the team.

Common sense says that there should be a positive correlation between the quality of requirement ideas generated by a team and each of the CR and the REXP of the team. Nevertheless, neither CR nor REXP showed any significant effect in the ANOVA test. However, as shown in Table 5.7, the ANOVA test reveals that the interaction between the CR and the MIX of a team and the interaction between the CR and the IEXP of a team are statistically significant for the means of the AVG_R generated by the teams and of the AVG_F generated by the teams.



(a) AVG_R (b) AVG_F

Figure 5.4: Mean Number of Ideas vs. MIX vs. CR



(a) AVG_R (b) AVG_F

Figure 5.5: Mean Number of Ideas vs. IEXP vs. CR

5.4.2 Pairwise Comparisons

For pairwise comparisons, since the numbers of teams with different mixes of domain familiarities are not equal, the Least Significant Difference (LSD) method is used [45].

Carrying out the LSD method on the MIX values of teams shows that:

- the means of the AVG_R generated by a team, shown in Figure 5.2(a), differ significantly ($p = 0.048$) for the *3I* and *2I* teams.
- the means of the AVG_R generated by a team, shown in Figure 5.2(a), differ rather significantly ($p = 0.053$ with $\alpha = 0.10$) for the *3I* and *0I* teams.
- the means of the AVG_F generated by a team, shown in Figure 5.2(b), differ significantly ($p = 0.008$) for the *3I* and *0I* teams.
- the means of the AVG_F generated by a team, shown in Figure 5.2(b), differ rather significantly ($p = 0.055$ with $\alpha = 0.10$) for the *2I* and *0I* teams.
- the means of the AVG_F generated by a team, shown in Figure 5.2(b), differ rather significantly ($p = 0.018$) for the *3I* and *1I* teams.

Carrying out the LSD method on the IEXP values of teams shows that:

- the means of the AVG_R generated by a team, shown in Figure 5.3(a), differ significantly ($p = 0.026$) for the teams with an IEXP value of more than 2 years and the teams with an IEXP value of "none".
- the means of the AVG_R generated by a team, shown in Figure 5.3(a), differ significantly ($p = 0.025$) for the teams with an IEXP value of more than 2 years and the teams with an IEXP value of 1–2 years.
- the means of the AVG_F generated by a team, shown in Figure 5.3(b), differ rather significantly ($p = 0.089$ with $\alpha = 0.10$) for the teams with an IEXP value of "none" and the teams with an IEXP value of 1–2 years.
- the means of the AVG_F generated by a team, shown in Figure 5.3(b), differ rather significantly ($p = 0.061$ with $\alpha = 0.10$) for the teams with an IEXP value of more than 2 years and the teams with an IEXP value of 1–2 years.

Independent Variable	*p*
MIX	.966
CR	.996
IEXP	.240
REXP	.749

Table 5.8: Results of Applying Kruskal-Wallis Test on AVG_I

5.4.3 Non-Parametric Test on AVG_I

Recall that the values of the AVG_I generated by teams, as one of the dependent variables, did not meet the normal-distribution requirement for the use of an ANOVA. Moreover, an attempt to transform these data into normally distributed data did not work. So, it was necessary to apply a non-parametric test, the Kruskal-Wallis test.

The results of the Kruskal-Wallis test are presented in Table 5.8. The test does not show any significant effect for any of the four independent variables on the medians of the AVG_I generated by teams.

5.5 Conclusion

This chapter describes a controlled experiment to test the hypothesis that a team consisting of a mix of DIs and DAs is more effective in requirement idea generation than a team consisting of only DAs. The results, although not conclusive, show some support for accepting the main hypothesis. The data were analyzed also to determine the effect of the participants' creativity, industrial experience, and RE experience on their teams' requirements elicitation effectiveness.

The results of analysis of the experimental data are summarized as follows:

MIX The MIX of a team significantly affects the quality of the ideas generated by the team. Therefore, hypothesis H_{MIX_1} is strongly accepted.

CR The CR of a team does not significantly affect the quality of the ideas generated by the team. However, each of:

1. a team's CR interacting with its MIX, and
2. a team's CR interacting with its IEXP

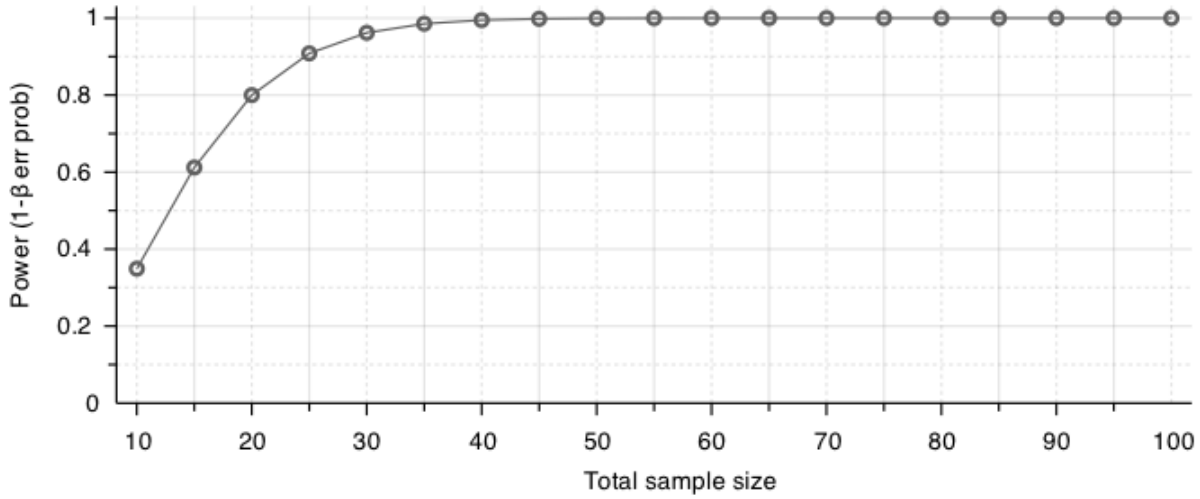


Figure 5.6: Statistical Power as a Function of Sample Size

affects the effectiveness of the team's requirement idea generation.

Therefore, it is not certain that a team's CR does not have any impact on the team's effectiveness in requirement idea generation. Therefore, hypothesis H_{CR_0} is very weakly accepted.

REXP The REXP of a team does not affect the quality of the ideas generated by the team. Therefore, hypothesis H_{REXP_0} is accepted.

IEXP The IEXP of a team does affect the quality of the ideas generated by the team. Therefore, hypothesis H_{IEXP_1} is accepted.

From these results, considering the threats, the main hypothesis, that a team consisting of mix of DIs and DAs is more effective in requirement idea generation than a team consisting of only DAs, appears to be weakly supported.

Regarding the threats to validity, I assumed that $\beta = 0.20$. Thus the minimum acceptable power in E1 is 0.80. A post-hoc power analysis was performed to detect the minimum sample size that would have yielded power of at least 0.80. Figure 5.6 shows that a sample size of 35–40 yields maximum power, considering the α, β, f^2, and dependent variables of E1. Also, the imbalance of the number of teams per MIX value accounts for some of the lost of power, and therefore, it is necessary to run E1 again with more teams of each type.

6

Controlled Experiment II

The experiment (E2) described in this chapter is a replicating extension of the experiment (E1) described in Chapter 5 to provide more data points. The main problems with E1 were the small number of teams and the unbalanced numbers of teams with each mix of domain familiarity. The small number of teams reduces the strength of any statistical results. The lack of balance reduces the power of the statistical results and limits the applicability of common statistical tests. E2 was conducted with the same procedure, problem domain, and evaluation process used in E1, in order to allow combining the results of both experiments. The results reported in the remainder of this chapter are from the combined data. Section 6.1 describes the participants of E2. Section 6.2 describes the data gathered from the participating teams. Section 6.3 describes the efforts taken to prepare the data for statistical analysis. Section 6.4 describes the hypotheses updated for E2. Section 6.5 describes the initial observations of the gathered data. Section 6.6 describes the results of the statistical analyses conducted on the gathered data. Section 6.8 summarizes the initial observations and the results of the statistical analyses and describes their effect on the hypotheses.

	Ideas				
	Relevant (C1 & C2)*	Feasible (C1 & C2)*	Innovative (C1 & C2)*	Innovative (C1 & C3)*	Innovative (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 6.1: Correlation between the Classifiers' Classifications of the Combined Ideas

6.1 Participants

Participants in E1 were all computer science and software engineering students. Because not many of these students spoke any right-to-left language, most teams were *3I*, consisting of only DIs. For E2, I decided to allow participants other than computer science and software engineering students. I knew that this decision might introduce new variables to the study, but it was the only option left at the time. I did insist that each participant be in some high technology field of study.

6.2 Gathered Data

As with E1, for E2, the list of all ideas generated by all teams was created. The same two domain experts classified the ideas with the same classification procedure used in E1. 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 6.1 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.

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, I decided to compare the ratios of the numbers of

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

relevant, feasible, and innovative ideas to the number of raw ideas for E1 and E2. As shown in Table 6.2, the changes in the ratios between E1 and E2 are clearly significant. 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 6.7.

Information about the participating teams is shown in Table 6.3, and a summary of the classification of their generated ideas is shown in Table 6.4.

6.3 Data Preparation for Statistical analysis

Compared to E1, educational background has been added as a new independent variable. Fortunately, the information required to extract the value of the new independent variable was collected when E1 was conducted.

As with E1, the data from E1 and E2 combined are first converted into numerical values. This section describes how the required conversions are done for independent variables about a team. Note that there are overlaps with the variables described in Section 4.3, but for the sake of completeness, all variables are described in this section.

- Mix of Domain Familiarities (MIX): the team's MIX value is the number of DIs it has, which is the numeral in the name of its mix, *0I*, *1I*, *2I*, or *3I*.
- Creativity (CR): the average of the team members' creativity scores. As shown in Figure 6.1, the Williams creativity scores were distributed so that scores in the range of 66 through 76.40 were the central part of the distribution. Therefore, each score was converted into a nominal value:

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 6.3: 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 6.4: Combined Data of the Generated Ideas

- 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.

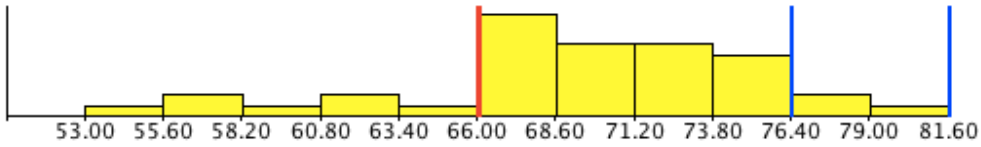


Figure 6.1: Distribution of the Teams' Average Creativity Scores of the Participating Teams

- RE Experience: the team members' 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. Based on the distribution shown in Figure 6.2, this number is transformed into four nominal values:
    * 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.

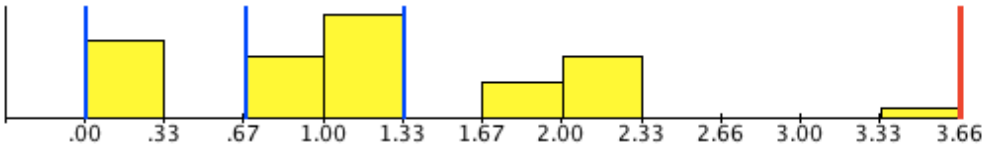


Figure 6.2: Distribution of the Teams' Average RE Experience

  - Industrial RE Experience (IREXP): the average number of industrial RE projects the members of the team have done in the past. Based on the distribution shown in Figure 6.3, this number is transformed into four numeral values:
    * 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

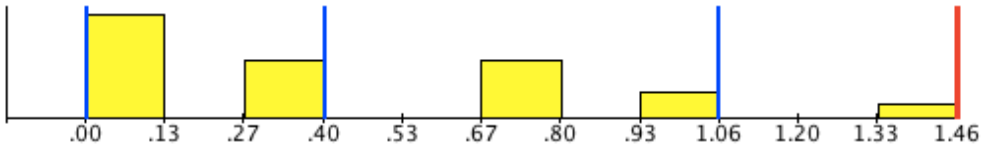


Figure 6.3: Distribution of the Teams' Average Industrial RE Experience

  - High: for a number greater than 1.06.

- Industrial Experience (IEXP): the average number of years of industrial software development experience of the members of the team. Based on the distribution shown in Figure 6.4, this number is transformed into four numeral values:
  - 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.

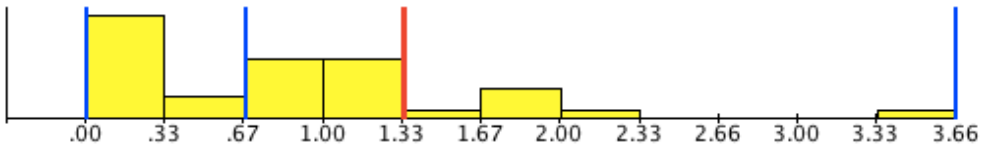


Figure 6.4: Distribution of the Teams' Average Industrial Experience

- Educational Background: the team members' educational background, which is divided into three subvariables:
  - Number of computer science student members (NCS): the number, between 0 and 3, of members in the team who are computer science students.
  - Number of software engineering student members (NSE): the number, between 0 and 3, of members in the team who are software engineering students.
  - Number of graduate student members (NGRAD): the number, between 0 and 3, of members in the team who are graduate students.

Table 6.5 summarizes the variables of the experiment.

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

Table 6.5: Variables of the Second Experiment

6.3.1 Data Normalization

Table 6.6 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 6.6: Test of Normality of the Dependent Variables

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 6.7: Acceptable Levels of Skewness (Adopted from [29])

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 [29], 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 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 [65]. 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 6.7 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 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 [64]. It shows whether a distribution has a greater or less than normal proportion of extreme scores in each tail [61]. 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

	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 6.8: Skewness and Kurtosis Test Results of the Dependent Variables

the skewness is 0. Table 6.8 shows that all standard skewness and kurtosis scores are outside of the acceptable ranges.

Now that the dataset is surely non-normal, with extreme skewness and kurtosis, it needs to be transformed in order to use ANOVA.

6.3.1.1 Transforming Data into Normal Distribution

Blom's formula [10] is a rank-based method that can be used to normalize non-normally distributed data. Table 6.9 shows that all the dependent variables, except NI, were successfully transformed into normal distributions. For NI, the Kolmogorov-Smirnov test result is 0.008 and the Shapiro-Wilk test result is 0.007, each of which is less than 0.05. Therefore, NI is not normalized.

Skewness and kurtosis are calculated once again. Table 6.10 shows that the skewness and kurtosis standard scores for all four dependent variables are within the acceptable range, even for NI.

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 6.9: Test of Normality of the Dependent Variables after Normalization

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

Table 6.10: Skewness and Kurtosis Test Results for the Dependent Variables after Normalization

Figure 6.5 shows on the left side, the plots for the original data for the dependent variables and on the right side, the plots for the normalized versions of the original data. It is evident that normalization has worked very well in transforming the data into normal distributions.

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 6.6 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 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 of 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, I decided to apply ANOVA to all dependent variables, and then, as an insurance policy, to apply to the AVG_I data a non-parametric test, which does not require the data to be normally distributed.

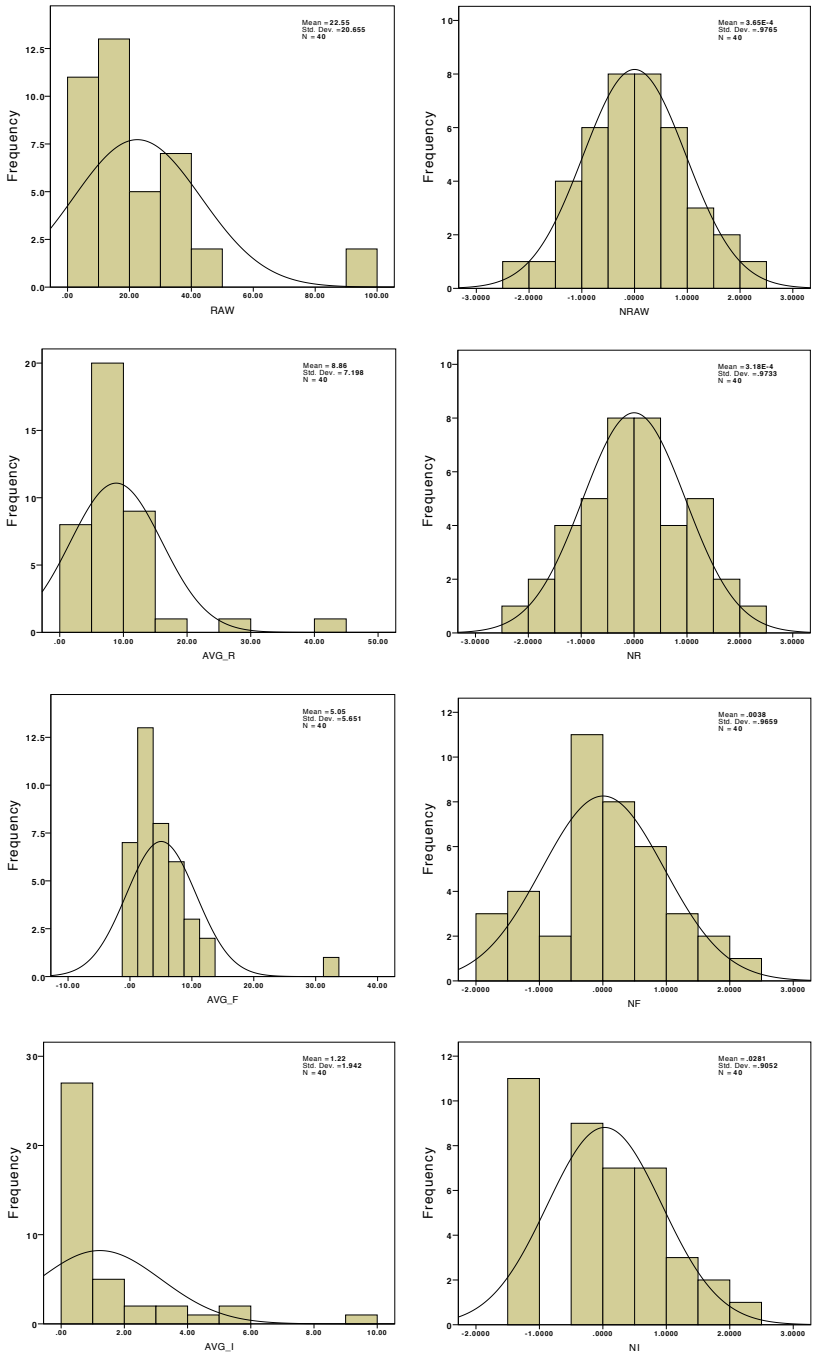
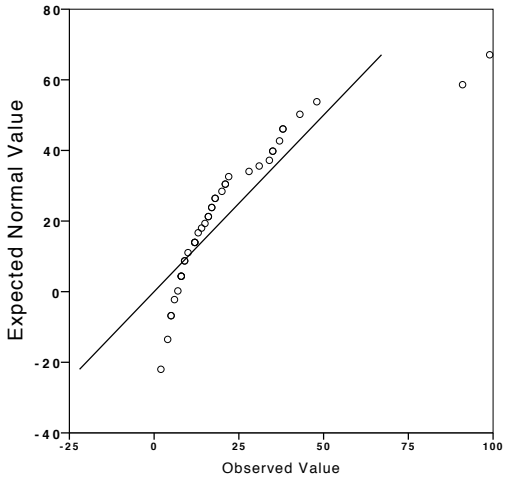
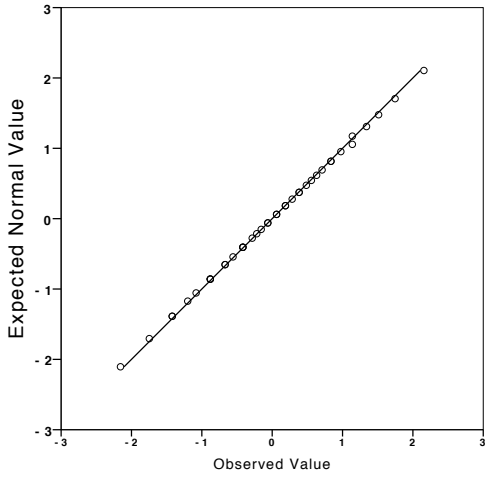


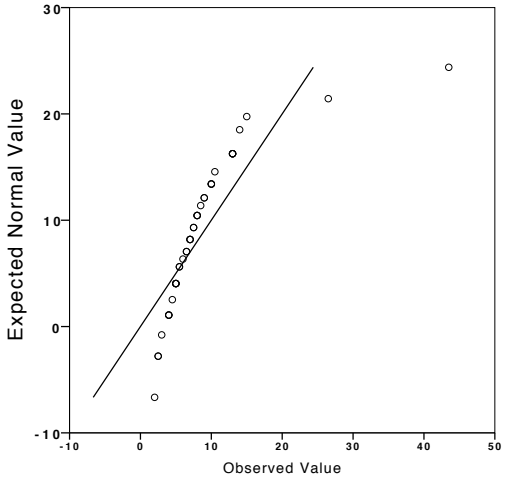
Figure 6.5: Normality Plots of the Dependent Variables



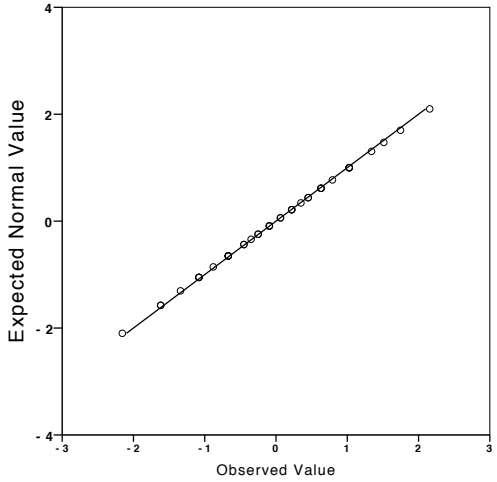
(a) RAW



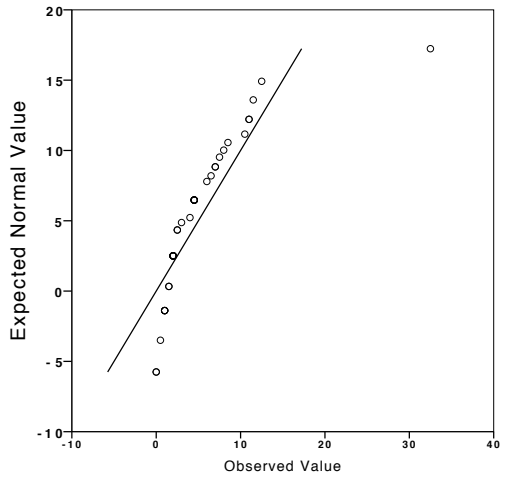
(b) NRAW



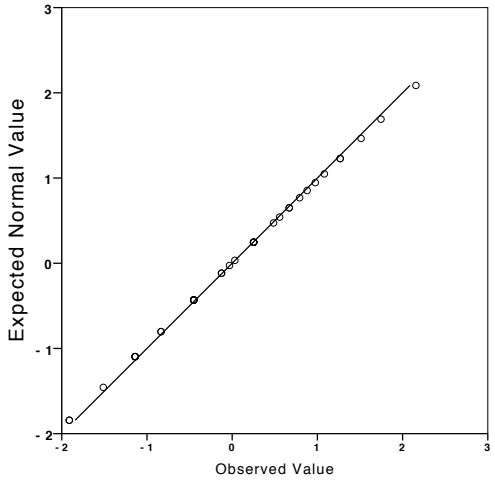
(c) AVG_R



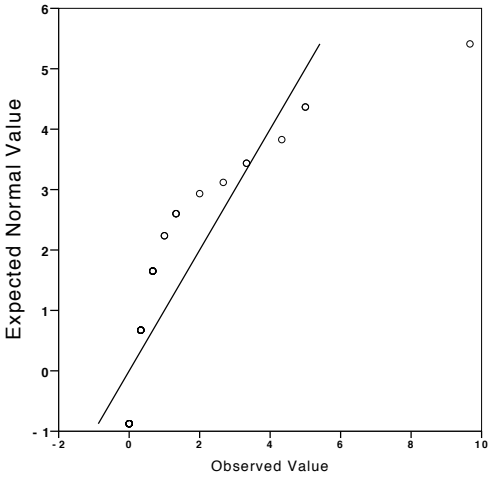
(d) NR



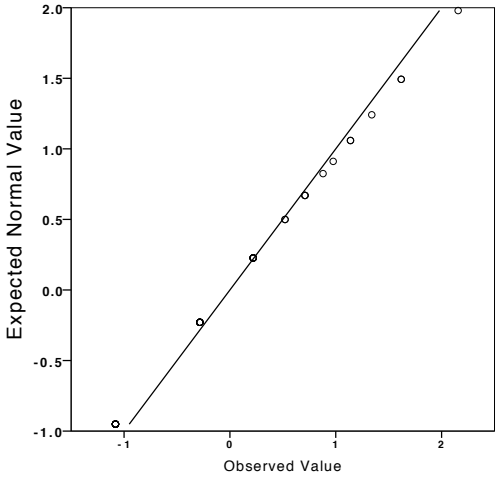
(e) AVG_F



(f) NF



(g) AVG_I



(h) NI

Figure 6.6: Q-Q Plots of the Dependent Variables

6.3.2 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 [59] 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 [30]. 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 6.3.2 shows the boxplot of the four dependent variables grouped by the main independent variable of the study, MIX. Figure 6.7(a) shows that the value of RAW is an outlier for Teams 8, 16, 32, and 34. Figure 6.7(b) shows that the value of AVG_R is an outlier for Teams 32 and 34. Figure 6.7(c) shows that the value of AVG_F is an outlier for Team 34. Figure 6.7(d) shows that the value of AVG_I is an outlier for Teams 19, 20, 24, and 34.

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 were marked as "Filtered". Otherwise, the results were marked as "Unfiltered", i.e., the study was done on the data including outliers.

6.3.2.1 Deeper Study of the Outliers

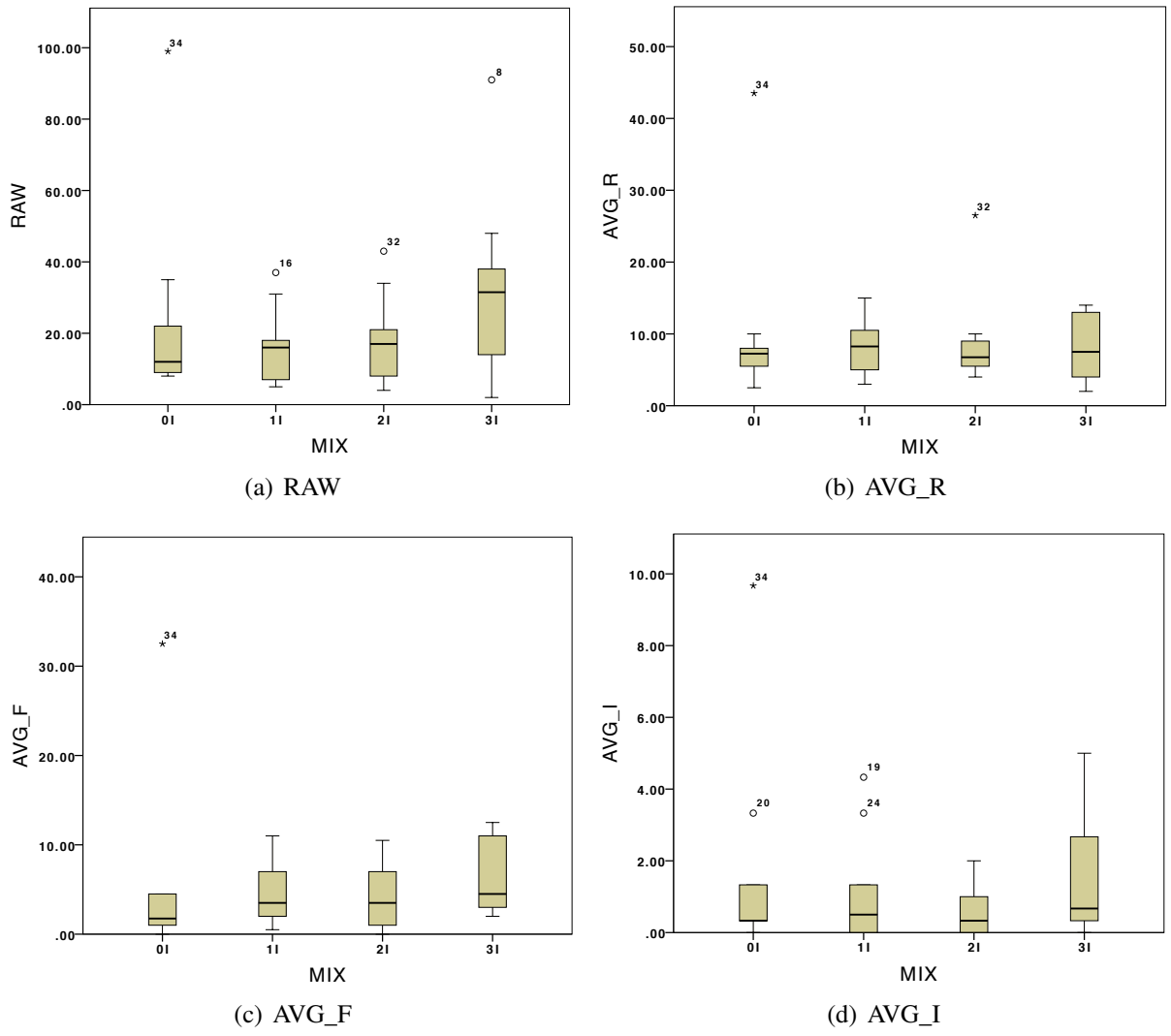
Outliers[1] produced about two times more RAW, AVG_R, AVG_F, and AVG_I compared to non-outliers.

When forming teams for E1 and E2, the only independent variable, beside MIX, that was considered in forming teams, was the teams' CR. Thus, the teams were balanced by their CR. 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, 2.29 for the outliers and 1.70 for the non-outliers,

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



(a) RAW (b) AVG_R

(c) AVG_F (d) AVG_I

Figure 6.7: Boxplots of the Dependent Variables

2. a lower IREXP, .71 for the outliers and 1.09 for the non-outliers,

3. a higher NGRAD, 2.14 for the outliers and 1.73 for the non-outliers, and

4. a lower NCS, 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 6.6 show that only two of these variables, NGRAD and NCS, have significant effects

on the effectiveness of the participating teams. I.e., 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.

6.3.3 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*[2]. The produced set of factors is supposed to be the main driver behind the dependent variables [33]. 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 (Table 6.5). 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)[3] is the most common method used in social sciences [63] to determine 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 KMO measure of a set of variables is greater than 0.5, factor analysis can be performed [34]. Table 6.11 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 6.11 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 6.12. The two factors indicated in Table 6.12 as Factor 1 and Factor 2 are the two factors identified by factor analysis. The numbers

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

[3]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 6.11: 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 6.12: Rotated Factor Matrix

in Table 6.12 are the loadings of each variable on each of the two identified factors. 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 6.8 plots the loading values of Table 6.12. 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 - 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.

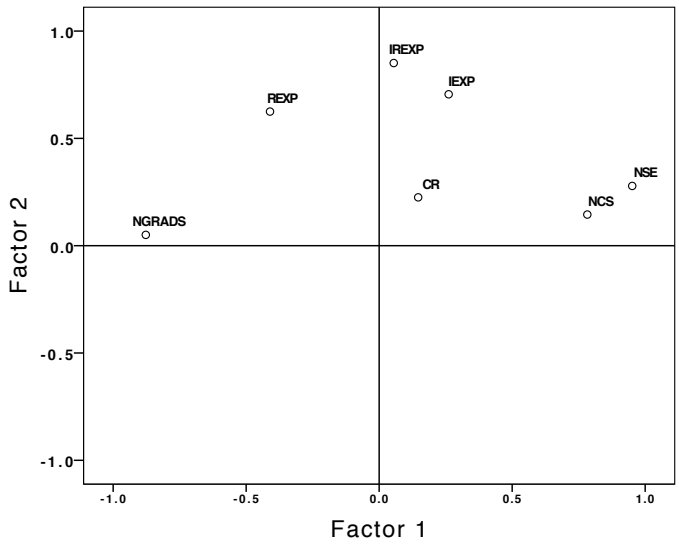


Figure 6.8: Factor Loading Plot

2. *Education (EDU)*: the sum of NSE and NCS. The resulting value is in the range of 0 - 6. This value is binned into:

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

Now that 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 is given in Section 6.6.9. The two factors are also studied in detail by means of one-way ANOVA in Section 6.6.

6.4 Hypotheses

Table 6.13 restates the hypotheses from Section 4.4 that this experiment is trying to answer. Note that the factor EDU has replaced NSE and NCS, and the factor EXP has replaced REXP, IREXP, and IEXP in these hypotheses. Each of the two factors and the remaining variables is addressed by its own pair of hypotheses.

Identifier	*Hypothesis*
H_{MIX_1}	The effectiveness of a team in requirement idea generation is affected by the team's mix of domain familiarities.
H_{MIX_0}	The effectiveness of a team in requirement idea generation is not affected by the team's mix of domain familiarities.
H_{CR_1}	The effectiveness of a team in requirement idea generation is affected by the team's creativity level.
H_{CR_0}	The effectiveness of a team in requirement idea generation is not affected by the team's creativity level.
H_{EDU_1}	The effectiveness of a team in requirement idea generation is affected by the team's educational background.
H_{EDU_0}	The effectiveness of a team in requirement idea generation is not affected by the team's educational background.
H_{NGRAD_1}	The effectiveness of a team in requirement idea generation is affected by the team's number of graduate participants.
H_{NGRAD_0}	The effectiveness of a team in requirement idea generation is not affected by the team's number of graduate participants.
H_{EXP_1}	The effectiveness of a team in requirement idea generation is affected by the team's experience.
H_{EXP_0}	The effectiveness of a team in requirement idea generation is not affected by the team's experience.

Table 6.13: List of the Updated Hypotheses

6.5 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 graphs 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 graphs 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 graphs.

As discussed earlier, due to the skewness of the distribution of the data, these graphs plot the medians, instead of the means, of the data.

6.5.1 Impact of MIX

Figure 6.9(a) shows that the medians of the unfiltered RAW generated by teams are positively correlated with the teams' MIX. Figure 6.10(a) shows that the plot of the medians of the filtered RAW generated by teams is similar to that of Figure 6.9(a). Thus, for the medians of the RAW values, removal of the outliers makes no real difference.

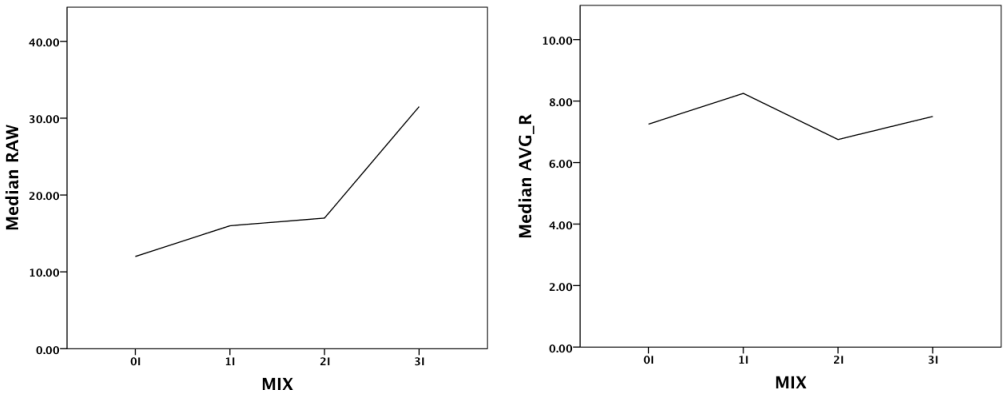
Figure 6.9(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' MIX. However, the figure shows that the medians unfiltered AVG_R generated by teams are highest for the teams with only one DI. Figure 6.10(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 6.9(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 6.9(c) shows that the medians of the unfiltered AVG_F generated by teams are positively correlated with the teams' MIX. Figure 6.10(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 6.9(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 6.9(d) shows that the medians of the unfiltered AVG_I generated by teams are not correlated with the teams' MIX. Figure 6.10(b) 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.

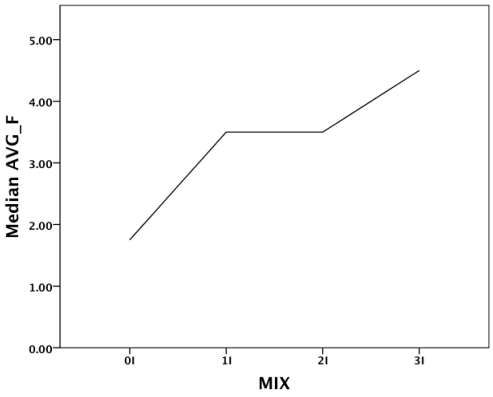
For the medians of the unfiltered and filtered RAW, AVG_F, and AVG_I, the corresponding plots show that the medians of the ideas generated by teams are highest for the teams with MIX = "*3I*".

Therefore, initially, it appears that H_{MIX_1} is supported and that hypothesis H_{MIX_0} is not supported.

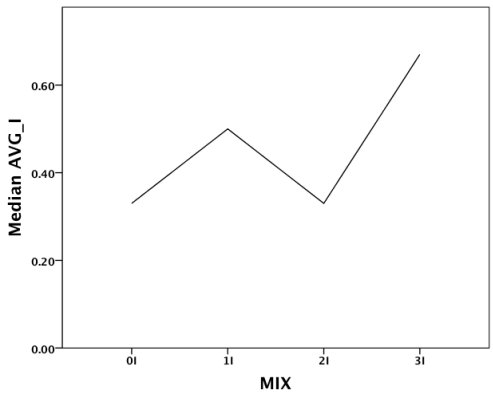


(a) RAW

(b) AVG_R

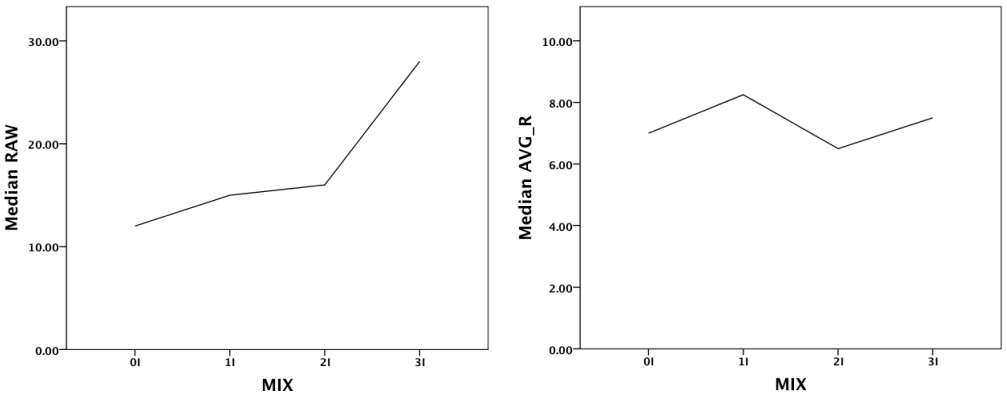


(c) AVG_F



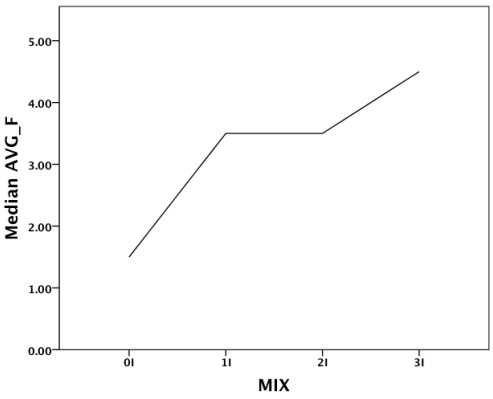
(d) AVG_I

Figure 6.9: Ideas vs. MIX (Unfiltered)

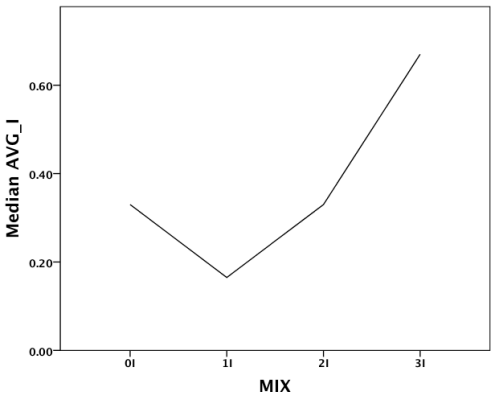


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Figure 6.10: Ideas vs. MIX (Filtered)

6.5.2 Impact of CR

Figure 6.11(a) shows that the medians of the unfiltered RAW generated by teams are partially negatively correlated with the teams' CR. Figure 6.12(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 AVG_R values, removal of the outliers makes a difference.

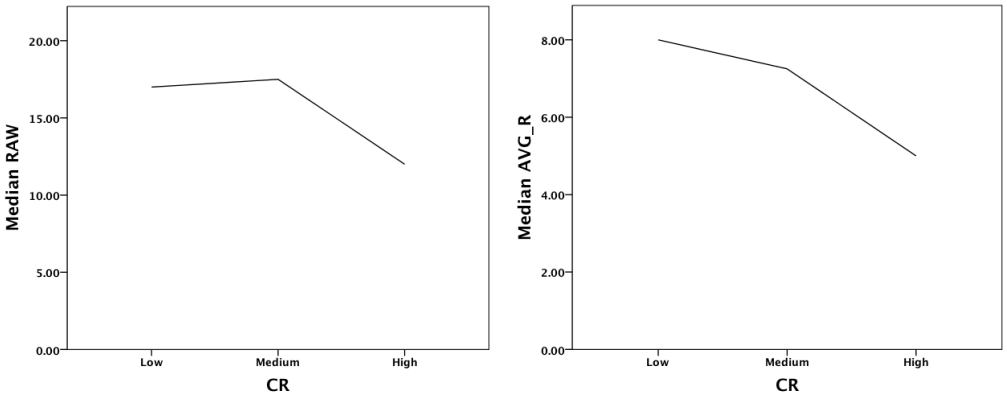
Figure 6.11(b) shows that the medians of the unfiltered AVG_R generated by teams are negatively correlated with the teams' CR. Figure 6.12(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 6.11(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 6.11(c) shows that the medians of the unfiltered AVG_F generated by teams are not correlated with the teams' CR. Figure 6.12(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 6.11(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 6.11(d) shows that the medians of the unfiltered AVG_I generated by teams are not correlated with the teams' CR. Figure 6.12(d) shows that the plot of the medians of the filtered AVG_I generated by teams is similar to that of Figure 6.11(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

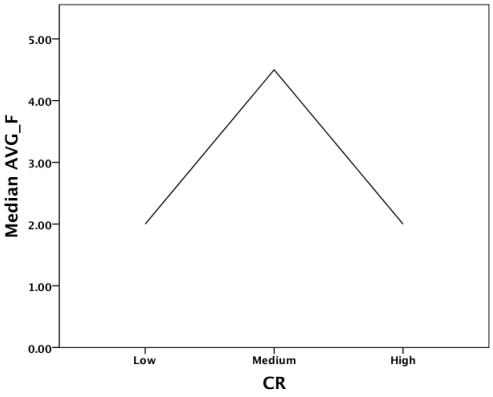
For the medians of the unfiltered and filtered RAW, AVG_F, and AVG_I, the plots show that the medians of the ideas generated by teams are highest for the teams with CR = "Medium".

Therefore, initially, it appears that H_{CR_0} is supported and that hypothesis H_{CR_1} is not supported.

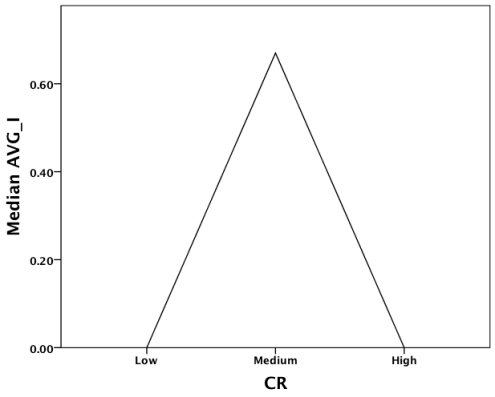


(a) RAW

(b) AVG_R

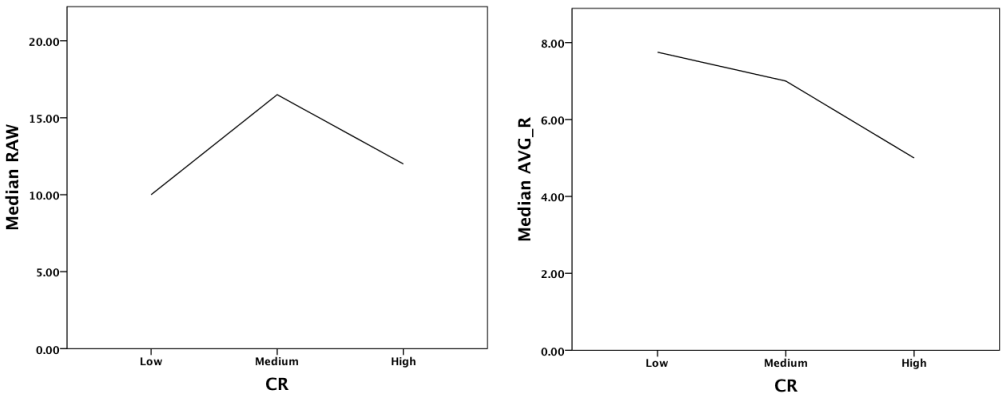


(c) AVG_F



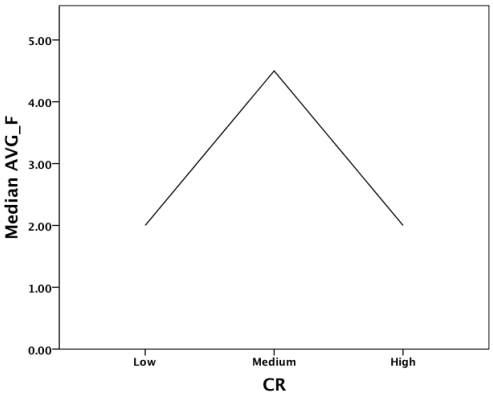
(d) AVG_I

Figure 6.11: Ideas vs. CR (Unfiltered)

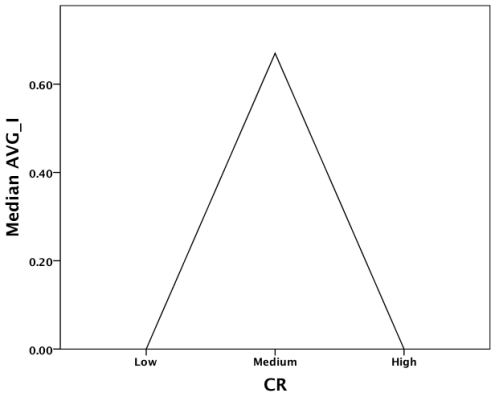


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Figure 6.12: Ideas vs. CR (Filtered)

6.5.3 Impact of REXP

Figure 6.13(a) shows that the medians of the unfiltered RAW generated by teams are not correlated with the teams' REXP. Figure 6.14(a) shows that the plot of the medians of the filtered RAW generated by teams is quite similar to that of Figure 6.13(a). Thus, for the medians of the RAW values, removal of the outliers makes no real difference.

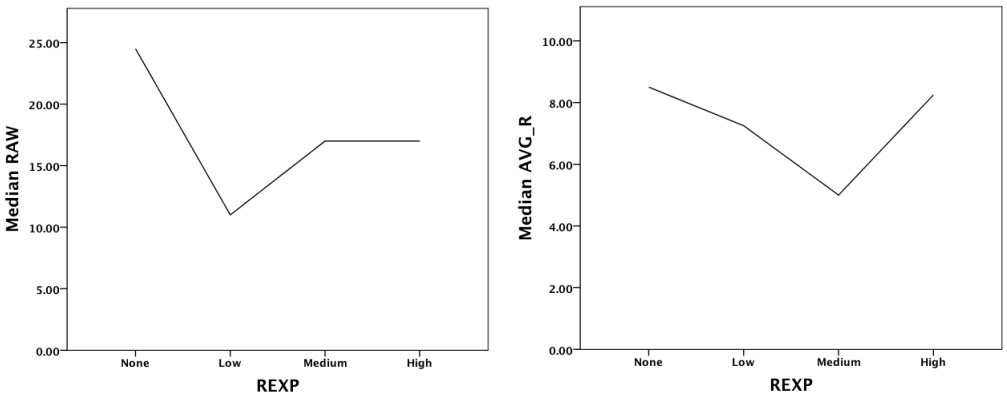
Figure 6.13(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' REXP. Figure 6.14(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 6.13(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 6.13(c) shows that the medians of the unfiltered AVG_F generated by teams are not correlated with the teams' REXP. Figure 6.14(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 6.13(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 6.13(d) shows that the medians of the unfiltered AVG_I generated by teams are not correlated with the teams' REXP. Figure 6.14(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.

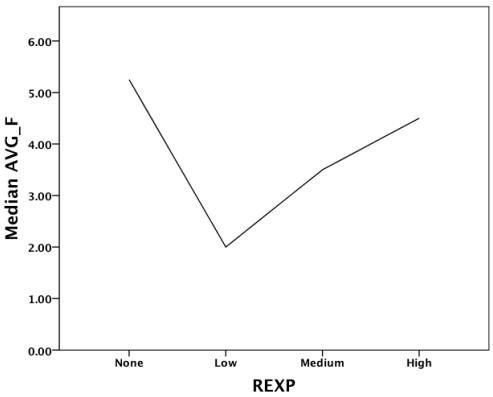
For the medians of the unfiltered and filtered RAW, AVG_R, and AVG_F, the plots show that the medians of the ideas generated by teams are highest for the teams with REXP = "None".

Therefore, initially, it appears that H_{REXP_0} is supported and that hypothesis H_{REXP_1} is not supported.

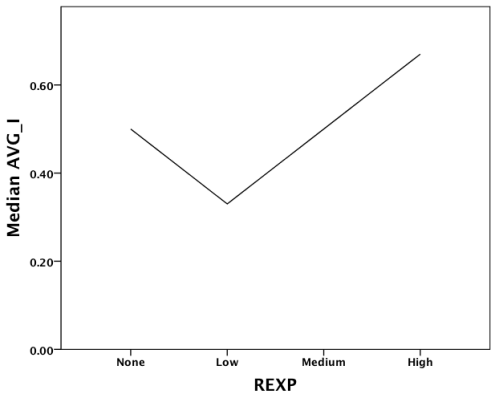


(a) RAW

(b) AVG_R

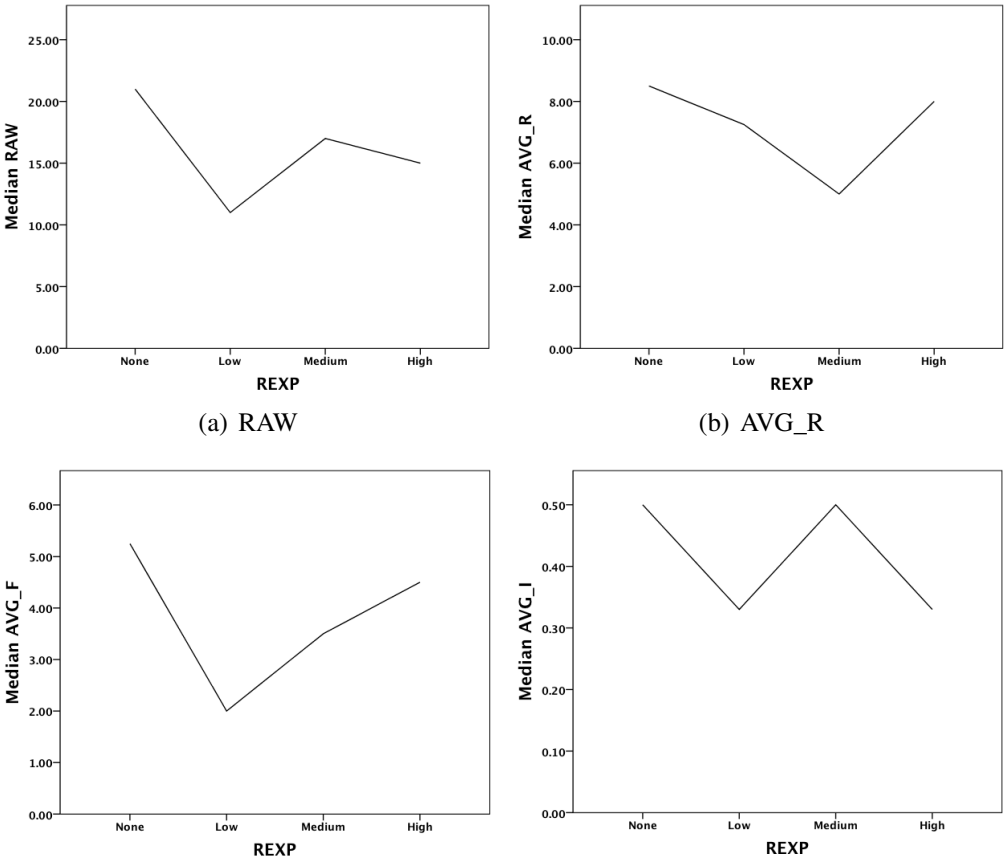


(c) AVG_F



(d) AVG_I

Figure 6.13: Ideas vs. REXP (Unfiltered)



(a) RAW (b) AVG_R

(c) AVG_F (d) AVG_I

Figure 6.14: Ideas vs. REXP (Filtered)

6.5.4 Impact of IREXP

Figure 6.15(a) shows that the medians of the unfiltered RAW generated by teams are positively correlated with the teams' IREXP. Figure 6.16(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 difference.

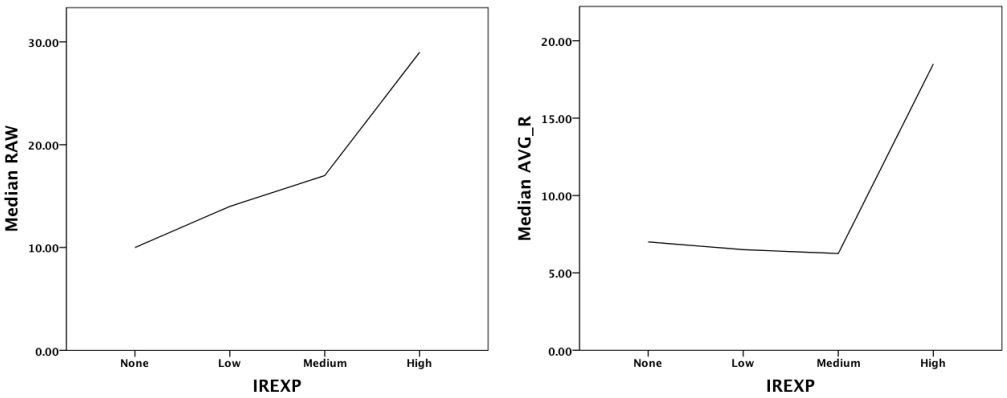
Figure 6.15(b) shows that the medians of the unfiltered AVG_R generated by teams are partially positively correlated with the teams' IREXP. Figure 6.16(b) shows that the medians of the filtered AVG_R generated by teams are partially positively correlated with the teams' IREXP. Thus, for the medians of the AVG_R values, removal of the outliers makes a difference.

Figure 6.15(c) shows that the medians of the unfiltered AVG_F generated by teams are partially positively correlated with the teams' IREXP. Figure 6.16(c) shows that the medians of the filtered AVG_F generated by teams are partially positively correlated with the teams' IREXP. Thus, for the medians of the AVG_F values, removal of the outliers makes a difference.

Figure 6.15(d) shows that the medians of the unfiltered AVG_I generated by teams are partially positively correlated with the teams' IREXP. Figure 6.16(d) shows that the plot of the medians of the filtered AVG_I generated by teams is similar to that of Figure 6.15(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

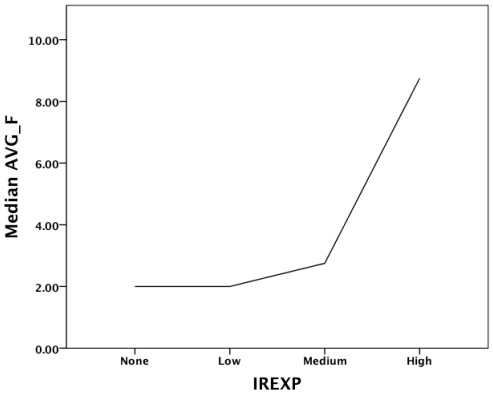
For the medians of the unfiltered and filtered AVG_R, AVG_F, and AVG_I, and the unfiltered RAW, the plots show that the medians of the ideas generated by teams are highest for the teams with IREXP = "High".

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

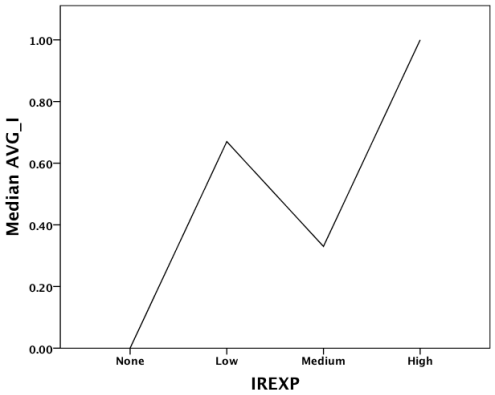


(a) RAW

(b) AVG_R

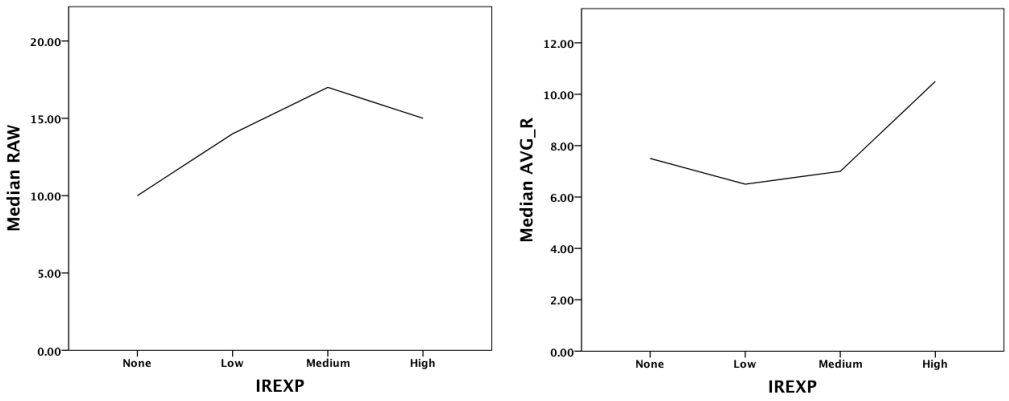


(c) AVG_F



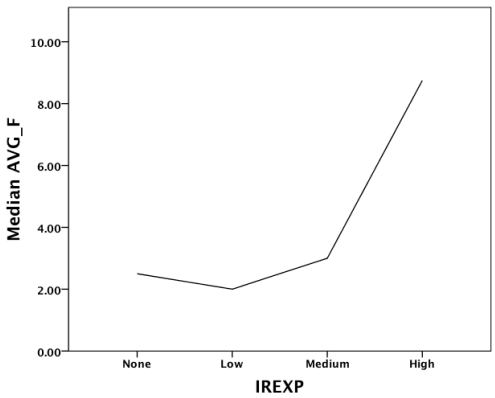
(d) AVG_I

Figure 6.15: Ideas vs. IREXP (Unfiltered)

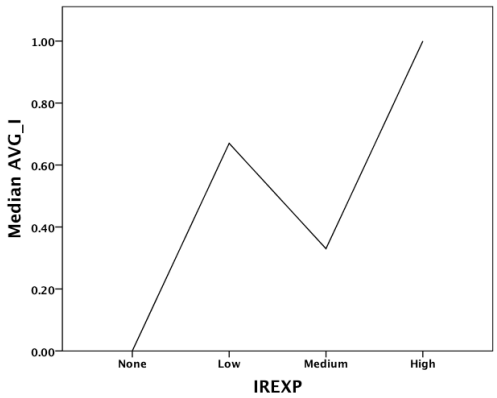


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Figure 6.16: Ideas vs. IREXP (Filtered)

6.5.5 Impact of IEXP

Figure 6.17(a) shows that the medians of the unfiltered RAW generated by teams are partially positively correlated with the teams' IEXP. Figure 6.18(a) shows that the medians of the filtered RAW generated by teams are partially positively correlated with the teams' IEXP. Thus, for the medians of the RAW values, removal of the outliers makes a difference.

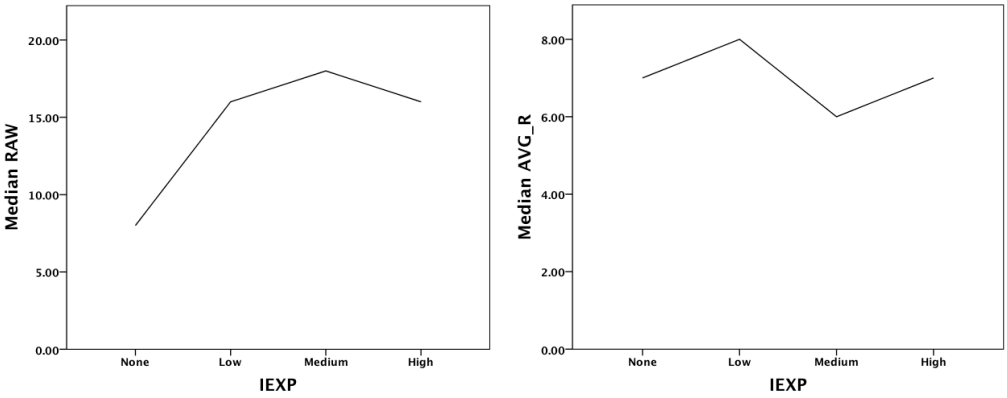
Figure 6.17(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' IEXP. Figure 6.18(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 6.17(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 6.17(c) shows that the medians of the unfiltered AVG_F generated by teams are partially positively correlated with the teams' IEXP. Figure 6.18(c) shows that the plot of the medians of the filtered AVG_F generated by teams is quite similar to that of Figure 6.17(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 6.17(d) shows that the medians of the unfiltered AVG_I generated by teams are not correlated with the teams' IEXP. Figure 6.18(d) shows that the medians of the filtered AVG_I generated by teams are not correlated with the teams' IEXP. Thus, for the medians of the AVG_I values, removal of the outliers makes a slight difference.

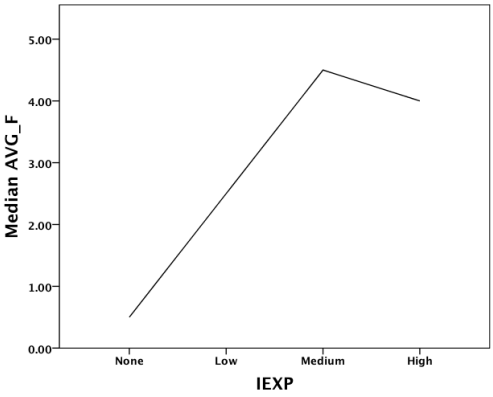
For the medians of the unfiltered and filtered RAW, AVG_F, and AVG_I, the plots show that the medians of the ideas generated by teams are highest for the teams with IEXP = "Medium".

Therefore, initially, it appears that H_{IEXP_1} is supported and that hypothesis H_{MIX_0} is not supported.

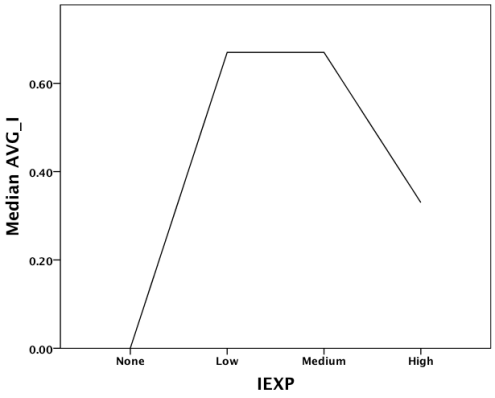


(a) RAW

(b) AVG_R

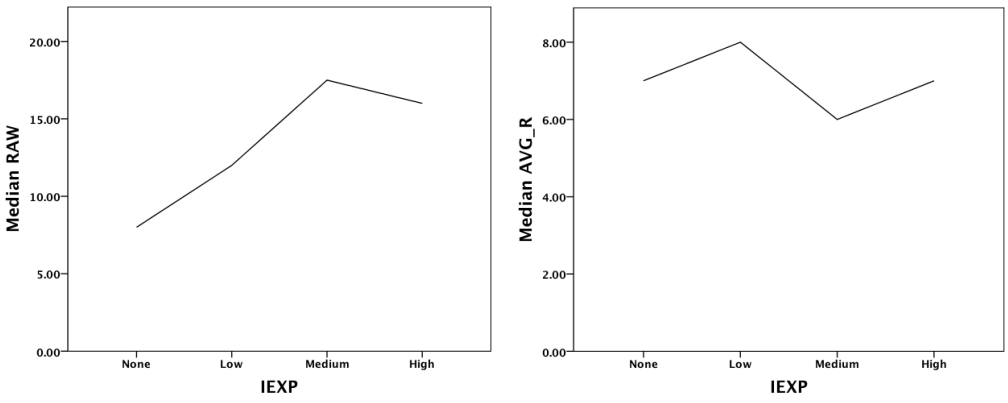


(c) AVG_F



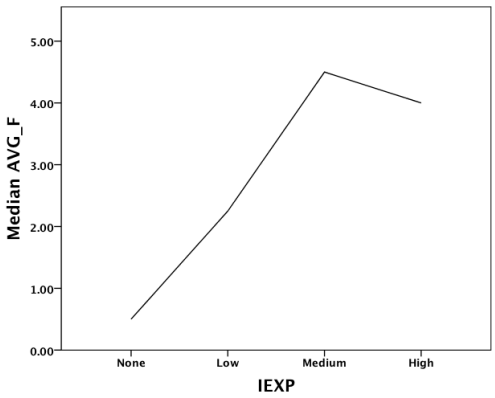
(d) AVG_I

Figure 6.17: Ideas vs. IEXP (Unfiltered)

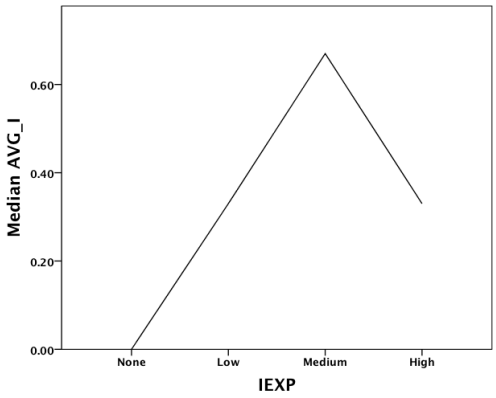


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Figure 6.18: Ideas vs. IEXP (Filtered)

6.5.6 Impact of NCS

Figure 6.19(a) shows that the medians of the unfiltered RAW generated by teams are positively correlated with the teams' NCS. Figure 6.20(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 difference.

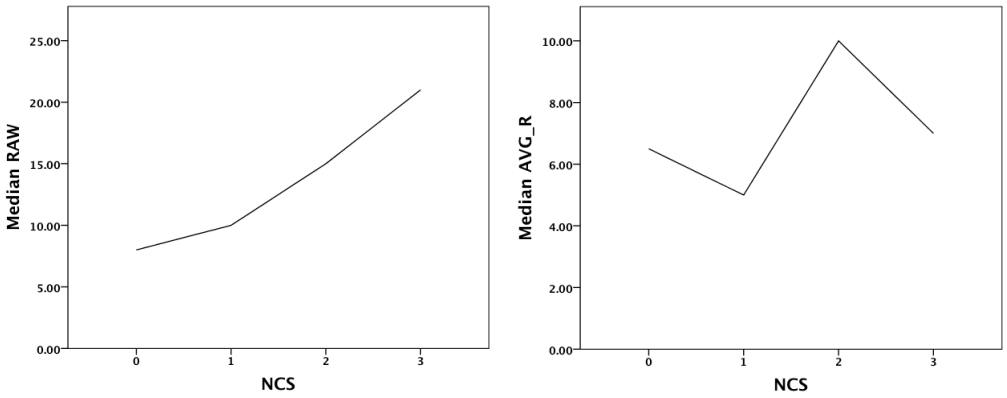
Figure 6.41(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' NCS. Figure 6.42(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 6.41(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 6.41(d) shows that the medians of the unfiltered AVG_F generated by teams are partially positively correlated with the teams' NCS. Figure 6.42(d) shows that the medians of the filtered AVG_F generated by teams are partially positively correlated with the teams' NCS. Thus, for the medians of the AVG_F values, removal of the outliers makes a difference.

Figure 6.41(f) shows that the medians of the unfiltered AVG_I generated by teams are partially positively correlated with the teams' NCS. Figure 6.42(f) shows that the plot of the medians of the filtered AVG_I generated by teams is similar to that of Figure 6.41(f). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

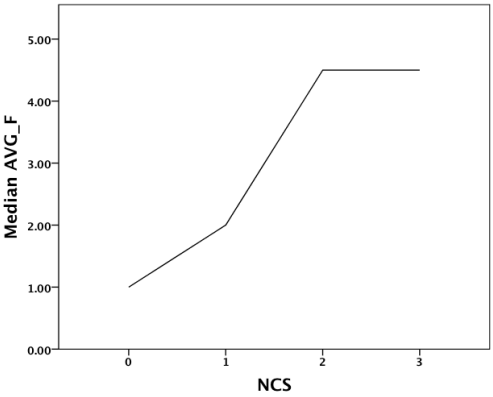
For the medians of the unfiltered and filtered RAW, AVG_R, AVG_F, and AVG_I, the plots show that the medians of the ideas generated by teams are highest for the teams with either NCS = 2 or NCS = 3.

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

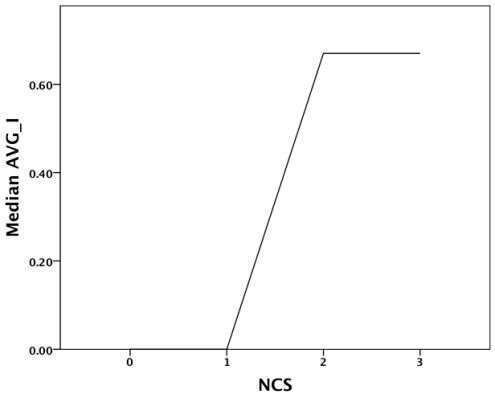


(a) RAW

(b) AVG_R

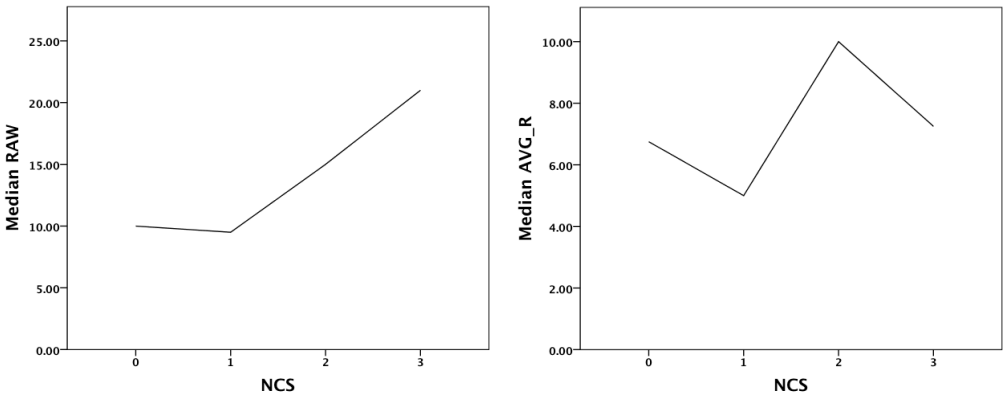


(c) AVG_F



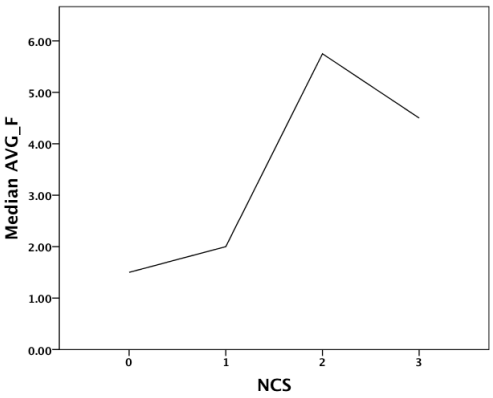
(d) AVG_I

Figure 6.19: Ideas vs. NCS (Unfiltered)

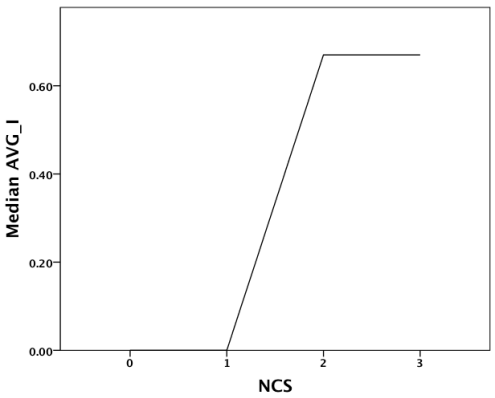


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Figure 6.20: Ideas vs. NCS (Filtered)

6.5.7 Impact of NSE

Figure 6.21(a) shows that the medians of the unfiltered RAW generated by teams are positively correlated with the teams' NSE. Figure 6.22(a) shows that the plot of the medians of the filtered RAW generated by teams is quite similar to that of Figure 6.21(a). Thus, for the medians of the RAW values, removal of the outliers makes no real difference.

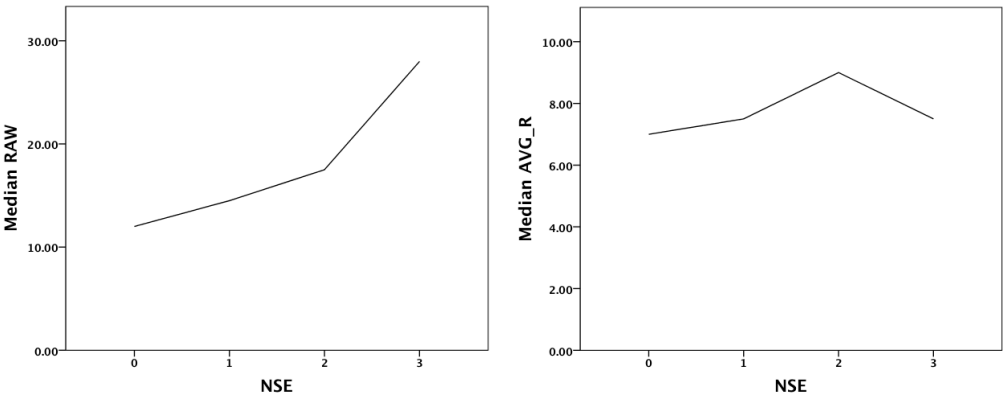
Figure 6.43(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' NSE. Figure 6.44(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 6.43(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 6.43(d) shows that the medians of the unfiltered AVG_F generated by teams are partially positively correlated with the teams' NSE. Figure 6.44(d) shows that the plot of the medians of the filtered AVG_F generated by teams is quite similar to that of Figure 6.41(d). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 6.43(f) shows that the medians of the unfiltered AVG_I generated by teams are partially positively correlated with the teams' NSE. Figure 6.44(f) shows that the plot of the medians of the filtered AVG_I generated by teams is quite similar to that of Figure 6.43(f). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

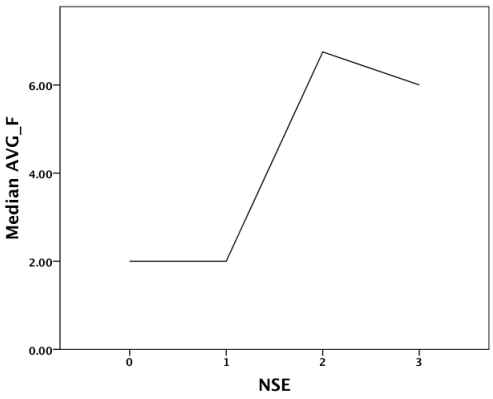
For the medians of the unfiltered and filtered AVG_R, AVG_F, and AVG_I, the corresponding plots show that the medians of the ideas generated by teams are highest for the teams with NSE = 2.

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

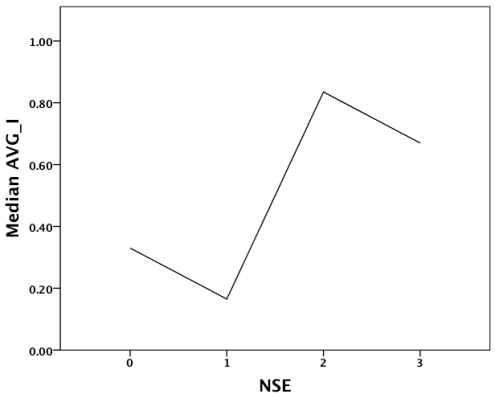


(a) RAW

(b) AVG_R

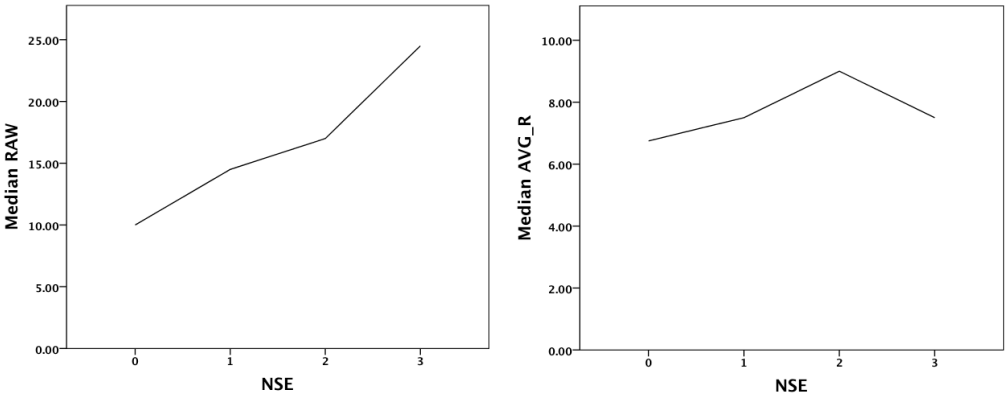


(c) AVG_F



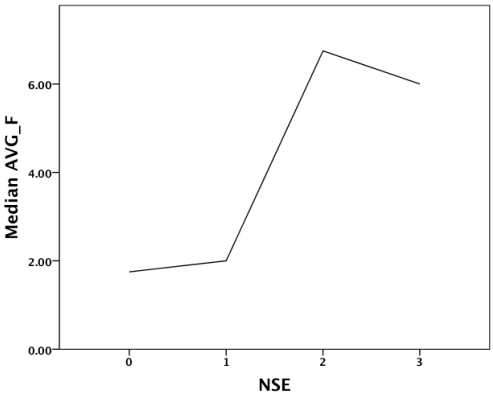
(d) AVG_I

Figure 6.21: Ideas vs. NSE (Unfiltered)

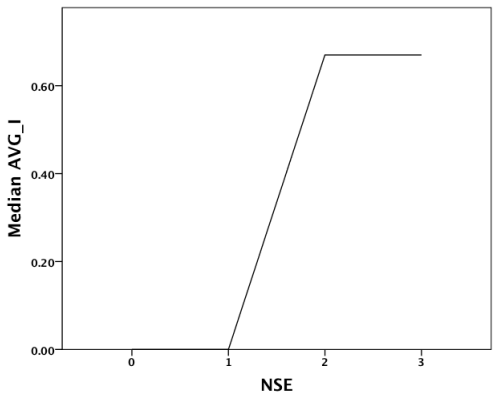


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Figure 6.22: Ideas vs. NSE (Filtered)

6.5.8 Impact of NGRAD

Figure 6.23(a) shows that the medians of the unfiltered RAW generated by teams are negatively correlated with the teams' NGRAD. Figure 6.24(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 difference.

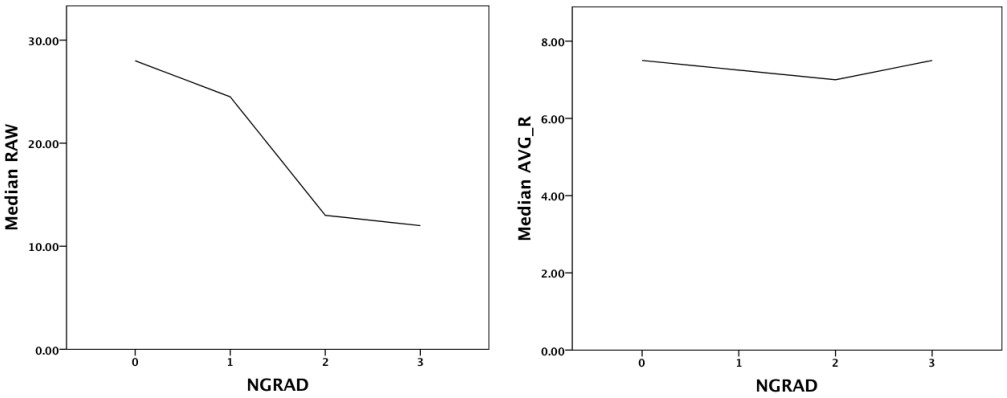
Figure 6.45(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' NGRAD. Figure 6.46(b) shows that the plot of the medians of the filtered AVG_R generated by teams is quite similar to that of Figure 6.45(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 6.45(d) shows that the medians of the unfiltered AVG_F generated by teams are negatively correlated with the teams' NGRAD. Figure 6.46(d) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 6.45(d). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 6.45(f) shows that the medians of the unfiltered AVG_I generated by teams are not correlated with the teams' NGRAD. Figure 6.46(f) shows that the medians of the filtered AVG_I generated by teams are partially negatively correlated with the teams' NGRAD. Thus, for the medians of the AVG_I values, removal of the outliers makes a difference.

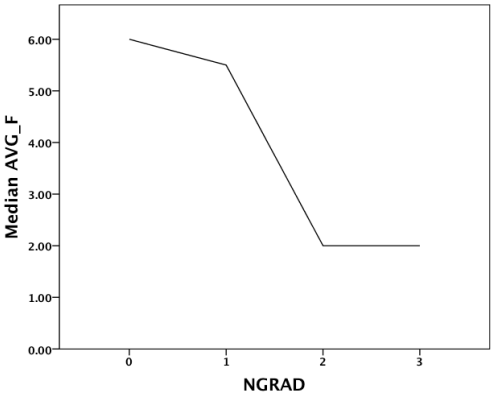
For the medians of the unfiltered and filtered RAW, AVG_R, and AVG_F, the plots show that the medians of the ideas generated by teams are highest for the teams with NGRAD = 0.

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

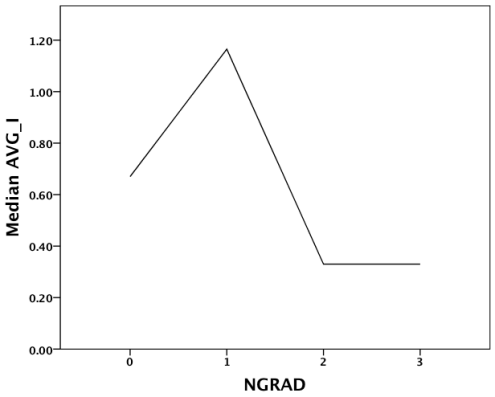


(a) RAW

(b) AVG_R

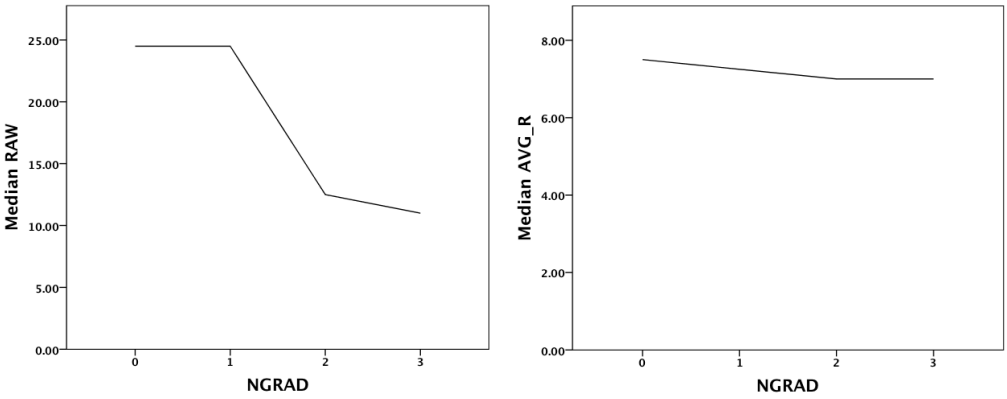


(c) AVG_F



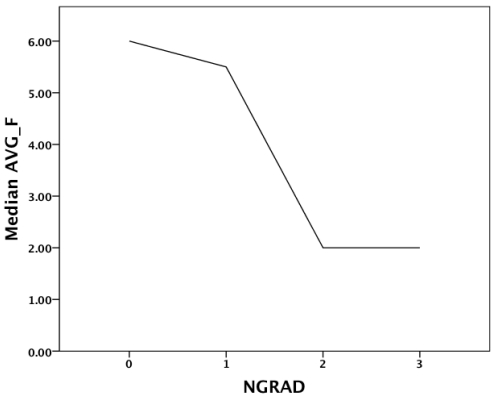
(d) AVG_I

Figure 6.23: Ideas vs. NGRAD (Unfiltered)

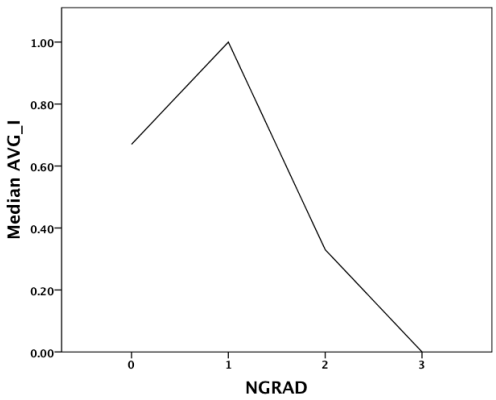


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Figure 6.24: Ideas vs. NGRAD (Filtered)

6.5.9 Impact of EDU

Figure 6.25(a) shows that the medians of the unfiltered RAW generated by teams are positively correlated with the teams' EDU. Figure 6.26(a) shows that the plot of the medians of the filtered RAW generated by teams is similar to that of Figure 6.25(a). Thus, for the medians of the RAW values, removal of the outliers makes no real difference.

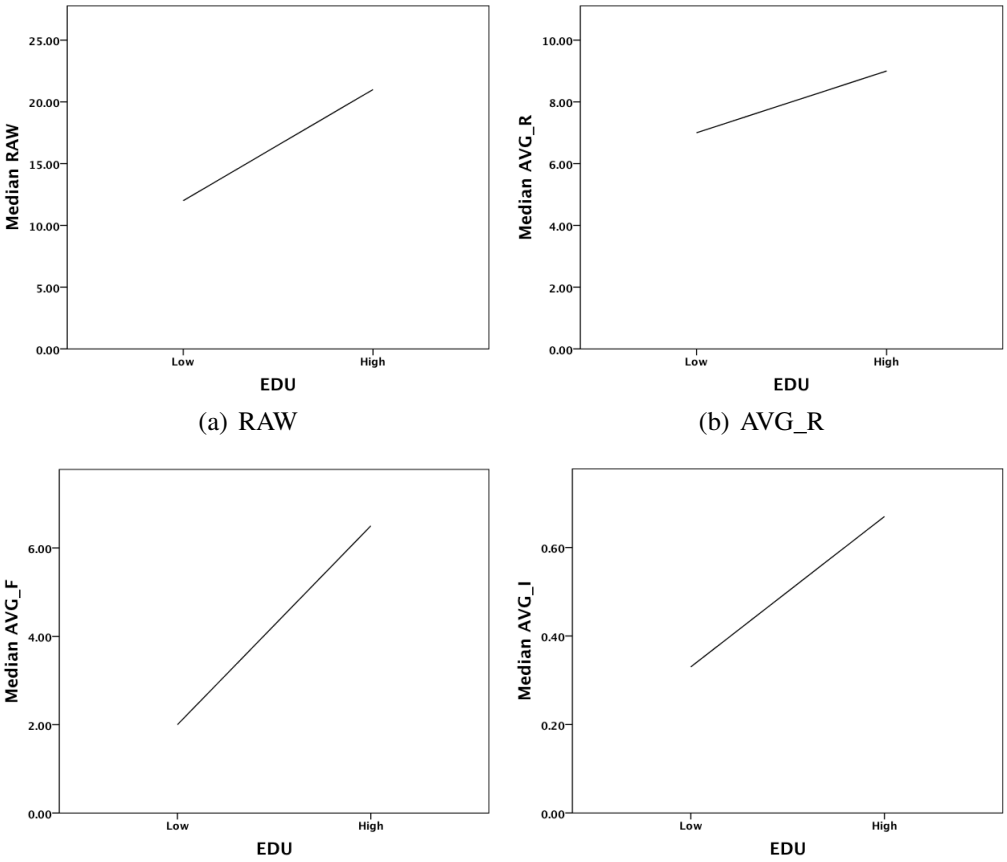
Figure 6.25(b) shows that the medians of the unfiltered AVG_R generated by teams are positively correlated with the teams' EDU. Figure 6.26(b) shows that the plot of the medians of the filtered AVG_R generated by teams is similar to that of Figure 6.25(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 6.25(c) shows that the medians of the unfiltered AVG_F generated by teams are positively correlated with the teams' EDU. Figure 6.26(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 6.25(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 6.25(d) shows that the medians of the unfiltered AVG_I generated by teams are positively correlated with the teams' EDU. Figure 6.26(d) shows that the plot of the medians of the filtered AVG_R generated by teams is quite similar to that of Figure 6.25(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

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

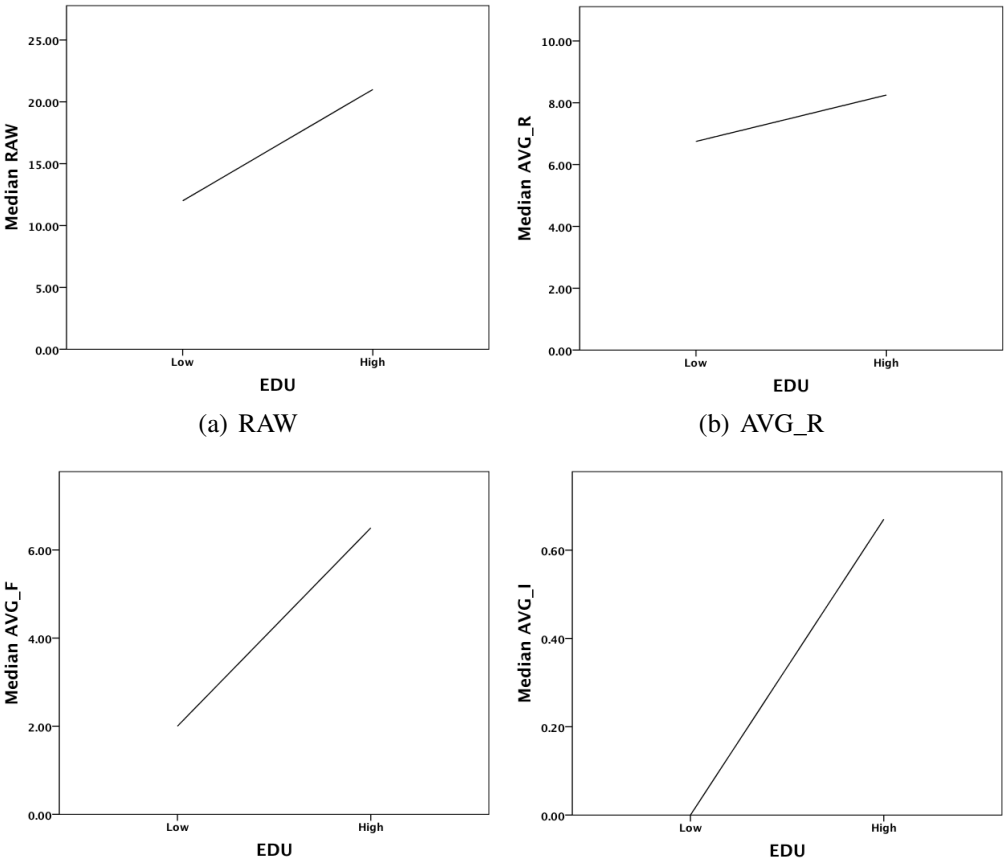
Therefore, initially, it appears that H_{EDU_1} is supported and that hypothesis H_{EDU_0} is not supported.



(a) RAW (b) AVG_R

(c) AVG_F (d) AVG_I

Figure 6.25: Ideas vs. EDU (Unfiltered)



(a) RAW (b) AVG_R

(c) AVG_F (d) AVG_I

Figure 6.26: Ideas vs. EDU (Filtered)

6.5.10 Impact of EXP

Figure 6.27(a) shows that the medians of the unfiltered RAW generated by teams are partially positively correlated with the teams' EXP. Figure 6.28(a) shows that the plot of the medians of the filtered RAW generated by teams is quite similar to that of Figure 6.27(a). Thus, for the medians of the RAW values, removal of the outliers makes no real difference.

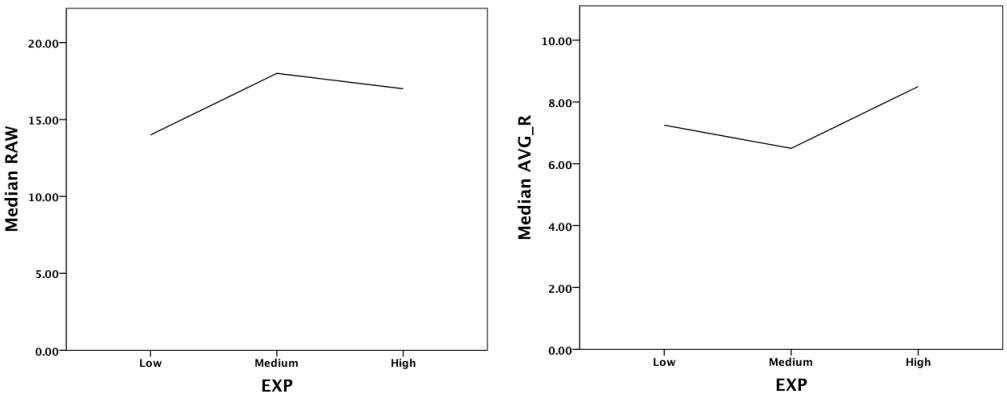
Figure 6.27(b) shows that the medians of the unfiltered AVG_R generated by teams are not correlated with the teams' EXP. Figure 6.28(b) shows that the plot of the medians of the filtered AVG_R generated by teams is quite similar to that of Figure 6.27(b). Thus, for the medians of the AVG_R values, removal of the outliers makes no real difference.

Figure 6.27(c) shows that the medians of the unfiltered AVG_F generated by teams are partially positively correlated with the teams' EXP. Figure 6.28(c) shows that the plot of the medians of the filtered AVG_F generated by teams is similar to that of Figure 6.27(c). Thus, for the medians of the AVG_F values, removal of the outliers makes no real difference.

Figure 6.27(d) shows that the medians of the unfiltered AVG_I generated by teams are not correlated with the teams' EXP. Figure 6.28(d) shows that the plot of the medians of the filtered AVG_I generated by teams is quite similar to that of Figure 6.27(d). Thus, for the medians of the AVG_I values, removal of the outliers makes no real difference.

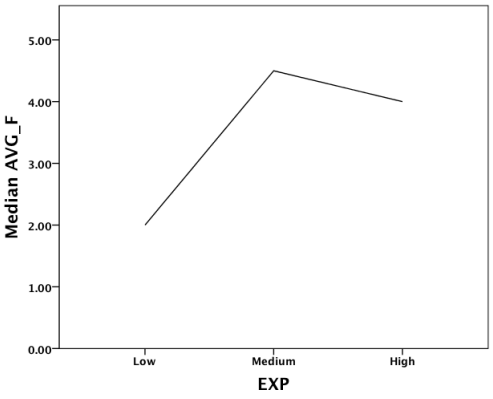
For the medians of the unfiltered and filtered RAW, AVG_F, and AVG_I, the plots show that the medians of the ideas generated by teams are highest for the teams with EXP = "Medium".

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

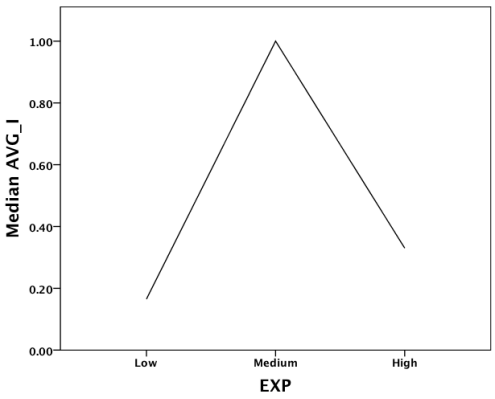


(a) RAW

(b) AVG_R

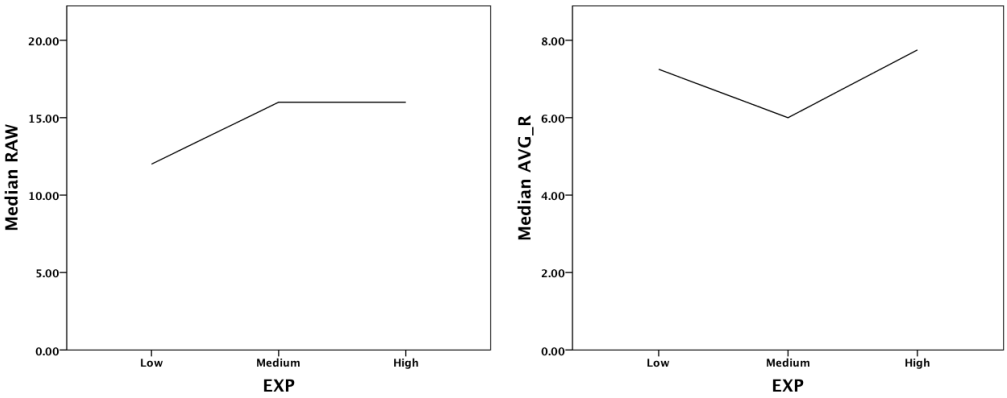


(c) AVG_F



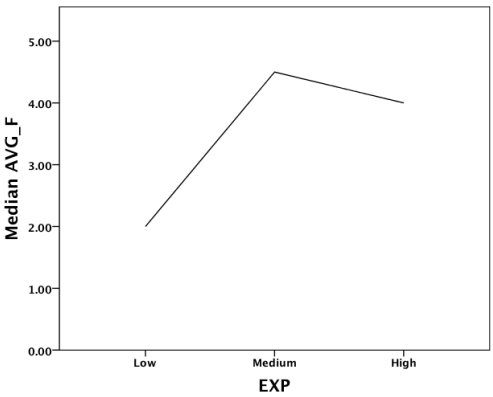
(d) AVG_I

Figure 6.27: Ideas vs. EXP (Unfiltered)

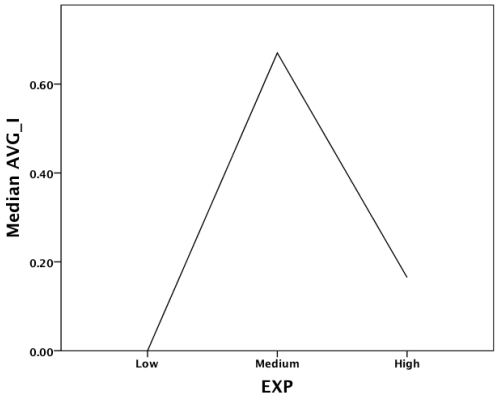


(a) RAW

(b) AVG_R



(c) AVG_F



(d) AVG_I

Figure 6.28: Ideas vs. EXP (Filtered)

6.6 Statistical Analysis[4]

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 6.3.3 to test the hypotheses given in Section 6.4. For the purposes of the following discussion, each factor is considered an independent variable.

Each of the subsections of this section attempts 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, but 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.

Then, the ANOVA itself is done to assess the impact of the chosen set of independent variables of a team on the subset of the team's dependent variables for which the ANOVA is reliable.

Then, for each of a team's dependent variables that the ANOVA test found to be significantly affected by the chosen set of independent variables, a Tukey HSD pairwise comparison test [66] 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 the ANOVA was 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 impact 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 [15] is conducted to compare all possible pairs of medians of the dependent variable to show which medians are significantly different from each other.

[4]The following analysis uses standard statistical terminology and notations.

Therefore, 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.

6.6.1 One-Way ANOVA on MIX

Table 6.14 shows that the Levene test result of the unfiltered dependent variables plotted against MIX is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these unfiltered variables.

Table 6.15 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. Therefore, 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 6.14: Results of the Levene Test for MIX (Unfiltered)

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 6.15: Results of the Levene Test for MIX (Filtered)

Table 6.16 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 6.17 shows the results of the ANOVA test of the filtered NRAW and NF plotted against MIX. The test shows no significant effect of the team's MIX on any of these variables.

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

[a] Type III sum of squares
[b] Degrees of freedom
[c] Value of the ANOVA's F-test
[d] p-value of the F-test
[e] Measure of effect size

Table 6.16: 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 6.17: Results of the One-Way ANOVA of the Effect of MIX (Filtered)

Table 6.18 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 6.19 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 6.18: Results of the Kruskal-Wallis Test of the Effect of MIX (Unfiltered)

Dependent Variable	*p*
AVG_R	.697
AVG_I	.264

Table 6.19: Results of the Kruskal-Wallis Test of the Effect of MIX (Filtered)

6.6.2 One-Way ANOVA on CR

Table 6.20 shows that the Levene test result of the unfiltered dependent variables plotted against CR is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these unfiltered variables.

Table 6.21 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. Therefore, 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 6.20: 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 6.21: Results of the Levene Test for CR (Filtered)

Table 6.22 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 6.23 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.

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 6.22: 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 6.23: Results of the One-Way ANOVA of the Effect of CR (Filtered)

Table 6.24 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 6.25 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	***p***
AVG_I	.102

Table 6.24: Results of the Kruskal-Wallis Test of the Effect of CR (Unfiltered)

Dependent Variable	***p***
AVG_R	.380
AVG_I	.060

Table 6.25: Results of the Kruskal-Wallis Test of the Effect of CR (Filtered)

6.6.3 One-Way ANOVA on REXP

Table 6.26 shows that the Levene test result of the unfiltered dependent variables plotted against REXP is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these unfiltered variables.

Table 6.27 shows that the Levene test result of the filtered dependent variables plotted against REXP is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these filtered variables.

Table 6.28 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 6.29 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.

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 6.26: 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 6.27: 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 6.28: 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 6.29: Results of the One-Way ANOVA of the Effect of REXP (Filtered)

Table 6.30 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 that the difference

between the means of the NR of the teams is significant when REXP = “Medium” and REXP = “High”.

Sample 1	*Sample 2*	*Mean Difference*	*Standard 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 6.30: Results of the Tukey HSD Pairwise Comparison Test of the Effect of REXP on NR (Unfiltered)

Tables 6.31 and 6.32 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 6.31: Results of the Kruskal-Wallis Test of the Effect of REXP (Unfiltered)

Dependent Variable	*p*
AVG_I	.948

Table 6.32: Results of the Kruskal-Wallis Test of the Effect of REXP (Filtered)

6.6.4 One-Way ANOVA on IREXP

Table 6.33 shows that the Levene test result of the unfiltered dependent variables plotted against IREXP is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these unfiltered variables.

Table 6.34 shows that the Levene test result of the filtered dependent variables plotted against IREXP is not significant for each of the four dependent variables. Therefore, 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 6.33: 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 6.34: Results of the Levene Test for IREXP (Filtered)

Table 6.35 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 6.36 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.

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 6.35: 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 6.36: Results of the One-Way ANOVA of the Effect of IREXP (Filtered)

Tables 6.37 and 6.38 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 6.37: Results of the Kruskal-Wallis Test of the Effect of IREXP (Unfiltered)

Dependent Variable	*p*
AVG_I	.060

Table 6.38: Results of the Kruskal-Wallis Test of the Effect of IREXP (Filtered)

6.6.5 One-Way ANOVA on IEXP

Table 6.39 shows that the Levene test result of the unfiltered dependent variables plotted against IEXP is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these unfiltered variables.

Table 6.40 shows that the Levene test result of the filtered dependent variables plotted against IEXP is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these filtered variables.

Table 6.41 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 6.42 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	*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 6.39: 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 6.40: Results of the Levene Test for IEXP (Filtered)

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 6.41: 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 6.42: Results of the One-Way ANOVA of the Effect of IEXP (Filtered)

Tables 6.43 and 6.44 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 6.43: Results of the Kruskal-Wallis Test of the Effect of IEXP (Unfiltered)

Dependent Variable	*p*
AVG_I	.504

Table 6.44: Results of the Kruskal-Wallis Test of the Effect of IEXP (Filtered)

6.6.6 One-Way ANOVA on NCS

Table 6.45 shows that the Levene test result of the unfiltered dependent variables plotted against NCS is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these unfiltered variables.

Table 6.46 shows that the Levene test result of the filtered dependent variables plotted against NCS is not significant for each of the four dependent variables. Therefore, 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 6.45: Results of the Levene Test for NCS (Unfiltered)

Table 6.47 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 6.48 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.

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 6.46: 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 6.47: 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 6.48: Results of the One-Way ANOVA of the Effect of NCS (Filtered)

Table 6.49 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 and NCS = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Standard 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 6.49: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NCS on NF (Filtered)

Table 6.50 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 and NCS = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Standard 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 6.50: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NCS on NI (Filtered)

Table 6.51 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 6.52 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 6.51: Results of the Kruskal-Wallis Test of the Effect of NCS (Unfiltered)

Dependent Variable	*p*
AVG_I	.010

Table 6.52: Results of the Kruskal-Wallis Test of the Effect of NCS (Filtered)

Table 6.53 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 and NCS = 3.

Sample 1	*Sample 2*	*Test Statistic*	*Standard Error*	*Standard 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 6.53: Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of NCS on AVG_I (Filtered)

6.6.7 One-Way ANOVA on NSE

Table 6.54 shows that the Levene test result of the unfiltered dependent variables plotted against NSE is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these unfiltered variables.

Table 6.55 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. Therefore, 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 6.54: 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 6.55: Results of the Levene Test for NSE (Filtered)

Table 6.56 shows the results of the ANOVA test of the unfiltered dependent variables plotted against NSE. The test shows no significant effect of the team's NSE on each of NRAW, NR, and NI, but shows a significant effect of the team's NSE on NF.

Table 6.57 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.

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 6.56: 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 6.57: Results of the One-Way ANOVA of the Effect of NSE (Filtered)

Table 6.58 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 and NSE = 2 and when NSE = 0 and NSE = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Standard 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 6.58: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NSE on NF (Unfiltered)

Table 6.59 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 and NSE = 2, when NSE = 0 and NSE = 3, and when NSE = 1 and NSE = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Standard 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 6.59: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NSE on NF (Filtered)

Table 6.60 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 and NSE = 2 and when NSE = 0 and NSE = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Standard 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 6.60: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NSE on NI (Filtered)

Table 6.61 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 6.62 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 6.61: Results of the Kruskal-Wallis Test of the Effect of NSE (Unfiltered)

Dependent Variable	*p*
AVG_R	.538
AVG_I	.005

Table 6.62: Results of the Kruskal-Wallis Test of the Effect of NSE (Filtered)

Table 6.63 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 and NSE = 3.

Sample 1	*Sample 2*	*Test Statistic*	*Standard Error*	*Standard 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 6.63: Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of NSE on AVG_I (Filtered)

6.6.8 One-Way ANOVA on NGRAD

Table 6.64 shows that the Levene test result of the unfiltered dependent variables plotted against NGRAD is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these unfiltered variables.

Table 6.65 shows that the Levene test result of the filtered dependent variables plotted against NGRAD is not significant for each of the four dependent variables. Therefore, 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 6.64: Results of the Levene Test for NGRAD (Unfiltered)

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 6.65: Results of the Levene Test for NGRAD (Filtered)

Table 6.66 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 6.67 shows the results of the ANOVA test 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.

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 6.66: 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 6.67: Results of the One-Way ANOVA of the Effect of NGRAD (Filtered)

Table 6.68 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 and NGRAD = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Standard 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 6.68: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NGRAD on NRAW (Filtered)

Table 6.69 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 and NGRAD = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Standard 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 6.69: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NGRAD on NF (Filtered)

Table 6.70 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 and NGRAD = 3.

Sample 1	*Sample 2*	*Mean Difference*	*Standard 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 6.70: Results of the Tukey HSD Pairwise Comparison Test of the Effect of NGRAD on NI (Filtered)

Table 6.71 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 6.72 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 6.71: Results of the Kruskal-Wallis Test of the Effect of NGRAD (Unfiltered)

Dependent Variable	*p*
AVG_I	.018

Table 6.72: Results of the Kruskal-Wallis Test of the Effect of NGRAD (Filtered)

Table 6.73 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 and NGRAD = 3.

Sample 1	*Sample 2*	*Test Statistic*	*Standard Error*	*Standard 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 6.73: Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of NGRAD on AVG_I (Filtered)

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

Table 6.74 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. Therefore, an ANOVA is applicable to the unfiltered NRAW, NR, and NF, but is not applicable to the unfiltered NI.

Table 6.75 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. Therefore, 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 6.74: 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 6.75: 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. I 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 6.76 shows the results of the three-way ANOVA test of the unfiltered dependent variables plotted against MIX, EXP, and EDU.

Source	*Dependent Variable*	*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*	*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 6.76: Results of the Three-Way ANOVA of the Effect of MIX, EXP, and EDU (Unfiltered)

Table 6.77 shows the results of the three-way ANOVA test of the filtered dependent variables plotted against MIX, EXP, and EDU.

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 is not enough, SPSS outputs a value of 0 for sum of squares and degrees of freedom and "." for the other fields.

Table 6.77: Results of the Three-Way ANOVA of the Effect of MIX, EXP, and EDU (Filtered)

Based on Table 6.76, the three-way ANOVA on the unfiltered dependent variables 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.
5. The rest of the interactions do not significantly affect any type of ideas.

Based on Table 6.77, the three-way ANOVA on the filtered dependent variables 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. For instance, 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.
6. The rest of the interactions do not significantly affect any type of ideas.

6.6.9.1 Interactions

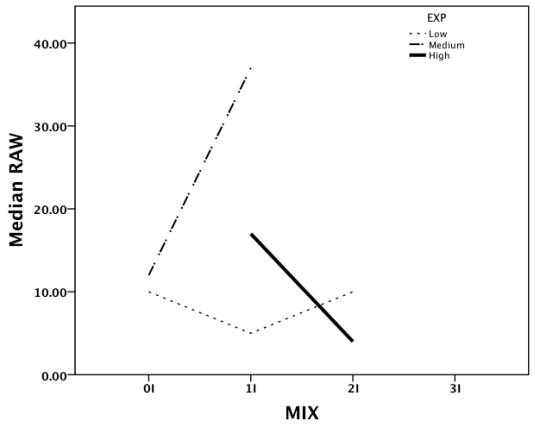
The three-way ANOVA results represented in Table 6.76 reveal that the interaction between MIX, EXP, and EDU on the unfiltered NRAW and NR is significant. Table 6.77 shows that also the interaction between EXP and EDU on the filtered NRAW is significant.

6.6.9.2 MIX * EXP * EDU (Unfiltered)

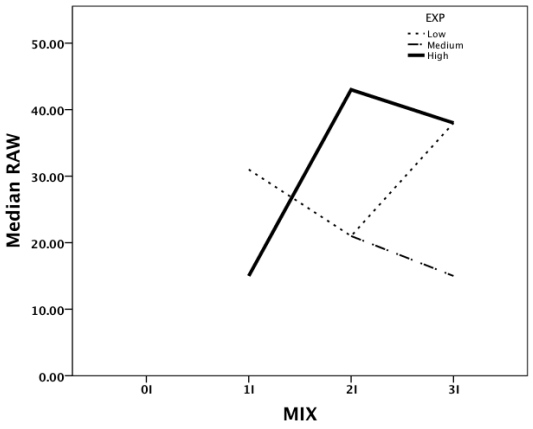
Figure 6.29 shows the interactions between three independent variables of MIX, EXP, and EDU on the unfiltered RAW and AVG_R. It is not possible to show interactions of three independent variables in a single plot. Thus, one of the independent variables, EDU, is fixed and the plots are provided for each value of EDU.

An issue with the sub-plots of Figure 6.29 is that there are not enough data points to show the interactions between all values of the affecting independent variables. Also, comparing Figure 6.29(a) with Figure 6.29(b) and Figure 6.29(c) with Figure 6.29(d), the correlations seem to be contradictory for EXP = "Low" and EXP = "High". All in all, the plots do not show anything interesting.

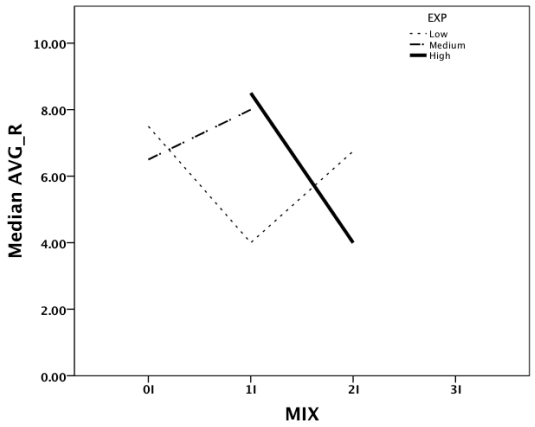
One possible explanation for the interactions shown in Figure 6.29 is that the less educated in computer science a team is, the more a higher level of overall experience helps in generating raw requirement ideas. Conversely the more educated in computer science a team is, the less a higher level of overall experience helps in generating raw requirement ideas.



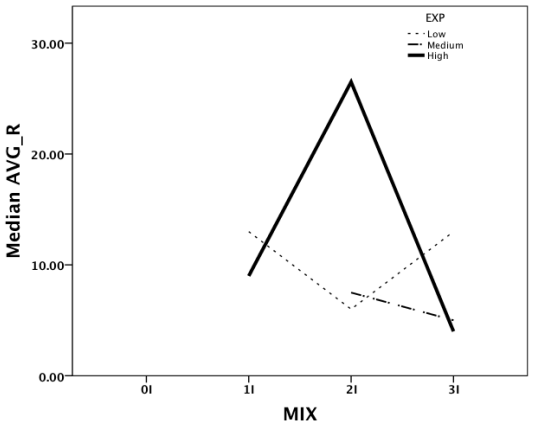
(a) RAW vs. MIX * EXP * EDU (EDU = Low)



(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)

Figure 6.29: Ideas vs. MIX * EXP * EDU (Unfiltered)

6.6.9.3 EXP * EDU (Filtered)

Figure 6.30 shows the interactions between two independent variables of EXP and EDU on the filtered RAW. The plot shows that the medians of the filtered RAW generated by teams with EDU = "Low" is positively correlated with the teams' EXP. On the other hand, the medians of the filtered RAW generated by teams' with EDU = "High" is negatively correlated with the teams' EXP.

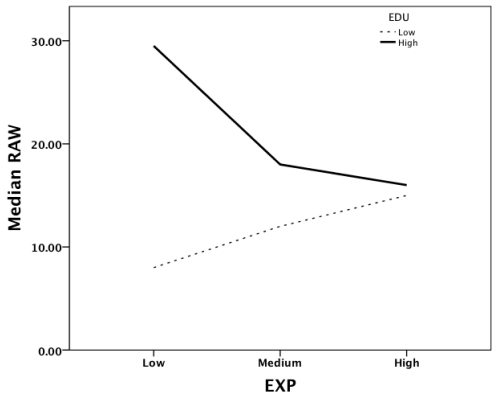


Figure 6.30: RAW vs. EXP * EDU (Filtered)

6.6.10 One-Way ANOVA on EDU

Table 6.78 shows that the Levene test result of the unfiltered dependent variables plotted against EDU is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these unfiltered variables.

Table 6.79 shows that the Levene test result of the filtered dependent variables plotted against EDU is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these filtered variables.

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 6.78: Results of the Levene Test for EDU (Unfiltered)

Table 6.80 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 6.81 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.

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 6.79: Results of the Levene Test for EDU (Filtered)

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 6.80: 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 6.81: Results of the One-Way ANOVA of the Effect of EDU (Filtered)

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 6.82 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 6.83 shows the results of the Kruskal-Wallis 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 6.82: Results of the Kruskal-Wallis Test of the Effect of EDU (Unfiltered)

Dependent Variable	*p*
AVG_I	.000

Table 6.83: Results of the Kruskal-Wallis Test of the Effect of EDU (Filtered)

6.6.11 One-Way ANOVA on EXP

Table 6.84 shows that the Levene test result of the unfiltered dependent variables plotted against EXP is not significant for each of the four dependent variables. Therefore, an ANOVA is applicable to each of these unfiltered variables.

Table 6.85 shows that the Levene test result of the filtered dependent variables plotted against EXP is not significant for each of the four dependent variables. Therefore, 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 6.84: Results of the Levene Test for EXP (Unfiltered)

Table 6.86 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 6.87 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.

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 6.85: 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 6.86: 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 6.87: Results of the One-Way ANOVA of the Effect of EXP (Filtered)

Table 6.88 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 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*	*Standard Error*	*p*
Low	Medium	-.875	.297	.015
	High	-.111	.352	.947
Medium	High	.764	.340	.076

Table 6.88: Results of the Tukey HSD Pairwise Comparison Test of the Effect of EXP on NI (Unfiltered)

Table 6.89 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" and EXP = "Medium" and when EXP = "Medium" and EXP = "High".

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

Table 6.89: Results of the Tukey HSD Pairwise Comparison Test of the Effect of EXP on NI (Filtered)

Table 6.90 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 6.91 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 6.90: Results of the Kruskal-Wallis Test of the Effect of EXP (Unfiltered)

Table 6.92 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

Dependent Variable	*p*
AVG_I	.013

Table 6.91: Results of the Kruskal-Wallis Test of the Effect of EXP (Filtered)

between the medians of the AVG_I of the teams is significant when EXP = "Low" and EXP = "Medium".

Table 6.93 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" and EXP = "Medium".

Sample 1 - Sample 2	*Test Statistic*	*Standard Error*	*Standard 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 6.92: Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of EXP on AVG_I (Unfiltered)

Sample 1 - Sample 2	*Test Statistic*	*Standard Error*	*Standard 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 6.93: Results of the Dunn-Bonferroni Pairwise Comparison Test of the Effect of EXP on AVG_I (Filtered)

6.7 Threats to Validity

As discussed in Section 6.2, there were differences between the ratios of the ideas classified in E1 and E2. In order to determine if these differences affected the results, I tried adjusting the E2

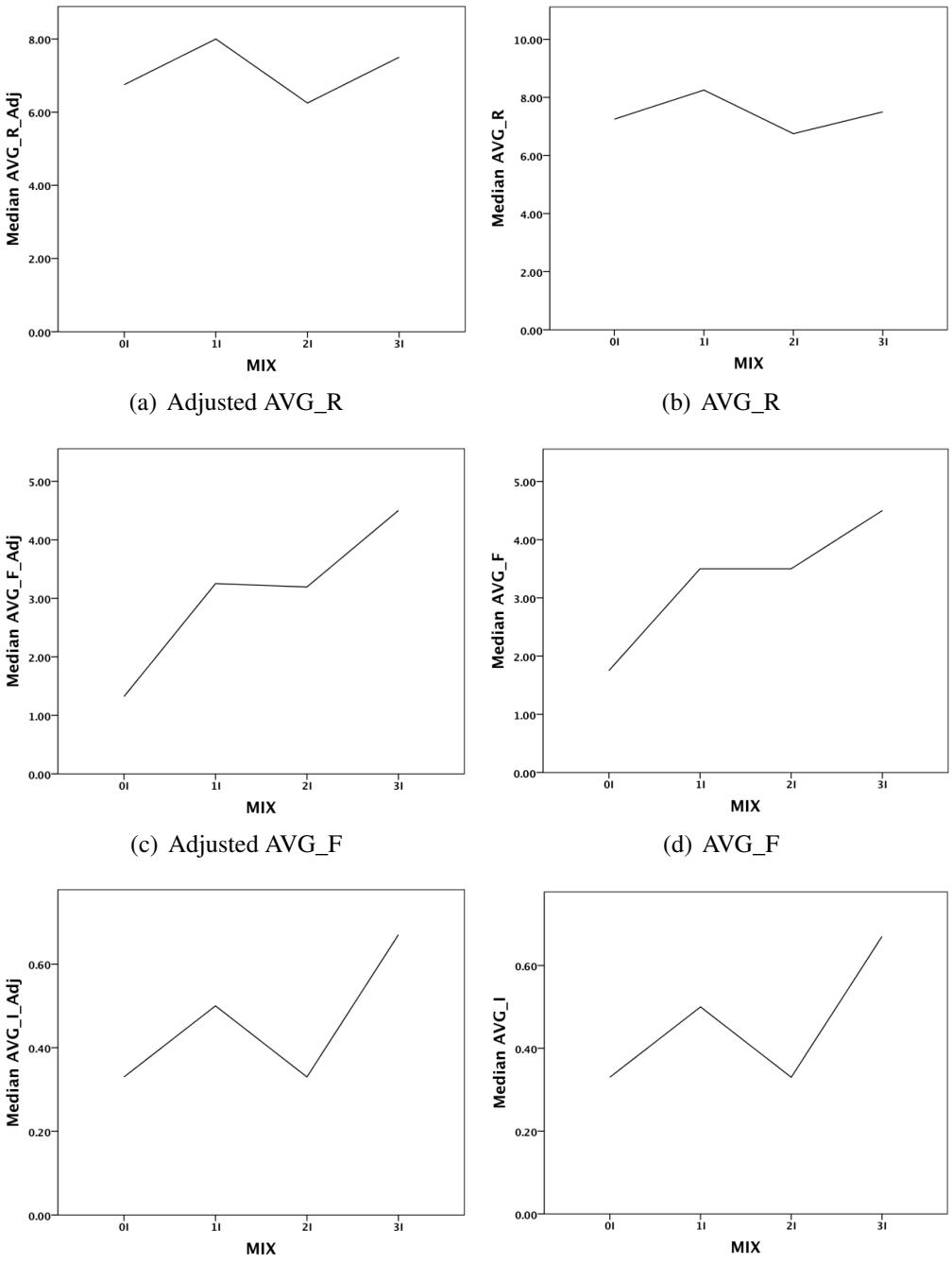
data to equalize the ratios between the two experiments. Therefore, the number of ideas of each type of idea, *T*, was multiplied by

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

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

Graphs of the correlations between the adjusted data generated by teams and the dependent variables are shown in the following figures. Figures 6.31 through 6.46 show each of the graphs from Figures 6.9 through 6.24 to the right of the corresponding graphs of the adjusted data. It is clear from examining these adjacent graphs that none of the correlations observed in Section 6.5 have changed to the point that the conclusions drawn in Section 6.8 would have to be changed. Specifically, these graphs 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 the same direction with the corresponding graphs of the unadjusted data. The new graphs 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 6.2 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. 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 subjective data.

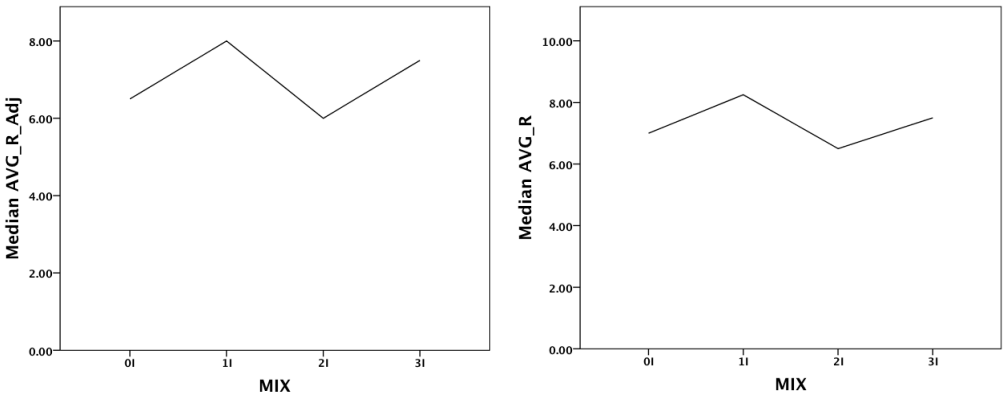


(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

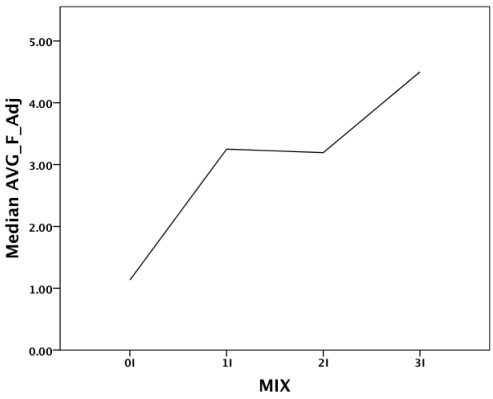
(e) Adjusted AVG_I (f) AVG_I

Figure 6.31: Adjusted Ideas vs. MIX – Ideas vs. MIX (Unfiltered)

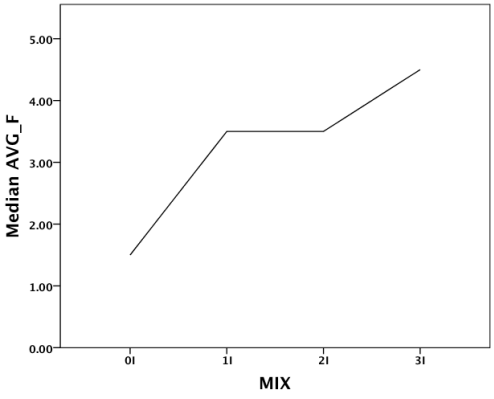


(a) Adjusted AVG_R

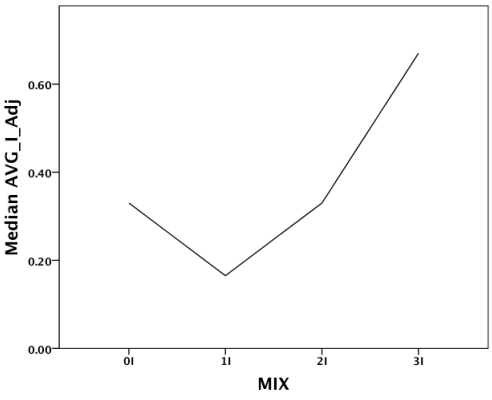
(b) AVG_R



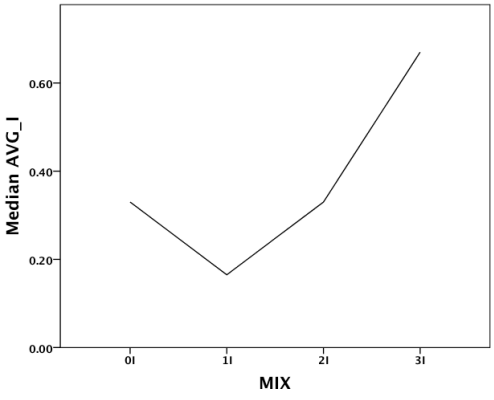
(c) Adjusted AVG_F



(d) AVG_F

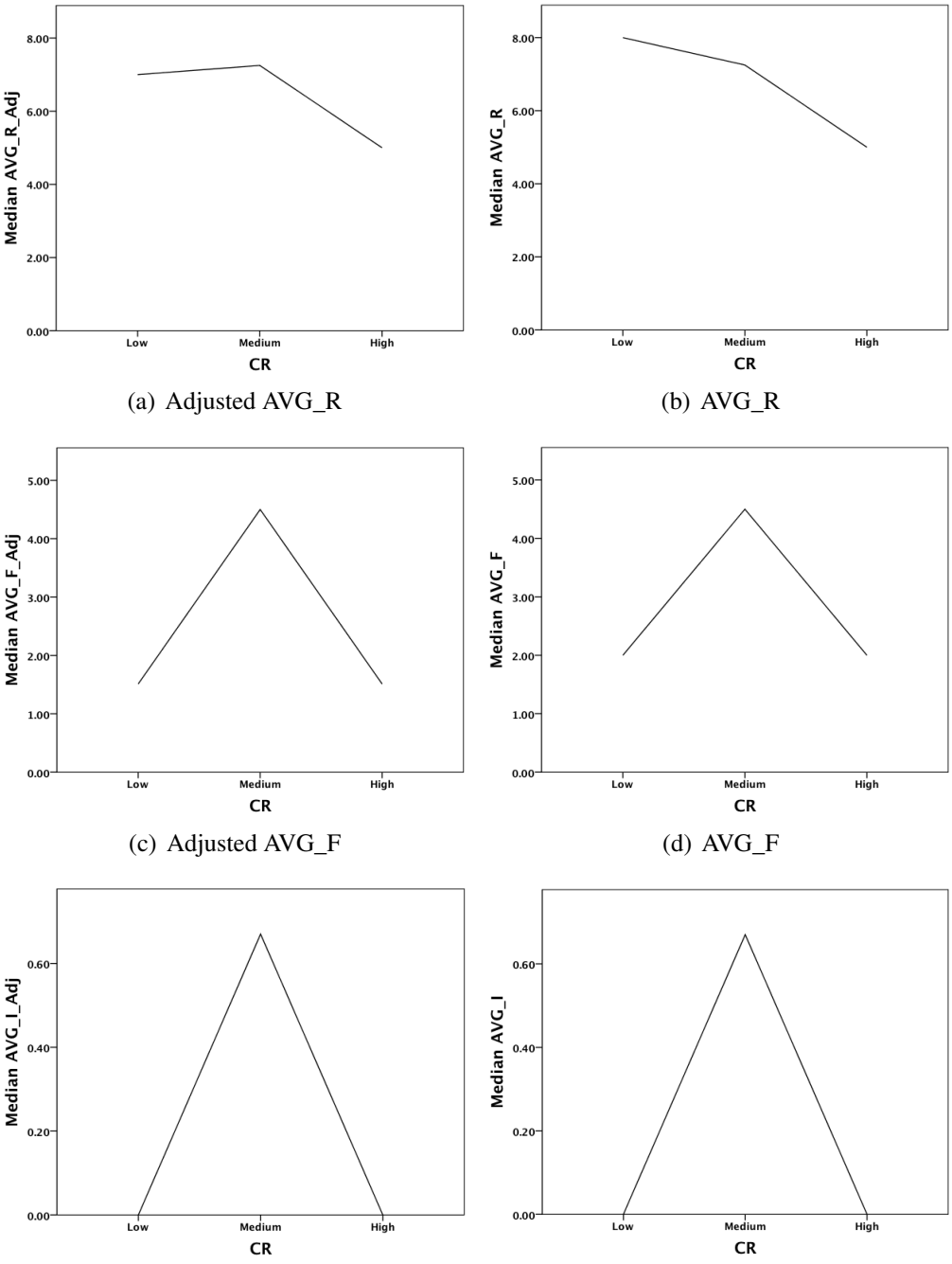


(e) Adjusted AVG_I



(f) AVG_I

Figure 6.32: Adjusted Ideas vs. MIX – Ideas vs. MIX (Filtered)

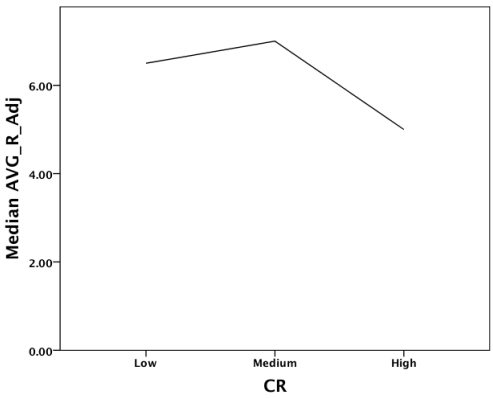


(a) Adjusted AVG_R (b) AVG_R

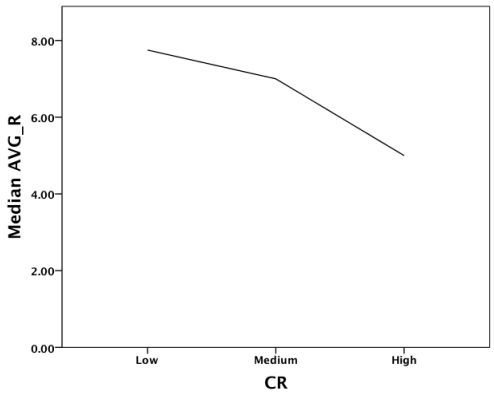
(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

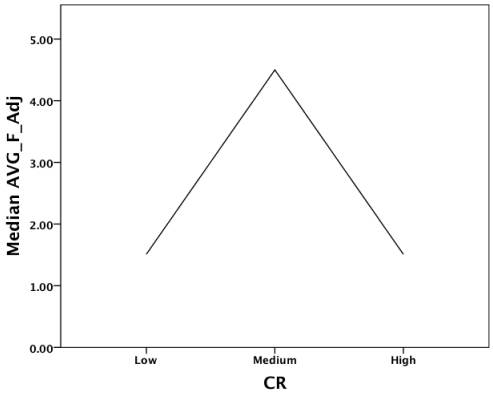
Figure 6.33: Adjusted Ideas vs. CR – Ideas vs. CR (Unfiltered)



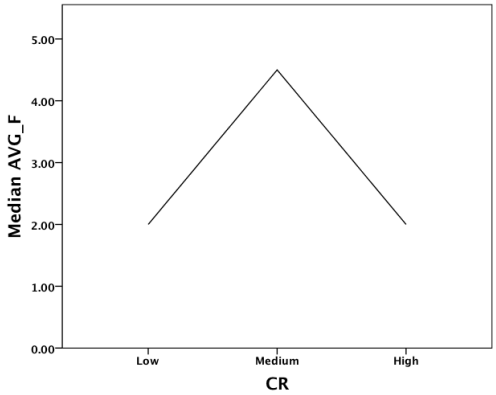
(a) Adjusted AVG_R



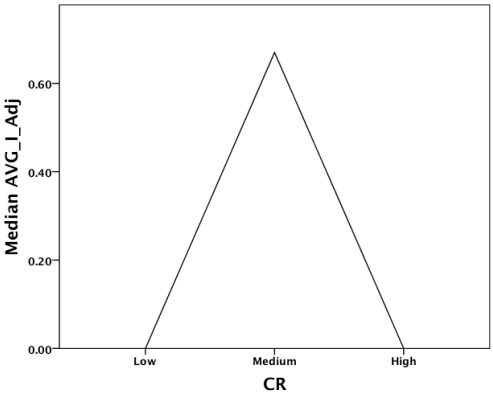
(b) AVG_R



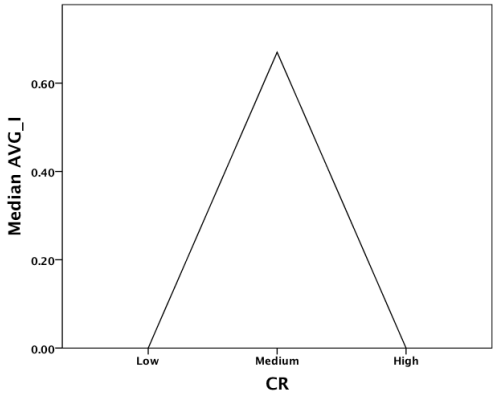
(c) Adjusted AVG_F



(d) AVG_F

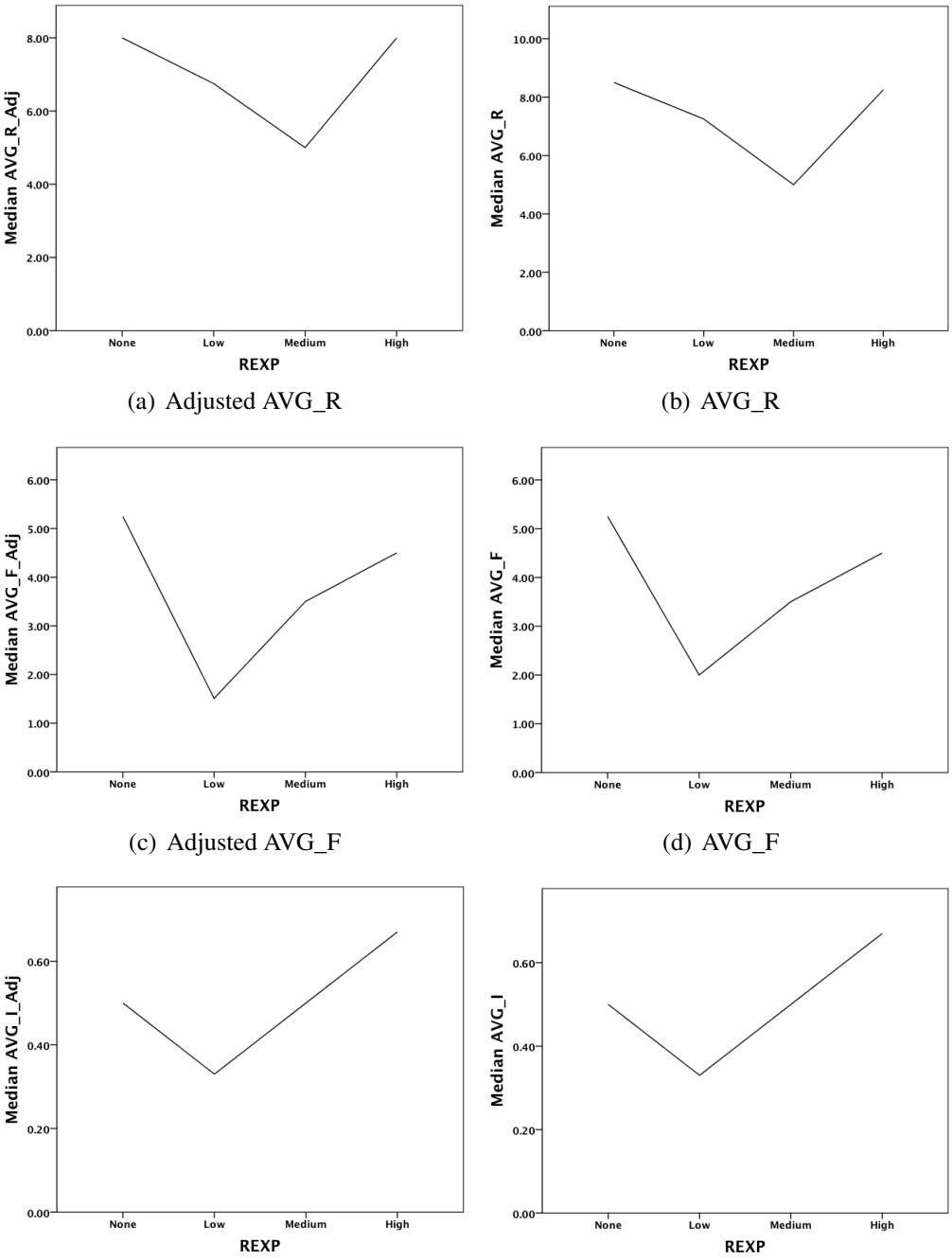


(e) Adjusted AVG_I



(f) AVG_I

Figure 6.34: Adjusted Ideas vs. CR – Ideas vs. CR (Filtered)

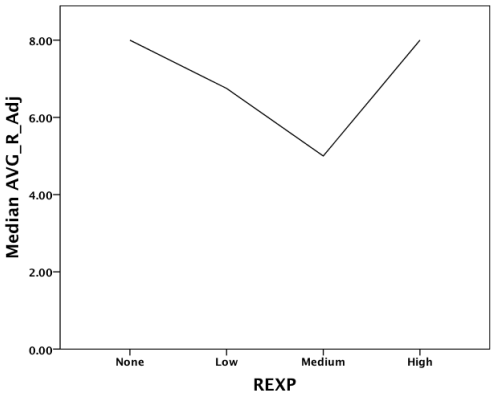


(a) Adjusted AVG_R (b) AVG_R

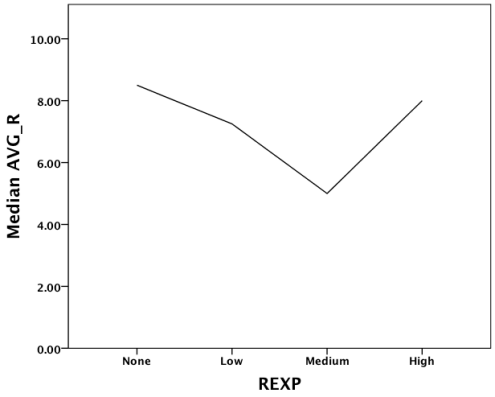
(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

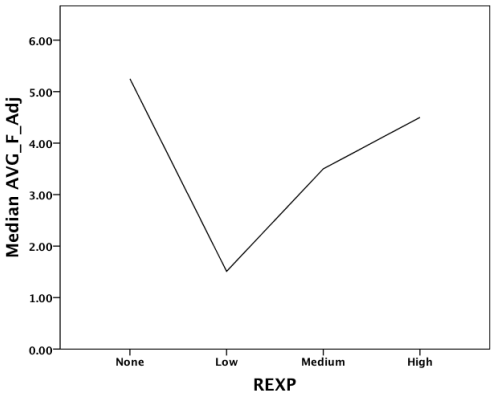
Figure 6.35: Adjusted Ideas vs. REXP – Ideas vs. REXP (Unfiltered)



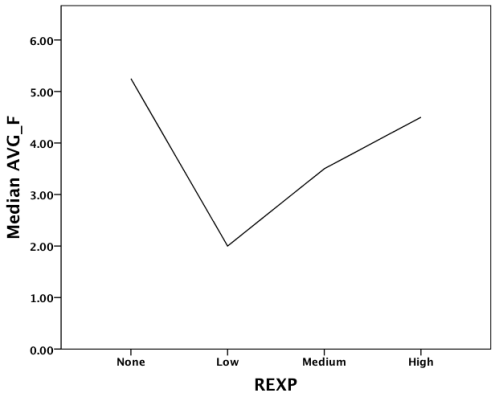
(a) Adjusted AVG_R



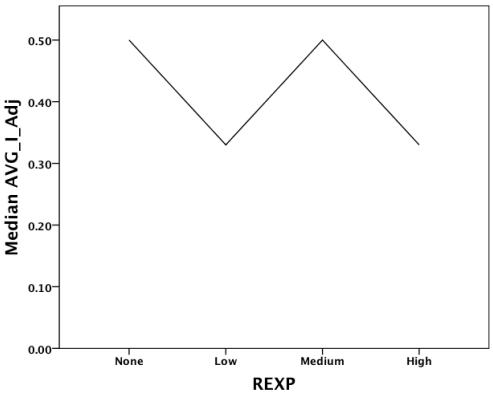
(b) AVG_R



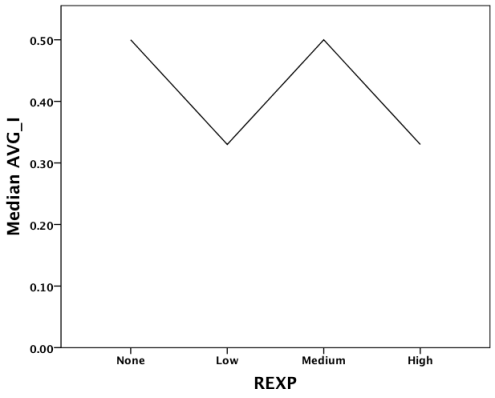
(c) Adjusted AVG_F



(d) AVG_F

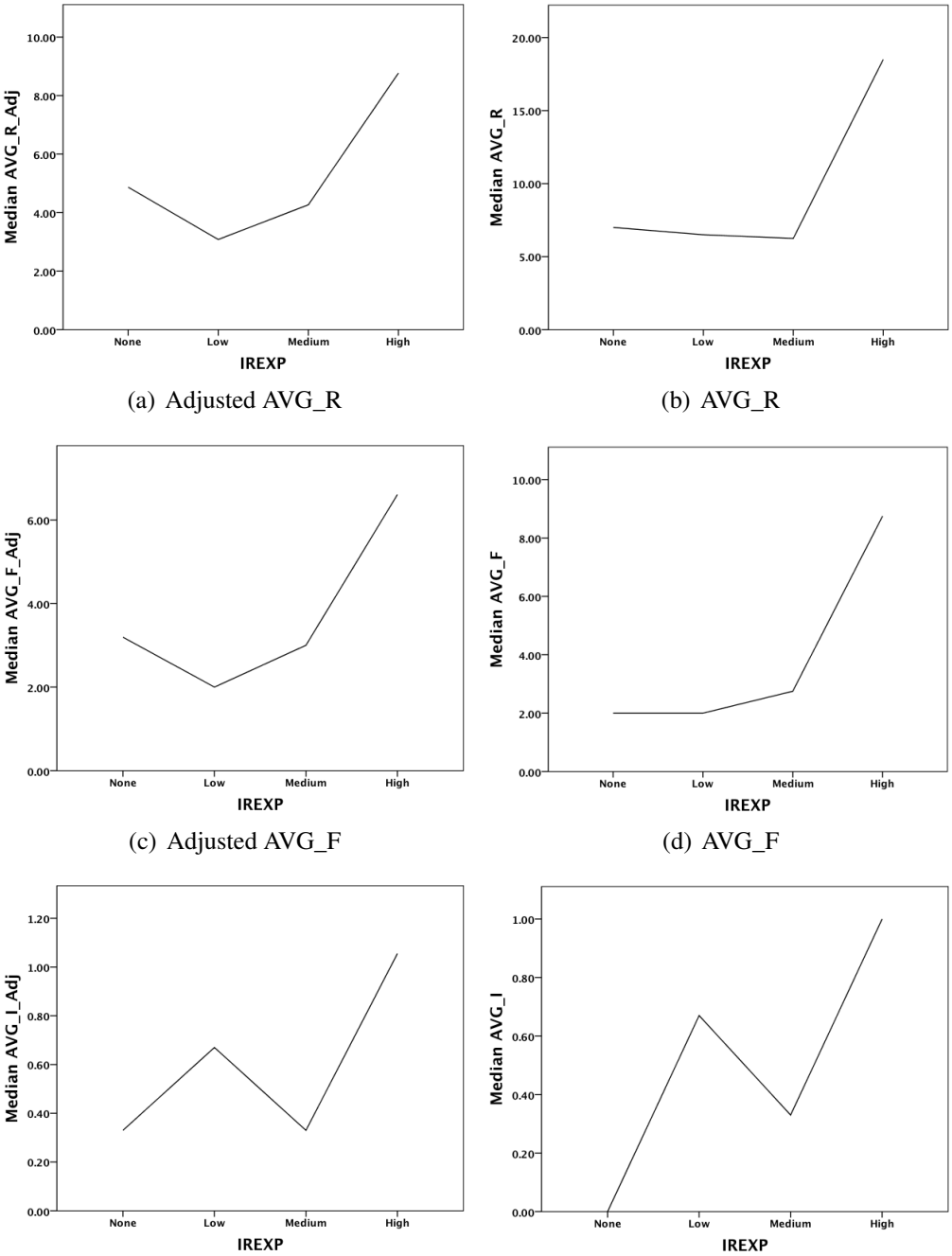


(e) Adjusted AVG_I



(f) AVG_I

Figure 6.36: Adjusted Ideas vs. REXP – Ideas vs. REXP (Filtered)

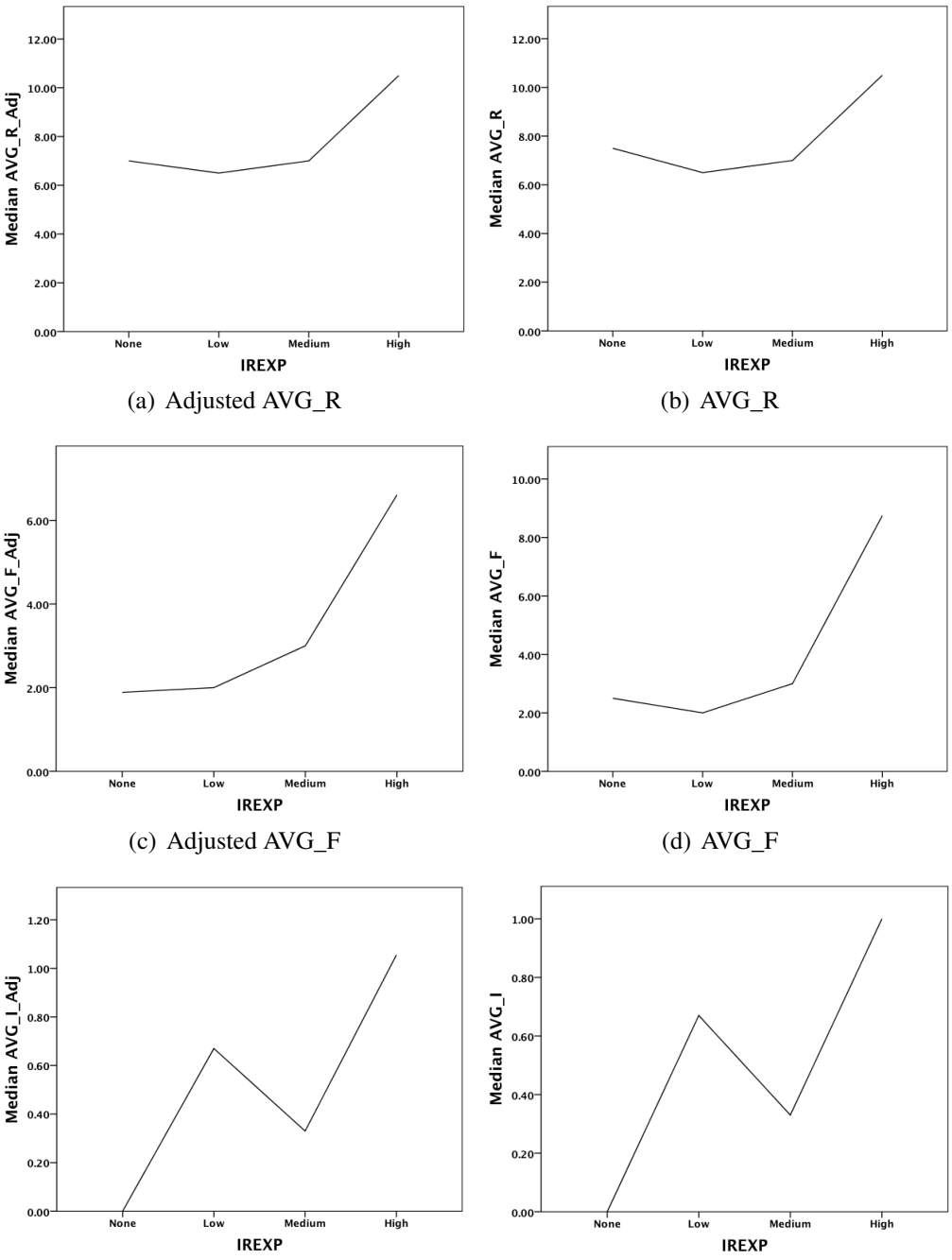


(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

Figure 6.37: Adjusted Ideas vs. IREXP – Ideas vs. IREXP (Unfiltered)

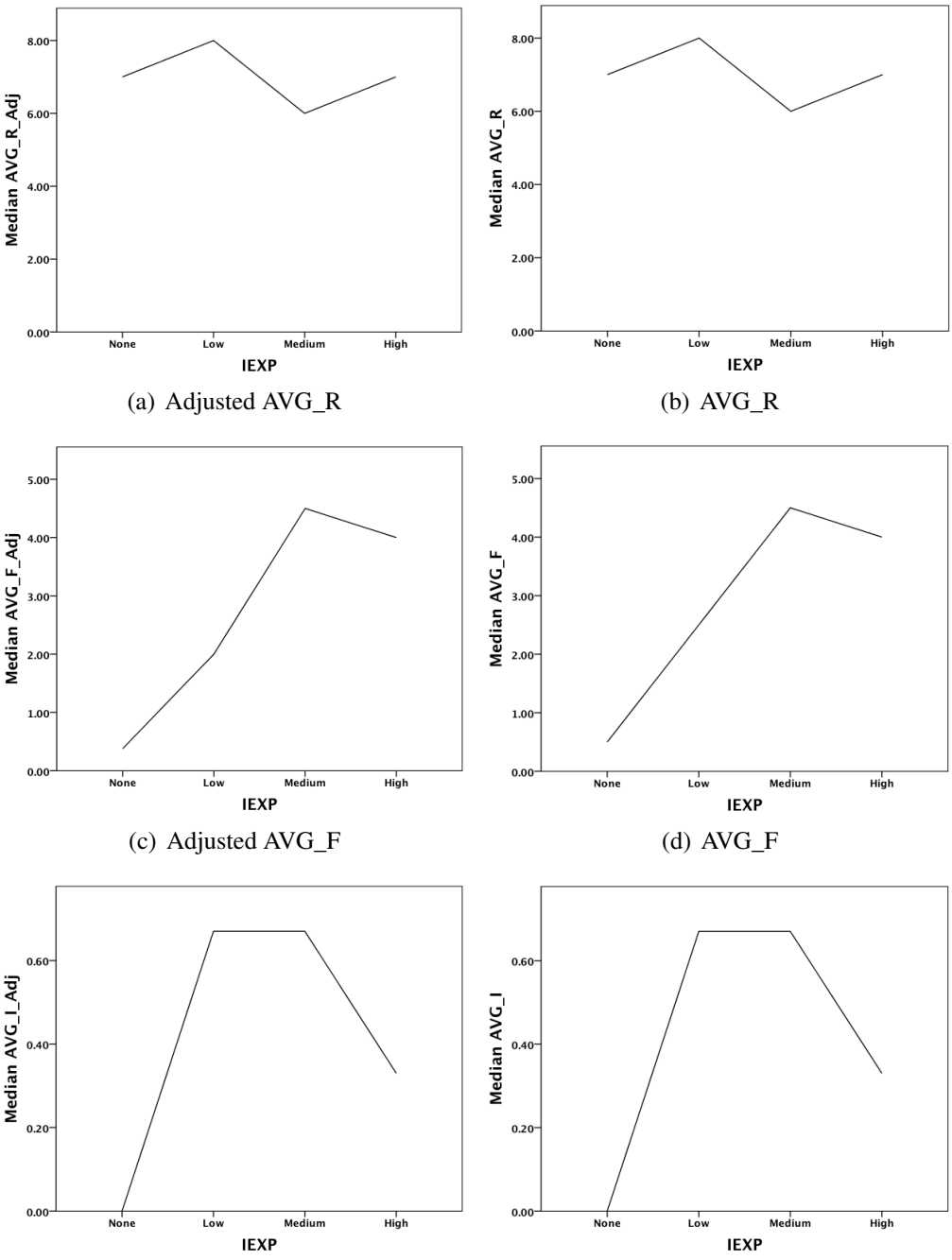


(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

Figure 6.38: Adjusted Ideas vs. IREXP – Ideas vs. IREXP (Filtered)

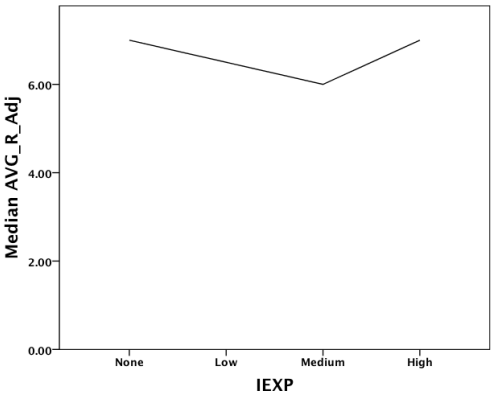


(a) Adjusted AVG_R (b) AVG_R

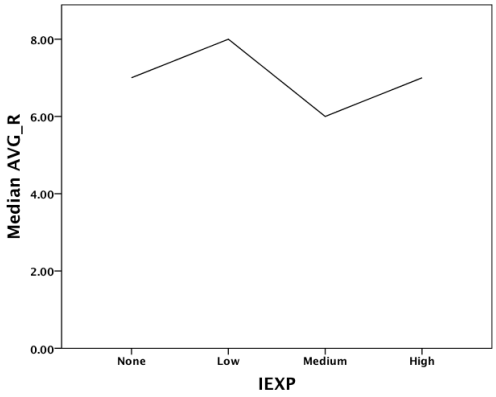
(c) Adjusted AVG_F (d) AVG_F

(e) Adjusted AVG_I (f) AVG_I

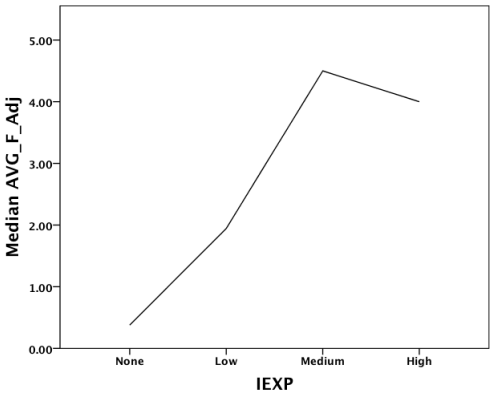
Figure 6.39: Adjusted Ideas vs. IEXP – Ideas vs. IEXP (Unfiltered)



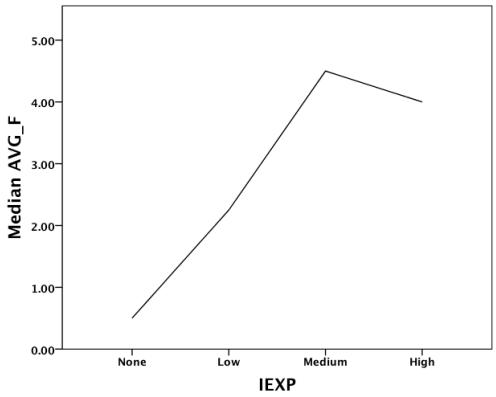
(a) Adjusted AVG_R



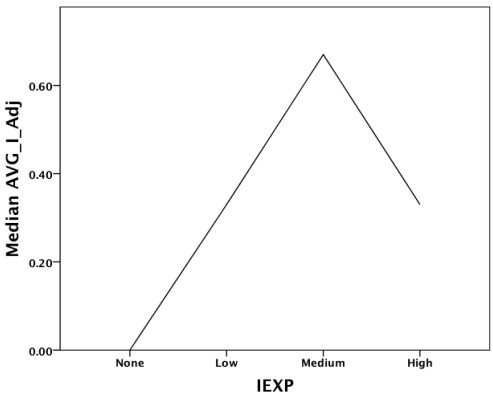
(b) AVG_R



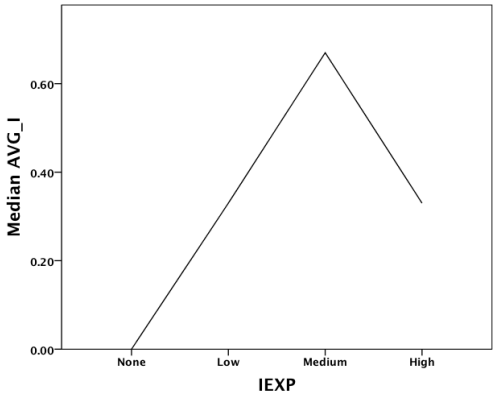
(c) Adjusted AVG_F



(d) AVG_F

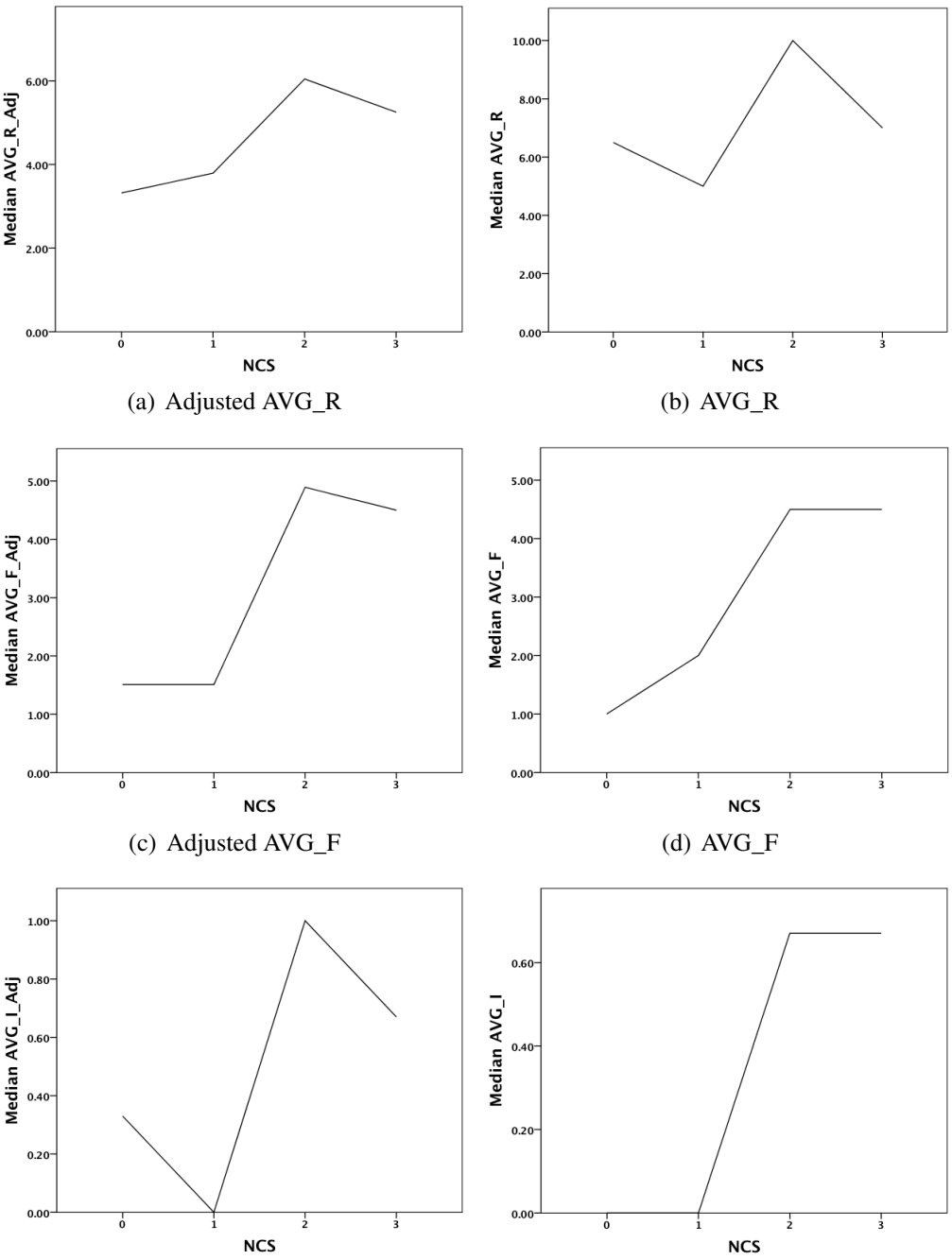


(e) Adjusted AVG_I



(f) AVG_I

Figure 6.40: Adjusted Ideas vs. IEXP – Ideas vs. IEXP (Filtered)

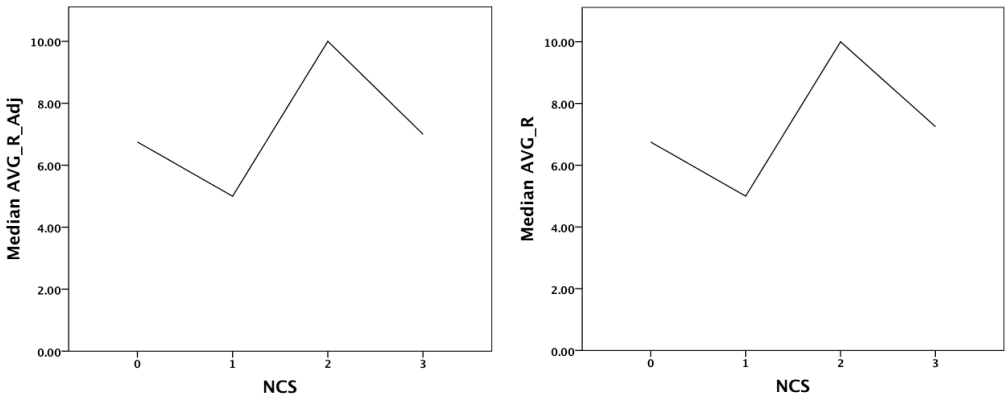


(a) Adjusted AVG_R (b) AVG_R

(c) Adjusted AVG_F (d) AVG_F

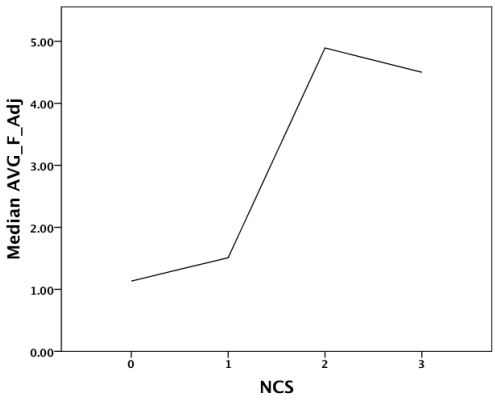
(e) Adjusted AVG_I (f) AVG_I

Figure 6.41: Adjusted Ideas vs. NCS – Ideas vs. NCS (Unfiltered)

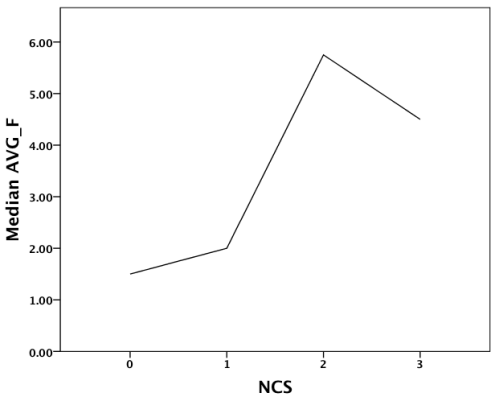


(a) Adjusted AVG_R

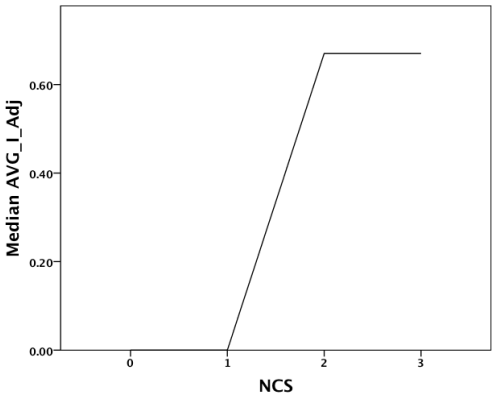
(b) AVG_R



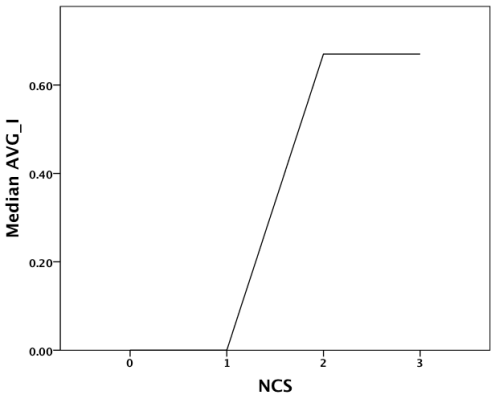
(c) Adjusted AVG_F



(d) AVG_F

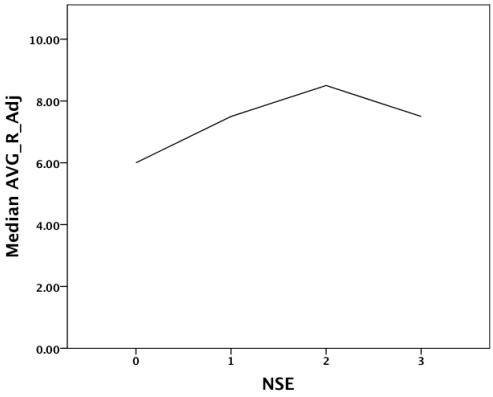


(e) Adjusted AVG_I

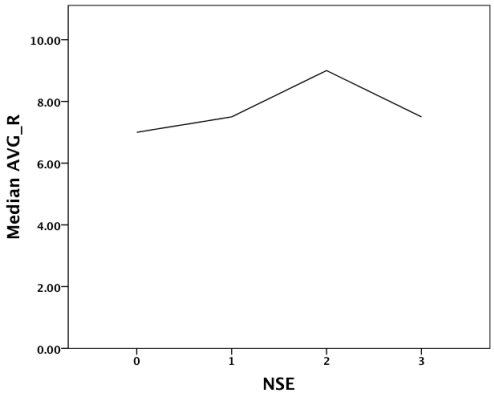


(f) AVG_I

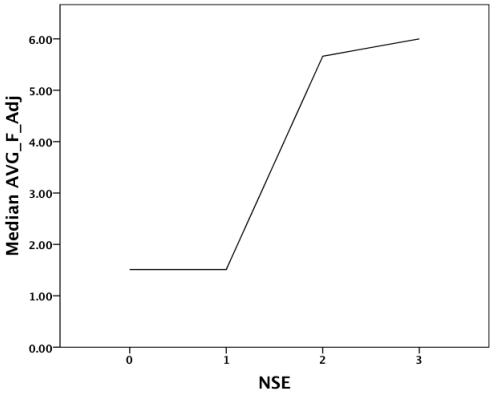
Figure 6.42: Adjusted Ideas vs. NCS – Ideas vs. NCS (Filtered)



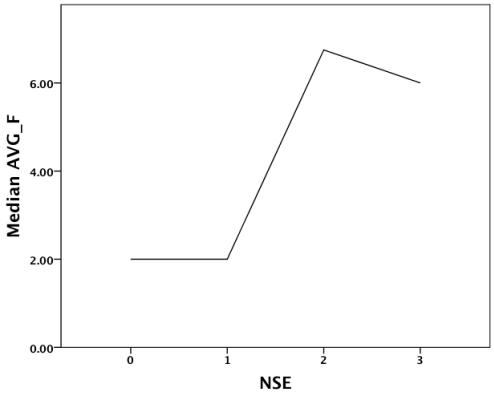
(a) Adjusted AVG_R



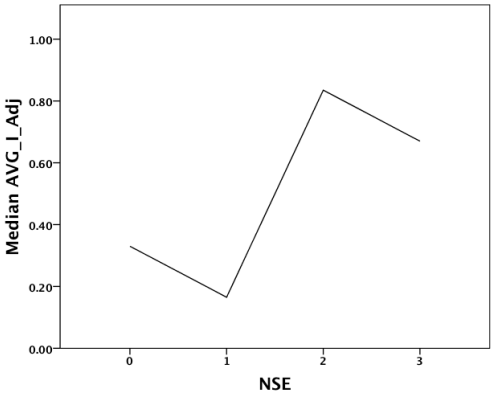
(b) AVG_R



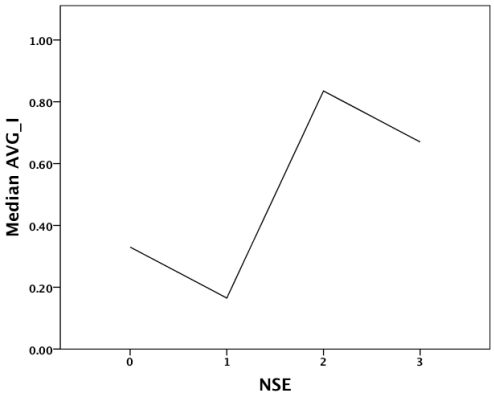
(c) Adjusted AVG_F



(d) AVG_F

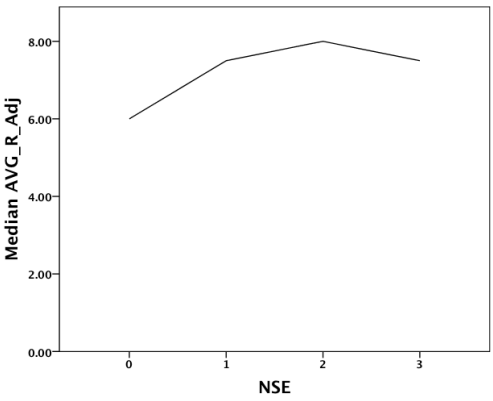


(e) Adjusted AVG_I

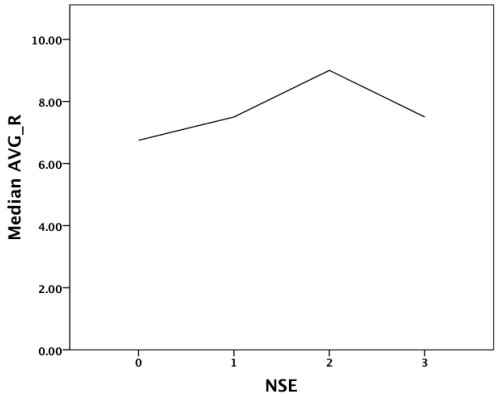


(f) AVG_I

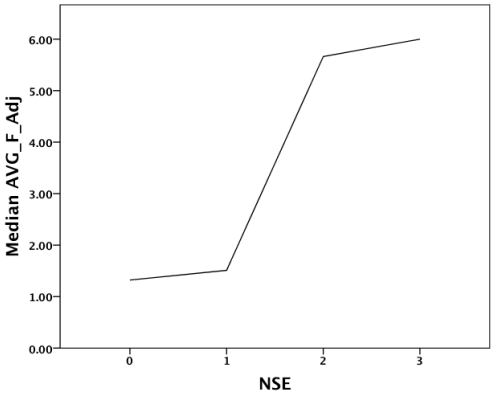
Figure 6.43: Adjusted Ideas vs. NSE – Ideas vs. NSE (Unfiltered)



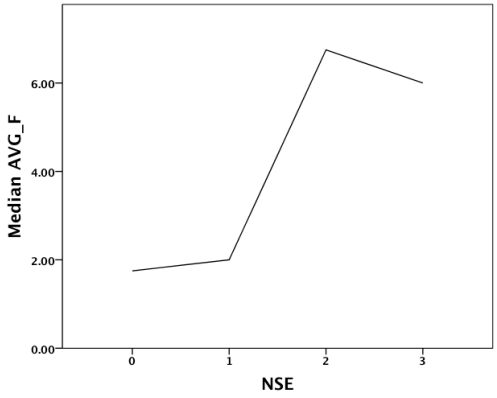
(a) Adjusted AVG_R



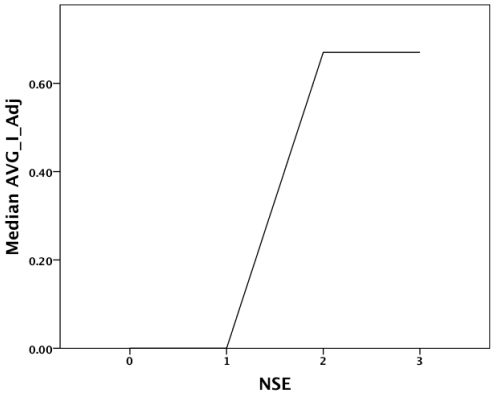
(b) AVG_R



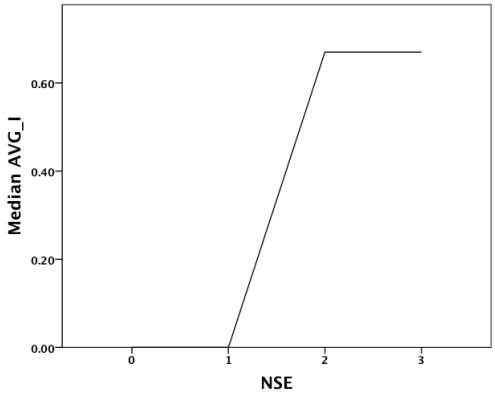
(c) Adjusted AVG_F



(d) AVG_F

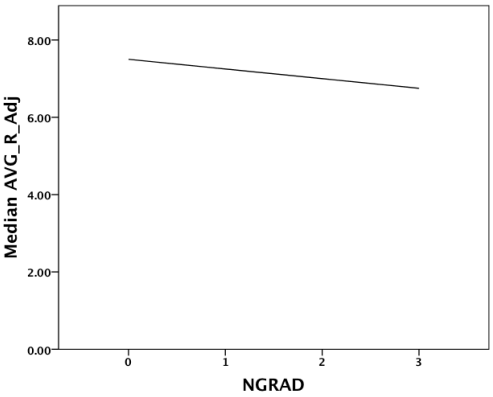


(e) Adjusted AVG_I

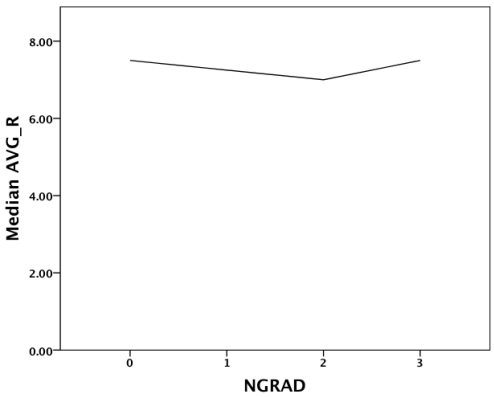


(f) AVG_I

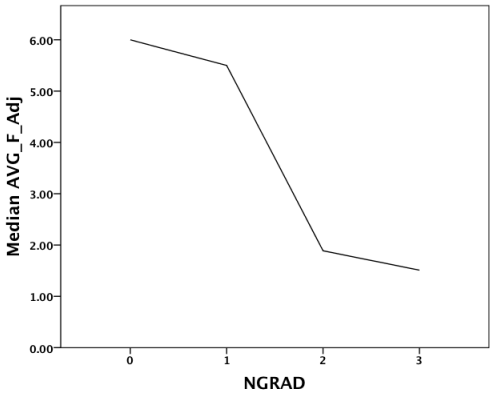
Figure 6.44: Adjusted Ideas vs. NSE – Ideas vs. NSE (Filtered)



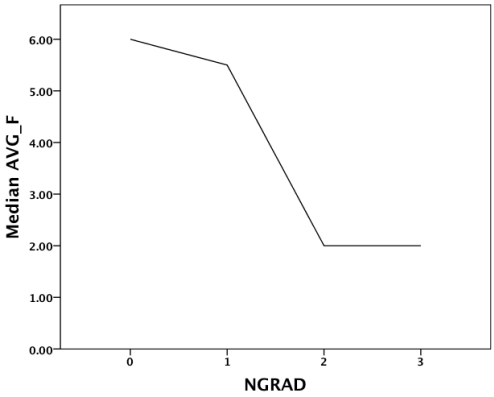
(a) Adjusted AVG_R



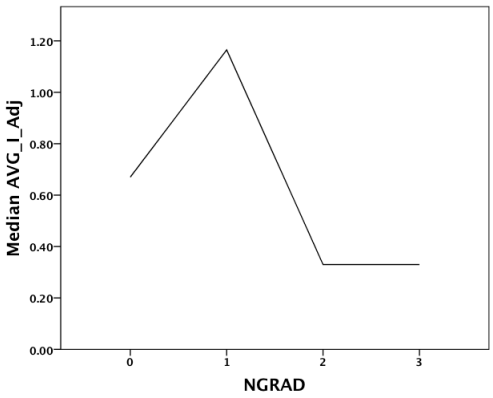
(b) AVG_R



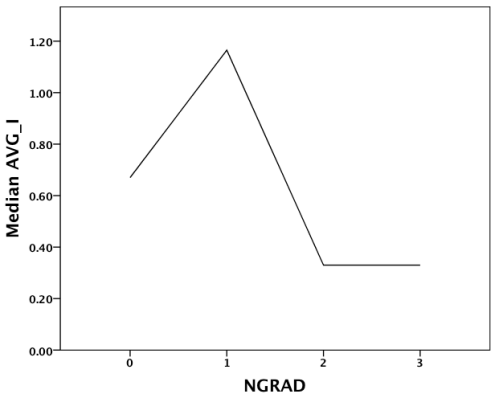
(c) Adjusted AVG_F



(d) AVG_F

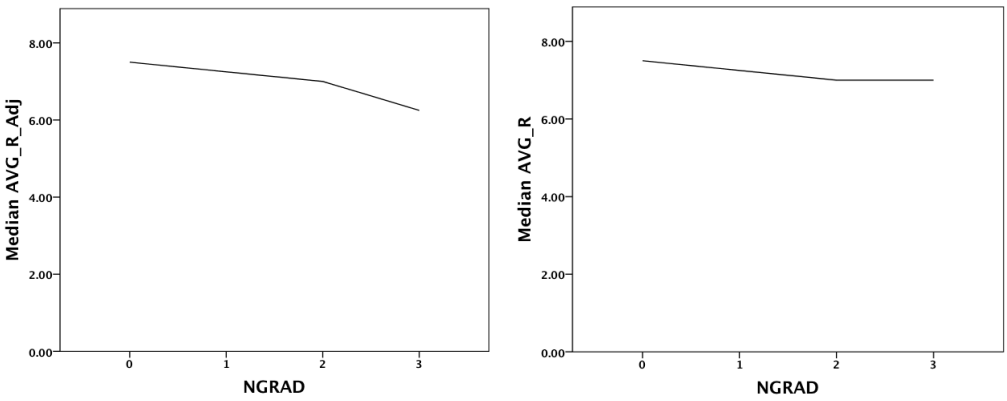


(e) Adjusted AVG_I



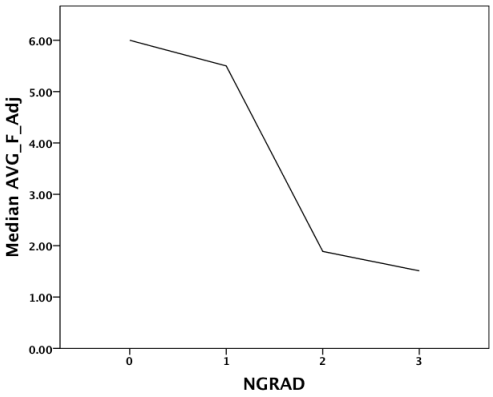
(f) AVG_I

Figure 6.45: Adjusted Ideas vs. NGRAD – Ideas vs. NGRAD (Unfiltered)

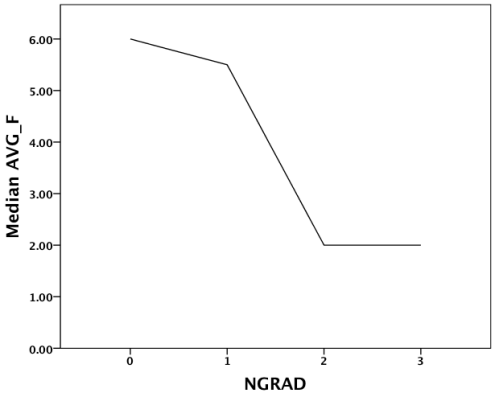


(a) Adjusted AVG_R

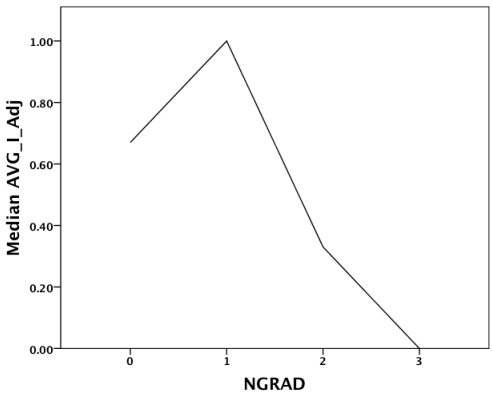
(b) AVG_R



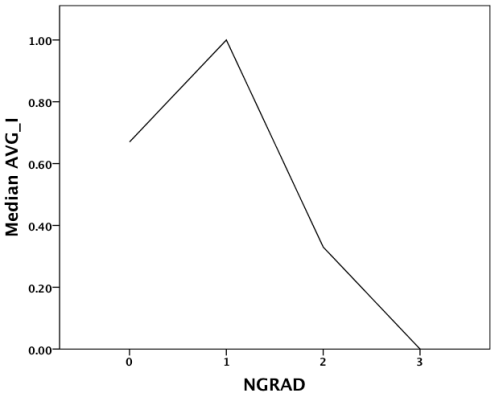
(c) Adjusted AVG_F



(d) AVG_F



(e) Adjusted AVG_I



(f) AVG_I

Figure 6.46: Adjusted Ideas vs. NGRAD – Ideas vs. NGRAD (Filtered)

6.8 Conclusion

The data of the aggregated results of the combined controlled experiments were analyzed to find any statistically significant results. A factor analysis was conducted first to reveal the most influential variables. The found factors replaced five variables to give the final set of four independent variables. Statistical analyses were performed next on the eight original independent variables plus the two factors identified by the factor analysis. Table 6.94 summarizes the initial observations of Section 6.5 and the statistical analysis results of Section 6.6.

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.

In general, teams with at least one DI were more effective than teams with no DIs. 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. 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. 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. Considering educational background, teams with NCS of 2 were generally most effective. Also CS knowledge is sort of domain knowledge, but it is different from problem domain knowledge. Similar to with NCS, teams with NSE of 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.

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 6.94: Summary of the Initial Observations and Statistical Analysis Results

7
Case Study

This chapter describes a case study [48] carried out to corroborate the results of the first experiment in an industrial setting. The case study involved the idea generation part of a requirement idea brainstorming session conducted in an industrial company to generate ideas for the company's future products. The idea generation was carried out by a team deliberately constructed with four domain experts supplied by the company and with four domain ignorants supplied by the University of Waterloo, including Berry and me.

7.1 Case Studies and the Design of this Case Study

As Dybå, Kitchenham, and Jørgensen have noted [21], laboratory-scaled, controlled experiments are not sufficient to prove a theory about software engineering. Industrial case studies are, therefore, a must in order to complement the laboratory experiments. Indeed, corroborating industrial case studies were among the suggested future work of the paper [47] that reported the results of the first controlled experiment. This chapter describes one particular industrial case study.

This case study was designed and is reported in accordance with many of the recommendations of Runeson and Höst [53].

7.1.1 Research Goals and Questions

Since the purpose of the case study is to corroborate the results of the controlled experiments reported in Chapters 4, 5, and 6, the research goals and research questions governing the case study are those given in Chapter 1.

7.1.2 Context of the Case Study

The study was the idea generation part of a brainstorming session for a company, *C*, to generate requirement ideas for a system situated in a specific domain, *D*, to be developed as one of *C*'s products.

An important provision in the ground rules for the case study was that none of *C*'s intellectual property was to leave the room. The only things that I was to study were the mechanism of the brainstorm and how the mix of DIs and DAs affected it. Therefore, there was no recording of the session, and only people from *C* wrote down any of the ideas. Since I was not a *C* employee, I was simply unable to evaluate the innovativeness of any idea. I could evaluate the newness of an idea only relative to the brainstorming session. In the rest of this chapter, a *new* idea is one that has just appeared in the brainstorming session, and a built-on idea from which a new idea is derived is necessarily an *old* idea. Follow-up questions asked after the session was over were to focus on the mechanism of brainstorming, who generated new and old ideas, and the way new ideas were built on. In particular, no actual generated ideas were to be discussed.

The session had eight participants, four of which came from *C* and four of which came from the University of Waterloo. I assumed by default that the four developers from *C* were DAs with respect to *C*'s domain *D*, and that the four University of Waterloo people, who were not working at *C*, were DIs with respect to *D*. The DAs were members of a team of software developers at *C* who had developed several *C* products in *D*. Thus, the DAs were, in fact, DEs. The DA team included one supervisor and three people working under him. The DIs were I, Berry, and two PhD students in computer science.

The fact that Berry and I actively participated in the brainstorming session of the study makes the case study an example of action research [67, 22]. Rory O'Brien describes *action research*, also known as *participatory research*, as [49] "'learning by doing' — a group of people identify a problem, do something to resolve it, see how successful their efforts were, and if not satisfied, try again." Therefore, it is *expected* that Berry and I did our best in the brainstorming to generate innovative ideas. On the other hand, it was necessary to find ways that my part in any evaluation of the brainstorming session would be as objective as possible. As a matter of fact, the only

evaluation of the *quality* of the session and of the ideas came from the participants from the company.

7.1.3 Basic Limitations of Case Studies and Threats

There is no way in a case study to *prove* anything even closely resembling the hypothesis that drives a controlled experiment. It is possible to get a few teams doing different treatments, but there is no way to factor out the differences between the teams to say with confidence that the difference in results is *caused* by the difference in the treatments. So the best that can be done is to have a team or a few try out the treatment that is being tested and to have the teams report how they believed that they did in comparison to their past efforts with different treatments.

Therefore, the plan for the case study was to get one team to carry out the idea generation part of a requirement idea brainstorm with a mix of DAs and DIs, and then to later ask the DA members of the team to compare what happened in the case study session to what happened in previous sessions they had done with only DAs.

A problem with this and every case study is the large dependence of the results on the context of the study. Therefore, the results of case studies have low external validity and are not generalizable [62]. However, repeated non-generalizable case studies all showing the same conclusion leads to generalizing over time.

Other potential threats to the validity of the results are the researchers' biases both in participating and in doing any evaluation. Since the case study was designed as action research, it is fine for the researchers to participate in the research. However, when a researcher participates, his evaluation of the ideas is likely to be biased. In this study, the quality of the session and of the generated ideas were evaluated by only the DAs from *C*. The only evaluation I did was to determine which generated ideas were built on previously generated ideas.

7.1.4 Data Collection and Evaluation

During the session, I monitored the conversations only in order to analyze some aspects of the discussions. I noted which ideas were generated by whom and whether an idea was a new one or one built on an idea originated by another participant. Another DI was monitoring the session for another research study. Therefore, two DIs were fully focused on generating requirement ideas, while each of the other two had a responsibility besides generating ideas.

In order to analyze the results, I focused on the individuals to see whether and how their domain knowledge affected their participation. I used a first-degree data collection technique,

Participants	*New Ideas* (Discussion initiator)	*Extensions*	*Total*
DAs	12 (3)	15	27
Fully focused DIs	14 (7)	2	16
Partly focused DIs	8 (2)	1	9
Total	34 (12)	18	52

Table 7.1: Summary of the Ideas Generated

a.k.a. "observations in meetings" [53], in which Berry and I participated in the study session and I collected data during the session.

The follow-up questions that I asked of the DAs after the session tried to get the DAs to offer their opinions about the overall quality of the ideas generated during the session, recognizing that they would not comment on any specific idea. The questions tried also to get the DAs to compare the just-completed idea generation session with past idea generation sessions conducted in *C*, using only *C* employees who were DAs.

7.2 Conduct of the Session

At the beginning of the session, the DIs asked the DAs for someone to give a short description of the system to be built, explaining that in a brainstorming session, everyone must be aware of the problem in *D* that is to be solved by the system. One of the DAs, the supervisor, gave a brief overview of the problem. Although none of the DIs knew about the problem beforehand, each had some familiarity with *D* and all had some acquaintance with *C* products. However, none had familiarity with the inner workings of these products, and none had any idea what new products *C* might be planning for the future. Certainly, the DAs were looking for requirements for a system the likes of which they had a lot of experience with. Given that each of the DAs was a *domain expert (DE)*, there was, thus, a significant difference in the domain knowledge of the DAs and the DIs. Thus, the default classification of the participants was, in retrospect, correct.

7.3 Results

As shown in Table 7.1, 52 ideas were generated during the session; 27 were generated by DAs and 25 were generated by DIs. Sixteen ideas were generated by DIs who were fully focused on generating ideas, whereas nine ideas were generated by DIs who were analyzing the ideas during the session. This difference is not surprising, considering the difference in focus.

Besides counting the raw number of ideas, I tracked the initiator of each idea. I found that nine of the ideas that were initiated by DIs led to extension ideas, and that three of the ideas that were initiated by DAs led to extension ideas. So, if we consider each idea and its extensions as a single idea, then there were 34 new ideas, 22 of which were generated by DIs. Nine of these 22 led to 18 extended ideas.

7.4 Observations

I noticed that the DAs were not as active in the beginning of the session as the DIs. They gradually became active after DIs threw out a few ideas. Another interesting phenomenon was that many ideas were initiated by a DI bringing up something that might or might not have been outside *D*. It seemed that in many such cases, a DA built on these ideas.

7.5 DAs' Feedback and Evaluation

After the session was over, I sent some questions by e-mail to the supervisor DA. I got answers from the supervisor, hereinafter called "*A*" and one other DA, hereinafter called "*B*". I asked follow-up questions based on their responses. Appendix B shows sanitized versions of the original questions and the follow up questions. This subsection summarizes what I learned from the answers to these questions.

Each responding DA seemed to be generally satisfied with the session and the ideas generated. They indicated that they were eager to elaborate on the requirement ideas generated during the session. That they so indicated is a good sign. Normally, such elaboration is done in the second part of brainstorming, which did not happen in the case study. That they wanted to elaborate on their own suggests that they found some of the generated ideas worth pursuing and that they had not thought of them before.

When I asked the DAs whether they thought that the session went differently from what they do in *C*, *B* replied that it was different since wider areas were explored, but with less focus than

in normal brainstorming sessions at *C*. This statement by a *C* employee is an indication that the DIs may have generated some ideas that were innovative to *C*. One explanation for this observation is that the DAs are more interested in technical details, as they were looking for and considering only ideas that are feasible to implement. The DIs were oblivious to technical details and feasibility.

I asked the DAs whether there were ideas proposed by the DIs that the DAs would not have thought of. *B* replied that there were ideas totally new to them. This statement by a *C* employee is yet another indication that some of the ideas generated by DIs were innovative to *C*. This DA said that some of the ideas were not polished, but some might possibly become useful after some refinement and fleshing out of details. Considering that the main purpose of the idea generation part of brainstorming is to generate as many ideas as possible, it seems that what the DA observed is about normal for idea generation in brainstorming.

When I asked the DAs why they were not active in the beginning, each of the respondents replied that the four of them wanted to hear about the ideas DIs have and to learn the DIs' level of knowledge of *D*, and only then to throw in their own ideas. The DAs said also that they found ideas generated by DIs interesting enough that the DAs felt compelled to build on them.

I asked the DAs whether they had noticed that many of the ideas were initiated by a DI. Each of the respondents said "Yes" and explained that DIs have perspectives that are different from those of DAs. *A* said some of the ideas that DIs brought up were irrelevant to the problem. However, the DAs were often able to make an irrelevant idea worthy of follow up by modifying the idea to fit what they perceive as the domain *D* of *C*.

7.6 Conclusion

The results show very similar numbers of raw ideas generated by the DAs and the DIs. However, DIs were better in generating new ideas. Many of the new ideas were nevertheless considered useful by DAs, and the DAs tried to extend them. Therefore, we can say that DIs helped provide new ideas.

DIs seemed to generate ideas that the DAs were not expecting. The DAs seemed stuck in the rut of their domain's box.

The results support the conclusion that having a team consisting of a mix of DAs and DIs improves the effectiveness of the idea generation part of requirement idea brainstorming.

The results of the case study are in basic agreement with results of the controlled experiments in that,

1. if the statements of the DAs can be taken at face value, the session with a mix of DIs and DAs seemed to generate more innovative ideas than are generated during in-house sessions in which only DAs from *C* participate, and

2. each DI had been educated as a computer scientist, as a software engineer, or as both.

7.7 Future Work

There are a number of research questions that could be the subject of related future work. As asked by an anonymous referee of the paper published on this case study [48], does a DI eventually become a DA? Probably, the answer is "Yes". If so, then a DI on a project has a limited life as a DI. On the assumption that a DI does become a DA, perhaps a team consisting of only DIs might work as well as a team of mixed DIs and DAs. The data from the first controlled experiment indicate that on average, among three-person teams, a DI-only team or a mixed team generates more raw ideas than a DA-team. However, a mixed team generates more high quality ideas than a DI-only team. It would seem, as suggested by the case study results, that DAs are needed to enhance innovative, but useless ideas generated by the DIs into useful and still innovative ideas. Only additional work can answer this question.

To learn the true quality of the ideas that were generated in the reported brainstorming session, it would be useful to ask the *C* participants of the session in about one or two years' time, if any of the ideas generated in the session have led to any actual *C* products.

So far, the work has focused on only one RE activity, requirement idea generation during requirements elicitation. Future case studies could involve other RE activities.

8

Conclusion and Future Work

After some refinement, the main objective of this research was to study the impact of lack of domain knowledge in requirements engineering. This study tested mainly the hypothesis that a team consisting of a mix of DIs and DAs generates more requirement ideas while performing the idea generation part of brainstorming for requirement ideas than does a team consisting of only DAs. Section 8.1 describes the results of the two controlled experiments. Section 8.2 describes the results of the case study. Section 8.3 draws the conclusions from these results.

8.1 Results of the Controlled Experiments

The primary research method used in this study is controlled experiments. Two pilot studies were conducted to learn a good design for the experiment. The first controlled experiment, E1, was conducted afterwards to test the main hypothesis, and several others identified during the pilot studies, using brainstorming for requirement ideas for a BDWP. Each of the participants was a computer science or software engineering student. The results 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.

A second controlled experiment, 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 computer science and software engineering 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 different conclusions?" In terms of types of experimental errors, two possibilities are that:

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 might 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 might 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.

8.2 Results of the Case Study

An industrial case study was conducted to corroborate the results of the controlled experiments. The main outcome of the case study was that the DIs were not only at least as effective as DAs, they were also probably more creative than the DAs, and therefore, brought many out-of-the-box ideas to the tables.

The results of the case study are clearly in line with the results of E1, which weakly support the main hypothesis. Since the results of E1+E2 do not support the main hypothesis, one might say that the results of the case study disagree with the results of E1+E2. However, one possible reason that the results of E1+E2 did not support the main hypothesis is the presence of many people not in computer science or software engineering among the participants of E2. In the case study, *each* participant had a computer science or software engineering background. So, the results of the case study are not inconsistent with the results of E1+E2.

8.3 Conclusion

The conclusions of the controlled experiments and the observations of the case study are used to answer the research questions presented in Section 2.3:

RQ_1 *Does a team consisting of a mix of DIs and DAs performing a GT generate more DGUs than a team consisting of only DAs?*

The one GT that was applied in the experiments was requirement idea generation, and the DGUs were requirement ideas. Due to the inherent difficulties of conducting controlled experiments, it was not practical to apply the experiment to more than one RE activity. The results of the controlled experiments, E1 in Section 5.5 and E1+E2 in Section 6.8, provide at best, weak support for the the hypothesis that the effectiveness of a team in requirement idea generation is affected by the team's mix of domain familiarities. However, there is evidence in Section 6.8 that the effect of the mix of a team was overshadowed by other independent variables, especially the team's educational background. The case study, however, provided some evidence, in Section 7.6, of the positive effect of the team's mix.

Therefore, the results of the experiments and the case study allow me to say that having at least one DI with a computer science or software engineering background on a team that is brainstorming for requirement ideas probably increases the number of raw, relevant, feasible, and innovative requirement ideas generated by the team.

RQ_2 *Do factors other than a team's mix of DIs and DAs impact the number of DGUs generated by the team's performing a GT?*

There is evidence in Section 6.8 that factors other than the mix of a team impact the number of requirement ideas generated by the team, with the most significant one being the team members' educational backgrounds.

The ultimate objective of the research that motivated this thesis is to help RE managers in forming more effective teams for doing 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 any RE activity: Before any experiment, my bet was that in a three-person team, a team with precisely one DI was the most effective in requirement idea generation. However, the E1 and E2 results show that teams with more than one DI were more effective.
- providing some RE activities for which domain ignorance is at least helpful: This thesis showed that domain ignorance is at least helpful in requirement idea generation.
- 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: The experience together with Mehrotra's results [44] suggest that in company *C*, brainstorming teams should be composed of DEs and new employees. Each new employee would serve as a roving DI until he learned too much about *C*'s domain *D* to be a true DI. Moreover, serving as a roving DI would allow him to be immediately useful while leisurely learning *D* in circumstances in which a lack of knowledge of *D* both is useful and cannot cause any damage.

8.4 Future Work

There are many activities other than requirement idea generation that could benefit from domain ignorance. One such requirements engineering activity is requirement ideas inspection.

As mentioned in Chapter 1, one of the 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 [17], data mining [40, 2], and exploratory software testing [36]. 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 [28]. 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.

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.

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.

The findings of the case study are promising to have potential significance in industry. Staffing RE teams can potentially benefit from the results of this study. Indeed, more work is required to strengthen the power of the results by generating more data points via replicating the controlled experiment described in this study. 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 software engineering.

Appendices

A

Experiment Materials

Ethics Clearance Certificate

UNIVERSITY OF WATERLOO

OFFICE OF RESEARCH ETHICS

Notification of Ethics Clearance of Application to Conduct Research with Human Participants

---

Faculty Supervisor: Daniel Berry
Department: Computer Science, School of

Student Investigator: Ali Niknafs
Department: Computer Science, School of

ORE File #:

Project Title: The impact of domain knowledge on the effectiveness of requirements engineering activities.

---

*This certificate provides confirmation that the additional information/revised materials requested for the above project have been reviewed and are considered acceptable in accordance with the University of Waterloo's Guidelines for Research with Human Participants and the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. Thus, the project now has received ethics clearance. This clearance is valid for a period of **four years** from the date shown below and is subject to an **annual ethics review process** (see Note 2). A new application must be submitted for on-going projects continuing beyond four years.*

***Note 1**: This project must be conducted in accordance with the description in the application and revised materials for which ethics clearance has been granted. All subsequent modifications to the application must be submitted for prior ethics review using ORE Form 104 and must not be intiated until notification of ethics clearance has been received.*

***Note 2**: All ongoing research projects must undergo annual ethics review. ORE Form 105 is used for this purpose and must be submitted by the Faculty Investigator/Supervisor (FI/FS) when requested by the ORE. Researchers must submit a Form 105 at the conclusion of the project if it continues for less than a year.*

***Note 3**: FIs and FSs also are reminded that they must immediately report to the ORE (using ORE Form 106) any events related to the procedures used that adversely affected the participants and the steps taken to deal with these.*

6/2/10
Date

Director, Office of Research Ethics

OR

Senior Manager, Research Ethics

OR

Manager, Research Ethics

Likert Checklist

The following checklist contains a list of problem domains which are the potential domains of the experiment task you will perform in the second meeting. Read each domain name and indicate how much knowledge and experience you have of that domain by putting an X in the corresponding box.

		Very Aware	Aware	Neutral	Ignorant	Very Ignorant
Version Control for Software Development						
Configuration Management for Software Development						
Traceability Management for Software Development						
Multilingual Formatting						
Multilingual Text Editing						
Bidirectional Formatting						
Bidirectional Text Editing						
Outline Fonts						
Compilation of Programming Languages						
Operating Systems						
Graphical User Interface Builders						
HTML						
XML						
Browsers						
Search Engines						
E-mail Handling						
E-store						
E-auctioning						
Airline Flight Reservation, Ticketing, Boarding Passes						
Music	General					
	Rock and Roll					
	Jazz					
Dictionary						
Art						
Architecture						
Bridge building						
Highway building						
Photography						
Web Publishing						
University Admissions						
Online Review						
Social Networking						
Classified Ads						

Consent Form

Consent Form

Project Title: A Study of the Effectiveness of Brainstorming

Student Investigator: Ali Niknafs, School of Computer Science, [redacted]

Faculty Advisor: Dr. Daniel Berry, School of Computer Science, [redacted]

Study Purpose:

The goal of this research is to empirically study the effectiveness of brainstorming in requirements elicitation.

The research will be conducted in the context of a small requirements elicitation session. In this study you will participate in a brainstorming activity to generate the requirements for a computer-based system.

Study Procedures:

The study takes place during one session of 1 hour and 30 minutes. During the first hour, you will be asked to sign a consent form, to write the Williams Creativity Test, and then hear a description of a computer-based system (CBS). The goal of the creativity test is to help the investigator to explain and analyze the outcomes of the brainstorming session. The goal of hearing the description of a CBS is to acquaint you with the system you will be brainstorming about.

In the last 30 minutes, you will be put in a team that will participate in a brainstorming session to generate requirements for the CBS.

This study will take approximately 90 minutes of your time.

Risks and Benefits to the Participant:

There are no known or anticipated risks to participation in this study. Your participation in the research will significantly contribute to the body of knowledge in requirements engineering. Your time and effort will help to further mature the field of requirements engineering. At the end of the study, you will be provided with a summary of the findings if you provide your e-mail address on the consent form. You may benefit from being exposed to new techniques in requirements elicitation that you will be able to use effectively in your professional career.

Participant's Right to Withdraw from the Study:

You have the right to refuse to participate or to withdraw at any time, without penalty.

Remuneration to the Participant:

You will be paid an honorarium of $30 for the approximate 90 minutes for participating.

Confidentiality and Privacy:

All data collected will be regarded as confidential, and every effort will be made to ensure that you will not be identified either directly or indirectly in either verbal or written reports.

Once the data are logged, each name will be replaced by a unique number in the data and the relationship between the number and the name will not be recorded.

Then the original data will be shredded. The information will be kept in a locked desk in the principal investigator's office, and in password-protected computer accounts. Paper records will be confidentially shredded after one month. Anonymized electronic data will be kept indefinitely.

Report of results

You may receive a copy of the report describing the results. To receive a copy, please fill in your e-mail address below your signature on the consent form.

Contact information and Ethics Clearance for the study

If you have any questions regarding this study, or would like additional information to assist you in reaching a decision about participation, please contact me at [redacted] or by e-mail at [redacted] or by contacting Dr. Berry at [redacted].

I would like to assure you that this study has been reviewed and received ethics clearance through the Office of Research Ethics at the University of Waterloo. However, the final decision about participation is yours. If you have any comments or concerns resulting from your participation in this study, please contact [redacted] of this office at [redacted] or [redacted].

Consent of Participant:

I have read the information presented in the information letter about a study being conducted by Ali Niknafs and Dr. Daniel Berry of the School of Computer Science at the University of Waterloo. I have had the opportunity to ask any questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted. I am aware that I may withdraw from the study without penalty at any time by advising the researchers of this decision.

This project has been reviewed by, and received ethics clearance through, the Office of Research Ethics at the University of Waterloo. I was informed that if I have any comments or concerns resulting from my participation in this study, I may contact the Director, Office of Research Ethics at [redacted] or [redacted].

With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

Participant's Print Name: _____

Participant's Signature: _____

Dated at Waterloo, Ontario: _____

Witness's Signature: _____

Dated at Waterloo Ontario: _____

E-mail address to be sent a copy of results report: _____

Recruitment Email

Hello,

My name is Ali Niknafs and I am a PhD student working under the supervision of Dr. Daniel Berry <e-mail: [redacted] > in the School of Computer Science at the University of Waterloo.

I am currently working on a study to investigate the effectiveness of brainstorming. The study takes place during one session of 1 hour and 30 minutes.
During the first hour, you will be asked to sign a consent form, to write a short creativity test, and then hear a description of a computer-based system (CBS).

The goal of the creativity test is to help the investigator to explain and analyze the outcomes of the brainstorming session. The goal of hearing the description of a CBS is to acquaint you with the system you will be brainstorming about.

In the last 30 minutes, you will be put in a team that will participate in a brainstorming session to generate requirements for the CBS.

You will not need any special experience in software engineering to be able to participate.

This study will take approximately *90 minutes* of your time, and in appreciation of your time commitment, you will receive an honorarium of $30.

We are looking for people who speak right-to-left languages (e.g., Arabic, Hebrew, Persian, and Urdu), but we want others as well. We need to have a mix of people who speak right-to-left languages and those who do not. Therefore, please accept our apologies if after indicating that you want to participate but before even signing the consent form, you are not asked to participate in the study.

If you are interested in participating, please fill out the registration. I will then send an e-mail confirming that you have been signed up and provide you with further information concerning the location of the study. You are free to withdraw from the study at any time by sending me e-mail.

This study has been reviewed and received ethics clearance through the Office of Research Ethics at the University of Waterloo.
Sincerely,

Ali Niknafs
David R. Cheriton School of Computer Science
University of Waterloo
[redacted]

William's Test

EXERCISE IN DIVERGENT FEELING

BY

DR. FRANK WILLIAMS

ID _____

DATE _____

Instructions:

This is an exercise which will help you find out how creative you feel about yourself. Among the following short sentences, you will find some that definitely fit you better than others. These should be marked with an X in the "MOSTLY TRUE ABOUT ME" column. Some sentences may fit you only part of the time and should be marked with an X in the "PARTLY TRUE OR UNTRUE ABOUT ME" column. Other sentences will not fit you at all and should be marked with X in the "MOSTLY UNTRUE (FALSE) ABOUT ME" column. Those sentences that cannot make a decision about should be marked with an X in the "CANNOT REALLY DECIDE" column. Mark every sentence and do not think a long time about the sentence. *There are no right or wrong answers*. Mark your first feeling as you read each sentence. This is not timed exercise, but work as quickly as you can. Remember to try and answer each sentence by the way you really feel about yourself. Place an X in the column which you feel is the most nearly like you.

	Mostly True About Me (yes)	Partly True or Untrue About Me (maybe)	Mostly Untrue (False) About Me (no)	Cannot Really Decide (don't know)
1. In college or at work, I try to guess about things even if I'm not right.				
2. I like to look at things more closely to find details I haven't seen before.				
3. I usually ask questions when I don't know something.				
4. I never like to have a plan for doing things.				
5. I want to know that I am right before I try a new proposal.				
6. I like to dream about things I want to know or do.				
7. If I can't do something the first time, I keep trying until I can do it.				
8. I never choose to undertake an experience that no one else has done before.				
9. I like doing things the same way rather than finding new ways.				
10. I like to find out if things are really true.				
11. I like doing many new things.				
12. I like making new acquaintances and new friends.				
13. I like to think about things that have never happened to me.				

		Mostly True About Me (yes)	Partly True or Untrue About Me (maybe)	Mostly Untrue (False) About Me (no)	Cannot Really Decide (don't know)
14.	I do not like to dream about someday being a good artist, musician, or poet.				
15.	Some of my ideas are so exciting to me that I forget other things.				
16.	I would rather live and work on a space station than live here on earth.				
17.	I get nervous when I don't know what's going to happen next.				
18.	I like things that are different.				
19.	I often wonder what other people are thinking.				
20.	I like stories or TV about things that have already happened in the past.				
21.	I feel OK about joining some friends and sharing my ideas.				
22.	I usually keep quiet when things go wrong or when I make a mistake.				
23.	I would like to do or make something no one else has thought of before.				
24.	I like friends who do everything the same old way all the time.				
25.	I usually don't like most rules.				
26.	I like to solve a problem even if it has no right answer.				

		Mostly True About Me (yes)	Partly True or Untrue About Me (maybe)	Mostly Untrue (False) About Me (no)	Cannot Really Decide (don't know)
27.	There are many things I would like to experiment with.				
28.	Once I find an answer to a problem, I like to stick to it instead of trying to find other answers.				
29.	I don't like to talk in public, at assemblies, meetings, etc.				
30.	When I read or watch TV, I like to pretend I am someone in the story.				
31.	I like to dream about how people lived 200 years ago.				
32.	I don't like it when my friends cannot make a decision.				
33.	I like to explore old trunks and boxes just to see what might be in them.				
34.	I would like that in college or at work, the people around me keep doing things the same way instead of changing.				
35.	I know the way I feel about things is right.				
36.	It's fun to guess and see if I'm right.				
37.	It's fun to create situations that make me wonder about what's going to happen next.				
38.	I am interested in machines, and wonder what they look like inside and how they run.				

		Mostly True About Me (yes)	Partly True or Untrue About Me (maybe)	Mostly Untrue (False) About Me (no)	Cannot Really Decide (don't know)
39.	My best friends dislike having silly ideas.				
40.	I like to think about new ideas even if I can't use them.				
41.	I like to keep everything in a special place.				
42.	It would be exciting to try and find answers for problems in the future.				
43.	I like to try new things just to see what will happen.				
44.	I am usually more interested in doing things I enjoy than doing just things I do well.				
45.	I like to think about exciting things that no one has ever thought of.				
46.	When I see a picture of someone I don't know, I like to wonder what that person is like.				
47.	I like to look at books and magazines just to see what is in them.				
48.	I think there is just one right answer to most questions.				
49.	I like to ask questions about things other people don't think about.				
50.	I really like having lots of interesting things to do at home, in college, or at work.				

Feedback Letter

FEEDBACK LETTER

Project Title:	The impact of domain knowledge on the effectiveness of requirements engineering activities.
Student Investigator:	Ali Niknafs, School of Computer Science, [redacted]
Faculty Advisor:	Dr. Daniel Berry, School of Computer Science, [redacted]

Dear *(Name of Participant),*

We appreciate your participation in our study, and thank you for spending the time helping us with our research!

As a reminder, the goal of this research was to empirically investigate the impact of domain knowledge on the effectiveness of requirements engineering activities. In this study you participated in a requirements engineering activity to generate the requirements for a computer-based system. The study examined whether the prior experience of the problem domain affects the effectiveness of the requirements engineering activities. It is expected that participants with no prior experience with the problem domain would not only discover a good number of requirements, they would be more effective than those who are expert in the domain.

Based on the information you provided in the first session, you were assigned to one the two groups: domain-aware or domain-ignorant. Those who were assigned to the domain-aware group attended a tutorial session right before the task session. This tutorial was supposed to sharpen the difference of the two groups and to bring the knowledge of the members of domain-aware group to a common level. In the next stage three teams were formed: 1) a team of subjects all drawn from the domain-aware group, 2) a team of subjects all drawn from the domain-ignorant group, and 3) a team consisted of a blend of subjects from both groups.

Team assignments were not random in this study. The results of the Williams Creativity test were used to keep the balance from creativity point of view between the teams.

All information you provided is considered completely confidential; indeed, your name will not be included or in any other way associated, with the data collected in the study.

Once all the data are collected and analyzed for this research, I plan on sharing the analyzed results with the research community through seminars, conferences, presentations, and journals. If you are interested in receiving more information regarding the results of this study, or if you have any questions or concerns, please contact me via the e-mail address listed at the bottom of the page. If you would like a summary of the results and you did not request it when signing the consent form, please let me know now by providing me with your e-mail address. When the study is completed, I will send it to you.

As with all University of Waterloo projects involving human participants, this project was reviewed by, and received ethics clearance through, the Office of Research Ethics at the University of Waterloo. In the event you have any comments or concerns resulting from your participation in this study, please contact [redacted] at [redacted].

We really appreciate your participation, and hope that this has been an interesting experience for you.

Sincerely,

Ali Niknafs
David R. Cheriton School of Computer Science
University of Waterloo
[redacted]

B

Feedback from the Case Study

Here is the list of questions I asked from the experts participated in the study and their responses. *B* is one of the developers and *A* is the supervisor.

B (a developer)'s Answers:

1. What were your impressions of the brainstorming session you had at the University of Waterloo on Wednesday 4 Oct?

   [*B*] I thought it was pretty good, we had some interesting ideas. That being said, I would've liked some more solid focus on actual APIs/implementation details required for how we could handle some of the ideas.

2. Did it go better, the same, or worse than sessions you have had at C with only C people?

   [*B*] I'd say it didn't go any better or worse, but not the same as the C only brainstorming. There was a wider array of ideas (like for example, Dan mentioning that ...), but not as much depth to the ideas (like how it would be implemented, APIs required, etc)

3. Did the outsiders contribute anything to the session that you believe you would not have gotten had there been ONLY C people?

   [*B*] Yes, there were some things mentioned by the outsiders which were completely outside of the box. Taking some of these ideas we could run with them a bit.

4. Did the outsiders contribute anything of value to the session?

   [*B*] They helped - their contributions required some refinement but it was interesting to hear different perspectives on what people used their phone for and what they thought it could do.

5. I noticed that you (you experts) were not as active in the beginning as the outsiders.

   - What caused you not to participate?

     [*B*] I wanted to hear what the outsiders had to say first; I didn't want to poison their opinion/direction of conversation with my thoughts.

   - Were you inhibited? If so, by what?

     [*B*] I wasn't sure what level of knowledge/background the outsiders had. I didn't want to poison their opinion/direction of conversation with my thoughts.

   - What prompted you to begin to participate when you did?

     [*B*] Eventually got a general sense of knowledge/background the outsiders had.

6. Many ideas seemed to be initiated by an outsider bringing up something that might have been outside the domain (although probably the outsider thought they were relevant). It seemed that in many such cases, an insider built on these ideas.

   - Did you observe this phenomenon? If so, what were and are your impressions of the phenomenon?

     [*B*] Yes I observed this as well. I think it makes sense - often being an "insider", you have a strict way of looking at the subject matter. By bringing in an outsider, they no longer have the same restricted view of the subject matter and bring some things which the insider may not have observed before.

   - Did you participate in such a phenomenon? If so, what were your thoughts as you did so? [*B*] Yes, I did a couple times; though no notable examples stand out. That being said, I both tried to build upon the idea using language the outside would understand, as well as discuss possible implementation details with the insiders.

A (the supervisor)'s Answers:

1. I noticed that you (you experts) were not as active in the beginning as the outsiders.

- What caused you not to participate?

  [*A*] I did not want to influence the responses of the others. I wanted to see what ideas they had first and then I would try to offer ideas.

- Were you inhibited? If so, by what?

  [*A*] See above.

- What prompted you to begin to participate when you did?

  [*A*] The ideas generated sounded interesting and I wanted to push and prod to see what else there was.

2. Many ideas seemed to be initiated by an outsider bringing up something that might have been outside the domain (although probably the outsider thought they were relevant). It seemed that in many such cases, an insider built on these ideas.

   - Did you observe this phenomenon? If so, what were and are your impressions of the phenomenon?

     [*A*] I'd call this the spring board effect. An outsider brought up an idea that may have lied outside the domain and an expert would be able to link the idea to the domain a little bit more. Or even expand on the idea to bring more relevance to the topic.

   - Did you participate in such a phenomenon? If so, what were your thoughts as you did so?

     [*A*] If I recall, yes. There were quite a few instances of this where an idea was brought up, I didn't think it was immediately relevant so I prodded with some additional information to try to make the idea more relevant to the context.

References

[1] Amer Al-Rawas and Steve Easterbrook. Communication problems in requirements engineering: A field study. In *Proc. of Conf. on Prof. on Awareness in Software Engineering*, pages 47–60, 1996.

[2] Sarabjot S. Anand, David A. Bell, and John G. Hughes. The role of domain knowledge in data mining. In *Proceedings of the Fourth International Conference on Information and Knowledge Management*, CIKM '95, pages 37–43, New York, NY, USA, 1995. ACM. Available from: http://doi.acm.org/10.1145/221270.221321, doi:10.1145/221270.221321.

[3] Evan P. Apfelbaum, Katherine W. Phillips, and Jennifer A. Richeson. Rethinking the baseline in diversity research: Should we be explaining the effects of homogeneity? *Perspectives on Psychological Science*, 9(3):235–244, 2014. Available from: http://pps.sagepub.com/content/9/3/235.abstract, doi:10.1177/1745691614527466.

[4] Miles Barker. BA careers — domain knowledge versus analysis skills, 2010. Available from: http://www.modernanalyst.com/Community/CommunityBlog/tabid/182/articleType/ArticleView/articleId/1385/BA-Careers-Domain-knowledge-versus-analysis-skills.aspx.

[5] Victor R. Basili, Gianluigi Caldiera, and Dieter H. Rombach. *The Goal Question Metric Approach*, volume I. John Wiley & Sons, 1994.

[6] Brian Berenbach, Daniel J. Paulish, Juergen Kazmeier, and Arnold Rudorfer. *Software & Systems Requirements Engineering: In Practice*. Software & Systems Requirements Engineering: In Practice. McGraw-Hill, 2009. Available from: http://books.google.ca/books?id=Mjn53zPoDQIC.

[7] Daniel Berry, Khuzaima Daudjee, Jing Dong, Igor Fainchtein, Maria Augusta Nelson, Torsten Nelson, and Lihua Ou. User's manual as a requirements specification: Case studies. *Requirements Engineering Journal*, 9(1):67–82, 2004.

[8] Daniel M. Berry. The importance of ignorance in requirements engineering. *J. Syst. Softw.*, 28(2):179–184, February 1995. Available from: http://dx.doi.org/10.1016/0164-1212(94)00054-Q, doi:10.1016/0164-1212(94)00054-Q.

[9] Daniel M. Berry. The importance of ignorance in requirements engineering: An earlier sighting and a revisitation. *Journal of Systems and Software*, 60(1):83–85, 2002.

[10] Gunnar Blom. *Statistical Estimates and Transformed Beta-Variables*. Wiley, 1958. Available from: http://books.google.ca/books?id=rWoGAQAAIAAJ.

[11] Barry W. Boehm. *Software Engineering Economics*. Prentice Hall, 1 edition, November 1981. Available from: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0138221227.

[12] Frederick P. Brooks. *The Mythical Man-Month: Essays on Software Engineering, 20th Anniversary Edition*. Addison-Wesley Professional, August 1995. Available from: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20{&}path=ASIN/0201835959.

[13] Jeffrey C. Carver, Nachiappan Nagappan, and Alan Page. The impact of educational background on the effectiveness of requirements inspections: An empirical study. *Software Engineering, IEEE Transactions on*, 34(6):800–812, nov.-dec. 2008. doi:10.1109/TSE.2008.49.

[14] Thomas D. Cook and Donald T. Campbell. *Quasi-experimentation: design & analysis issues for field settings*. Rand McNally College Pub. Co., 1979. Available from: http://books.google.ca/books?id=6WxdAAAAIAAJ.

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

[16] Barthelemy Dagenais, Harold Ossher, Rachel K. E. Bellamy, Martin P. Robillard, and Jacqueline P. de Vries. A qualitative study on project landscapes. In *Proceedings of the 2009 ICSE Workshop on Cooperative and Human Aspects on Software Engineering*, CHASE '09, pages 32–35, Washington, DC, USA, 2009. IEEE Computer Society. Available from: http://dx.doi.org/10.1109/CHASE.2009.5071407, doi:10.1109/CHASE.2009.5071407.

[17] Daniela Damian, Remko Helms, Irwin Kwan, Sabrina Marczak, and Benjamin Koelewijn. The role of domain knowledge and cross-functional communication in socio-technical coordination. In *Proceedings of the 2013 International Conference on Software Engineering*, ICSE '13, pages 442–451, Piscataway, NJ, USA, 2013. IEEE Press. Available from: http://dl.acm.org/citation.cfm?id=2486788.2486847.

[18] Oscar Dieste, Natalia Juristo, and Forrest Shull. Understanding the customer: What do we know about requirements elicitation? *IEEE Softw.*, 25(2):11–13, mar 2008. Available from: http://dx.doi.org/10.1109/MS.2008.53, doi:10.1109/MS.2008.53.

[19] Kevin Dunbar. Scientists build models: Invivo science as a window on the science mind. In *Model-Based Reasoning in Scientific Discovery*, pages 85–99. Kluwer Academic/Plenum Publishers, 1999. Available from: http://www.utsc.utoronto.ca/~dunbarlab/pubpdfs/KDMBR99.pdf.

[20] Tore Dybå, Vigdis By Kampenes, and Dag I. K. Sjøberg. A systematic review of statistical power in software engineering experiments. *Information & Software Technology*, 48(8):745–755, 2006.

[21] Tore Dybå, Barbara A Kitchenham, and Magne Jørgensen. Evidence-based software engineering for practitioners. *IEEE Software*, 22(1):58–65, 2005.

[22] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. Selecting empirical methods for software engineering research. In Forrest Shull, Janice Singer, and DagI.K. Sjøberg, editors, *Guide to Advanced Empirical Software Engineering*, pages 285–311. Springer London, 2008. doi:10.1007/978-1-84800-044-5_11.

[23] Robert Feldt and Ana Magazinius. Validity threats in empirical software engineering research - an initial survey. In *SEKE*, pages 374–379, 2010.

[24] Remo Ferrari and Nazim H. Madhavji. 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 '07, page 16, Washington, DC, USA, 2007. IEEE Computer Society. Available from: http://dx.doi.org/10.1109/WICSA.2007.42, doi:10.1109/WICSA.2007.42.

[25] Anthony Finkelstein. Requirements engineering: a review and research agenda. In *Software Engineering Conference, 1994. Proceedings., 1994 First Asia-Pacific*, pages 10–19, Dec 1994. doi:10.1109/APSEC.1994.465278.

[26] Stuart Firestein. Ignorance (Course), August 2013. Available from: http://bioweb.biology.columbia.edu/firestein/?page_id=36.

[27] Gerhard Fischer. Symmetry of igorance, social creativity, and meta-design. In *Proceedings of the 3rd conference on Creativity & cognition*, C&C '99, pages 116–123, New York, NY, USA, 1999. ACM. Available from: http://doi.acm.org/10.1145/317561.317582, doi:10.1145/317561.317582.

[28] Carl Frappaolo. Implicit knowledge. *Knowledge Management Research and Practice*, 6(1):23–25, 2008.

[29] Gene V. Glass, Percy D. Peckham, and James R. Sanders. *Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance*. 1972. Available from: http://books.google.ca/books?id=f7BKtwAACAAJ.

[30] Karen Grace-Martin. Outliers: To drop or not to drop, 2013. Available from: http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/.

[31] Per Pascal Grube and Klaus Schmid. Selecting creativity techniques for innovative requirements engineering. In *Multimedia and Enjoyable Requirements Engineering - Beyond Mere Descriptions and with More Fun and Games, 2008. MERE '08. Third International Workshop on*, pages 32–36, sept. 2008. doi:10.1109/MERE.2008.6.

[32] Irit Hadar, Pnina Soffer, and Keren Kenzi. The role of domain knowledge in requirements elicitation via interviews: an exploratory study. *Requirements Engineering*, pages 1–17, September 2012. Available from: http://dx.doi.org/10.1007/s00766-012-0163-2, doi:10.1007/s00766-012-0163-2.

[33] Nadine Hanebutte, Carol S. Taylor, and Reiner R. Dumke. Techniques of successful application of factor analysis in software measurement. *Empirical Softw. Engg.*, 8(1):43–57, March 2003. Available from: http://dx.doi.org/10.1023/A:1021716832057, doi:10.1023/A:1021716832057.

[34] Perry R. Hinton, Isabella McMurray, and Charlotte Brownlow. *SPSS Explained*. Routledge, 2004. Available from: http://books.google.ca/books?id=DlAUuEbCdDIC.

[35] Martin Höst, Björn Regnell, and Claes Wohlin. Using students as subjects - a comparative study of students and professionals in lead-time impact assessment. *Empirical Softw. Engg.*, 5(3):201–214, November 2000. Available from: http://dx.doi.org/10.1023/A:1026586415054, doi:10.1023/A:1026586415054.

[36] Juha Itkonen, Mika V. Mantyla, and Casper Lassenius. The role of the tester's knowledge in exploratory software testing. *IEEE Transactions on Software Engineering*, 39(5):707–724, 2013. doi:http://doi.ieeecomputersociety.org/10.1109/TSE.2012.55.

[37] Matthias Jarke, Janis A. Bubenko Jr., Colette Rolland, Alistair G. Sutcliffe, and Yannis Vassiliou. Theories underlying requirements engineering: an overview of NATURE at Genesis. In *Requirements Engineering, 1993., Proceedings of IEEE International Symposium on*, pages 19–31, Jan 1993. doi:10.1109/ISRE.1993.324840.

[38] Robert B. Johnson and Larry B. Christensen. *Educational Research: Quantitative, Qualitative, and Mixed Approaches*. SAGE Publications, 2010. Available from: http://books.google.ca/books?id=sULKqmD6MkMC.

[39] Keren Kenzi, Pnina Soffer, and Irit Hadar. The role of domain knowledge in requirements elicitation: an exploratory study. In *MCIS*, 2010.

[40] Ioannis Kopanas, Nikolaos M. Avouris, and Sophia Daskalaki. The role of domain knowledge in a large scale data mining project. In Constantine D. Spyropoulos Ioannis P. Vlahavas, editor, *Methods and Applications of Artificial Intelligence*, volume 2308 of *Lecture Notes in Computer Science*, pages 288–299. Springer Berlin Heidelberg, 2002. Available from: http://dx.doi.org/10.1007/3-540-46014-4_26, doi:10.1007/3-540-46014-4_26.

[41] Per Kristensson, Anders Gustafsson, and Trevor Archer. Harnessing the creative potential among users. *Journal of Product Innovation Management*, 21(1):4–14, 2004. Available from: http://dx.doi.org/10.1111/j.0737-6782.2004.00050.x, doi:10.1111/j.0737-6782.2004.00050.x.

[42] Jonah Lehrer. Accept defeat: The neuroscience of screwing up, 2009. [Online; accessed 6-May-2014]. Available from: http://www.wired.com/2009/12/fail_accept_defeat.

[43] Jonah Lehrer. *Imagine: How Creativity Works*. Penguin Group (Canada), 2012. Available from: http://books.google.ca/books?id=orWo9clzaUQC.

[44] Gaurav Mehrotra. Role of domain ignorance in software development. Master's thesis, University of Waterloo, Waterloo, 2011. Available from: https://uwspace.uwaterloo.ca/handle/10012/5944.

[45] George A. Milliken and Dallas E. Johnson. *Analysis of Messy Data: Designed Experiments, Second Edition*. Number v. 1. Taylor & Francis, 2009. Available from: http://books.google.ca/books?id=loSQzOlY9LkC.

[46] Peter Naur and Brian Randell. *Software Engineering: Report of a conference sponsored by the NATO Science Committee, Garmisch, Germany, 7-11 Oct. 1968*. Scientific Affairs Division, NATO, Brüssel, 1969.

[47] Ali Niknafs and Daniel M. Berry. The impact of domain knowledge on the effectiveness of requirements idea generation during requirements elicitation. In *20th IEEE International Requirements Engineering Conference (RE)*, pages 181–190, Chicago, IL, USA, 2012 ©2012 IEEE. IEEE. Available from: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6345802.

[48] Ali Niknafs and Daniel M. Berry. An industrial case study of the impact of domain ignorance on the effectiveness of requirements idea generation during requirements elicitation. In *21st IEEE International Requirements Engineering Conference (RE)*, pages 279–283, Rio de Janeiro, Brazil, 2013 ©2013 IEEE. IEEE. Available from: http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=6636730.

[49] Rory O'Brien. Um exame da abordagem metodológica da pesquisa ação [An Overview of the Methodological Approach of Action Research]. *Teoria e Prática da Pesquisa Ação [Theory and Practice of Action Research]*, 2001. Available from: Englishversionavailable:http://www.web.ca/~robrien/papers/arfinal.html.

[50] Blaise Pascal and A. J. Krailsheimer. *Pensees: Translated with an Introduction by A.J. Krailsheimer*. Penguin classics. Penguin, 1968. Available from: http://books.google.ca/books?id=FEeKNAEACAAJ.

[51] Preethu Rose, Manish Kumar, Nirav Ajmeri, Manas Agrawal, V. Sivakumar, and Smita Ghaisas. A method and framework for domain knowledge assisted requirements evolution (K-RE). In *International Conference on Software Engineering, CONSEG 09*, pages 87–97, 2009.

[52] Jason Rubens. Business analysis and requirements engineering: the same, only different? *Requir. Eng.*, 12(2):121–123, May 2007. Available from: http://dx.doi.org/10.1007/s00766-007-0043-3, doi:10.1007/s00766-007-0043-3.

[53] Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical Softw. Engg.*, 14(2):131–164, April 2009. Available from: http://dx.doi.org/10.1007/s10664-008-9102-8, doi:10.1007/s10664-008-9102-8.

[54] Seung Sahn. Letters to a beginner, 2012. Available from: http://www.kwanumzen.org/about-zen/three-letters-to-a-beginner/.

[55] Nancy Van Schooenderwoert. Embedded agile project by the numbers with newbies. In *Agile Conference, 2006*, pages 351–366, July 2006. doi:10.1109/AGILE.2006.24.

[56] Helen Sharp. The role of domain knowledge in software design. *Behaviour and Information Technology*, 10(5):383–401, 1991. Available from: http://www.tandfonline.com/doi/abs/10.1080/01449299108924298, arXiv:http://www.tandfonline.com/doi/pdf/10.1080/01449299108924298, doi:10.1080/01449299108924298.

[57] Robert I. Sutton. *Weird ideas that work: 11 1/2 practices for promoting, managing, and sustaining innovation*. Free Press, 2002. Available from: http://books.google.ca/books?id=6UaSc8m8nSoC.

[58] Calvin W. Taylor and Frank E. Williams. *Instructional Media and Creativity: The Proceedings of the Sixth Utah Creativity Research Conference*. Distributed by ERIC Clearinghouse, [Washington, D.C.] :, 1965. Available from: http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED010651.

[59] BBN Technologies. Prophet 5: Statistical and sequence analysis software, August 2013. Available from: http://www.basic.northwestern.edu/biotools/prophet.html.

[60] Paul Thagard. Collaborative knowledge. Number 31, pages 242–261. Blackwell Publishers Inc., 1997. doi:10.1111/0029-4624.00044.

[61] Psychology Department University of Utah. Psystats, August 2013. Available from: http://psystats.wikispaces.com/.

[62] June M. Verner, Jennifer Sampson, Vladimir Tosic, Nur Azzah Abu Bakar, and Barbara Kitchenham. Guidelines for industrially-based multiple case studies in software engineering. In André Flory and Martine Collard, editors, *RCIS*, pages 313–324. IEEE, 2009. Available from: http://dblp.uni-trier.de/db/conf/rcis/rcis2009.html#VernerSTBK09.

[63] Rebecca M. Warner. *Applied Statistics: From Bivariate Through Multivariate Techniques: From Bivariate Through Multivariate Techniques*. SAGE Publications, 2012. Available from: http://books.google.ca/books?id=b1bXhepuJOEC.

[64] Wikipedia. Kurtosis — Wikipedia, the free encyclopedia, 2013. [Online; accessed 22-Aug-2013]. Available from: http://en.wikipedia.org/wiki/Kurtosis.

[65] Wikipedia. Skewness — Wikipedia, the free encyclopedia, 2013. [Online; accessed 22-Aug-2013]. Available from: http://en.wikipedia.org/wiki/Skewness.

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

[67] Wikipedia. Action research— Wikipedia, the free encyclopedia, 2014. [Online; accessed 2-May-2014]. Available from: http://en.wikipedia.org/wiki/Action_research.

[68] Wikipedia. Analysis of variance — Wikipedia, the free encyclopedia, 2014. [Online; accessed 7-Feb-2014]. Available from: http://en.wikipedia.org/wiki/Analysis_of_variance.

[69] Wikipedia. Internal validity — Wikipedia, the free encyclopedia, 2014. [Online; accessed 27-Jan-2014]. Available from: http://en.wikipedia.org/wiki/Internal_validity.

[70] Wikipedia. Lean software development — Wikipedia, the free encyclopedia, 2014. [Online; accessed 27-Jan-2014]. Available from: http://en.wikipedia.org/wiki/Lean_software_development.

[71] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in software engineering: an introduction*. Kluwer Academic Publishers, Norwell, MA, USA, 2000.