Jan 2022

2IMP40 2IMP40 Applications of Data Science to Software Engineering 21/22 · 3 exercises · 30.0 points

1 Part 1

20.0 points · 20 questions

The first part of the exam consists of **twenty** multiple choice questions. Each question has four answer options and exactly one is correct.

Text

Read the following fragment from Ralph (2018): "The problem is known and the goal of the system is clear. Analysts elicit comprehensive, unambiguous requirements which are agreed by the client. Designers search a conceptual solution space for design candidates that satisfy the requirements. They use logic and reason to deduce an appropriate architecture or user interface. Design decisions are concentrated in this phase of the project. Developers select an appropriate software development method and use it to build the system. Although perfect rationality is impossible, developers strive to be as structured, methodical and rational as they can. They plan development as a series of activities or phases with milestones and execute this plan. Unexpected events trigger re-planning. Teams understand and evaluate their progress in terms of the plan. The project is successful if it delivers the agreed scope within the allotted time and budget, at a reasonable quality. Researchers understand this process in terms of lifecycle models and software development methods."

Text

a The quote from Ralph (2018) above corresponds to

1.0 point · Multiple choice · 4 alternatives

●	rationalism	1.0
○	empiricism	0.0
○	both rationalism and empiricism	0.0
○	neither rationalism nor empiricism	0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

b Which of the following research questions **cannot** be answered empirically?

1.0 point · Multiple choice · 4 alternatives

○ What are the most common GitHub badges? 0.0

Feedback

Answering this question requires observing how badges are used on GitHub. This question can (and should) be answered empirically.

○ What skills do students think are best to develop in an online educational hackathon? 0.0

Feedback

To answer this question one has to observe students' opinions. This question can (and should) be answered empirically.

● **How can we model the domain of blazonry as a domain-specific language, using software language engineering methods?** 1.0

Feedback

This is a design question

○ What are the top topics that have been studied in Software Engineering? 0.0

Feedback

This is an empirical question about the topics that have been studied; it can be answered by analysing scientific publications or interviewing/surveying researchers.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

c Jiang et al. (2017) investigate why and how developers fork what from whom, using a collection of GitHub repositories involving over 236,000 developers and over 1.8m forks. This study can be classified as

1.0 point · Multiple choice · 4 alternatives

○ A field experiment 0.0

Feedback

No, no interventions

○ A field study 0.0

Feedback

No,

○ A judgement study 0.0

● A sample study 1.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

d Niknafs and Berry (2017) investigate the impact of domain knowledge on requirements engineering activities. Hypotheses are set up and evaluated through two controlled experiments with students. This study can be classified as

1.0 point · Multiple choice · 4 alternatives

○ An experimental simulation	0.0
○ A field experiment	0.0
○ A formal theory	0.0
● A laboratory experiment	1.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

e Consider the following abstract from Danilova et al.:

Recruiting professional programmers in sufficient numbers for research studies can be challenging because they often cannot spare the time, or due to their geographical distribution and potentially the cost involved. Online platforms such as Clickworker or Qualtrics do provide options to recruit participants with programming skill; however, misunderstandings and fraud can be an issue. This can result in participants without programming skill taking part in studies and surveys. If these participants are not detected, they can cause detrimental noise in the survey data. In this paper, we develop screener questions that are easy and quick to answer for people with programming skill but difficult to answer correctly for those without. In order to evaluate our questionnaire for efficacy and efficiency, we recruited several batches of participants with and without programming skill and tested the questions. In our batch 42% of Clickworkers stating that they have programming skill did not meet our criteria and we would recommend filtering these from studies. We also evaluated the questions in an adversarial setting. We conclude with a set of recommended questions which researchers can use to recruit participants with programming skill from online platforms.

What research strategy was used by the authors?

1.0 point · Multiple choice · 4 alternatives

○ Judgement Study	0.0
● **Laboratory Experiment**	1.0
○ Field Experiment	0.0
○ Sample Study	0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

f Consider the following excerpt from the paper by Papoutsoglou et al.:

> *On DEV, users typically tag their articles to improve the article’s visibility and attract readers. Users browsing DEV can search for articles by tag. We wrote and deployed a customized web crawler that searches for all tags on the platform and collects all articles using that tag. We stored metadata and author information for every article. For each unique user we crawled the information available on their public profile, including links to other platforms. Using the official DEV API (https://docs.forem.com/api/), we identified more than 53,000 posting users, which we used to verify the completeness of our collection of articles. Roughly 33% of users wrote 80% of articles in our corpus. As an article can have more than one tag, we removed duplicate articles based on their URL and title. We extracted the full text of 147,030 articles. Using a language detection method we filtered out non-English articles leaving a corpus of 138,925 articles.*

What sampling method was used by the authors?

1.0 point · Multiple choice · 4 alternatives

○	Random sampling	0.0
○	Quota sampling	0.0
○	Whole frame sampling	0.0
●	**Purposive sampling**	1.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

g Consider the following quote from a paper by Devanbu et al.:

Our target audience were people primarily in a software engineering discipline at Microsoft: this included developers, testers, program managers, and their immediate supervisors. These people we felt, would have the opportunity to form informed opinions about the claims that were offered to them. We identified about 2500 professionals, from various locations around the world, in various projects, and sent an email with a link to the survey and solicited a response.

Which sampling strategy is described in the quote above from Devanbu et al.?

1.0 point · Multiple choice · 4 alternatives

●	**Purposive sampling.**	1.0
○	Quota sampling	0.0
○	Random sampling	0.0
○	Snowball sampling	0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

h Consider the following list of numbered activities that one goes through when conducting an interview study.

1. Conduct interviews
2. Conduct pilot interviews
3. Study the lingo
4. Draft Interview Protocol
5. Get approval from the ERB
6. Draft Research Questions

What ordering is the correct ordering?

1.0 point · Multiple choice · 4 alternatives

○ 3, 6, 5, 4, 2, 1	0.0
● 6, 3, 4, 5, 2, 1	1.0
○ 6, 4, 3, 5, 2, 1	0.0
○ 3, 6, 4, 5, 2, 1	0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

i (based on Strandberg (2019)). Aglaya and Parfyon have conducted a series of interviews and divided transcription work among themselves. They transcribed half of the interviews each, and did a round of quality control on the other half. Parfyon found it very time consuming to transcribe so he recruited two students to do the transcription of the last couple of interviews. What is the most important thing Parfyon should have done from an ethical perspective?

1.0 point · Multiple choice · 4 alternatives

○ Parfyon should have informed Aglaya that he has recruited students. 0.0

Feedback

This is indeed a good idea but it is not strictly speaking necessary.

● **The students should have signed a non-disclosure agreement before starting the transcription task.** 1.0

Feedback

Yes. Students are not employees of the university and are not covered by generic agreements of the university with its employees.

● **Parfyon's behaviour is unethical and none of the actions discussed in the other answers can help.** 1.0

Feedback

Based on feedback after the exam we have decided to also award points for this answer.

○ Quality control performed by Aglaya on the transcripts produced by the students ensures that no ethical principles have been violated. 0.0

Feedback

Aglaya has performed quality control of the transcripts, not of the process that has produced them.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

j Consider the following survey question:

Suppose you believe that an issue requires extra priority, how would you usually indicate this in a source code comment?

To what category does this question belong?

1.0 point · Multiple choice · 4 alternatives

○ Beliefs	0.0
○ Attitudes	0.0
○ Facts	0.0
● **Experiences**	1.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

k Versions of a GNU/Linux distribution called Debian are known as Buzz, Rex, Bo, Hamm, ..., Stretch, Buster, Bullseye, Bookworm, Trixie (from the earliest to the planned ones). The scale of a variable representing a version of Debian is

1.0 point · Multiple choice · 4 alternatives

○	nominal Feedback there is a natural order of versions	0.0
●	**ordinal**	1.0
○	interval, ratio or absolute	0.0
○	none of the previous answers is correct	0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

l A measure of how often the values in one distribution are larger than the values in a second distribution is an example of

1.0 point · Multiple choice · 4 alternatives

○	a p-value	0.0
○	a blocking variable	0.0
○	an interaction term	0.0
●	**an effect size**	1.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

m Intraclass Correlation Coefficient (ICC) can be used to evaluate

1.0 point · Multiple choice · 4 alternatives

● **reliability of a survey instrument, i.e., that similar results would have been obtained if the same survey would have been administered to two similar but different groups** 1.0

○ validity of a survey instrument, i.e., what one measures agrees with what one wants to measure. 0.0

○ comprehensiveness of a survey instrument, i.e., that the questions capture information relevant for the research and the answer options are as complete as possible 0.0

○ clarity of a survey instrument, i.e., that the prospective respondents will understand the questions. 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

n Which one of the following statements about Grounded Theory is ***true***?

1.0 point · Multiple choice · 4 alternatives

○ Application of Grounded Theory requires statistical methods such as ANOVA/Kruskal-Wallis test. 0.0

○ Grounded Theory assumes that the labels are known in advance. 0.0

● **Grounded Theory requires constant comparison of the emergent categories.** 1.0

○ To ensure intercoder reliability, application of Grounded Theory requires as many coders as possible. 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

o When mining version control systems from a large forge such as GitHub, which one of the following statements is ***false***?

1.0 point · Multiple choice · 4 alternatives

● **the data collection plan should focus on collecting more data of a few projects rather than less data of many projects.** 1.0

Feedback

Indeed, false as this depends on the research question(s)

○ the data collection plan should state if and how programs that are irrelevant to the research questions (e.g., toy, inactive, irrelevant) have been excluded. 0.0

○ the data collection plan should state how the artifacts created by bots/automatic tools (e.g., automatically generated commits or pull request/issue comments) have been identified. 0.0

○ the data collection plan should reflect on the impact of forks, rebasing, merging, history rewriting 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

p A student has designed a classifier to detect presence of bullying in code review comments. To evaluate the classifier the student has collected and manually labelled 723 comments corresponding to 25 largest pull requests, and used a 10-fold cross-validation. The classifier achieves 85% accuracy. Below we list three statements, select the right answer:

1. To ensure the reliability of the manual labeling, more than one labeler should have been involved.
2. The problem is unbalanced, accuracy is not an appropriate metric to evaluate the classifier.
3. Only selecting the largest pull requests is likely to bias the sample.

1.0 point · Multiple choice · 4 alternatives

○ Statement 1 and 2 are true, statement 3 is false. 0.0

○ Statement 1 is true, statement 2 and 3 are false. 0.0

○ All statements are false. 0.0

● **All statements are true.** 1.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

q What variable does one model as a random effect?

1.0 point · Multiple choice · 4 alternatives

○	Dependent variables	0.0
○	Parameters	0.0
○	Independent variables	0.0
●	**Blocking variables**	1.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

r Developers rarely express confusion in Code Review comments. Suppose a team of researchers wants to build a classifier that classifies whether confusion is present Code Review comments. What metric is most suited to measure the performance of their classifier?

1.0 point · Multiple choice · 4 alternatives

○	Gain Ratio	0.0
○	F1	0.0
○	Accuracy	0.0
●	**AUC (Area Under Curve)**	1.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

s Consider the following excerpt from the threats to validity section of Ju et al.:

This study is conducted in one company. While we recognize that similar studies from more companies are desired, we believe our study should generalize to most modern software engineering teams. Microsoft is a large company with teams working on multiple technical areas serving various customers. Its teams are self-organized and thus have the autonomy to practice various development methods and to different degrees. Furthermore, we sampled participants from two company divisions that have different products, businesses, and cultures.

To what category does the threat discussed in this excerpt belong?

1.0 point · Multiple choice · 4 alternatives

○ Internal	0.0
● **External**	1.0
○ Conclusion	0.0
○ Construct	0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

t Consider the following excerpt from the threats to validity section of Haering et al.:

We evaluated DeepMatcher by manually annotating 600 suggested bug reports for 200 problem reports. We performed two annotation tasks. One task to verify that the automatically classified app reviews are problem reports, and one to annotate whether DeepMatcher's suggested matches are relevant for developers. As in every other manual labeling study, human coders are prone to errors. Additionally, their understanding of "a relevant match" may differ, which could lead to disagreements. To mitigate this risk, we designed both annotation tasks as peer-coding tasks. Two coders, each with several years of app development experience, independently annotated the bug report matches. For the verification of problem reports, we used a well-established coding guide by Maalej et al. [27], which Stanik et al. [44] also reused for the automatic problem report classification. To mitigate the threat to validity regarding the annotation of relevant matches, we performed test iterations on smaller samples of our collected dataset and discussed different interpretations and examples to create a shared understanding.

To what category does the threat discussed in this excerpt belong?

1.0 point · Multiple choice · 4 alternatives

●	**Internal**	1.0
○	External	0.0
○	Conclusion	0.0
●	**Construct**	1.0

Feedback

Based on feedback after the exam we have decided to also award points for this answer.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

2 Part 2

10.0 points · 1 question

As appendix to this exam you have received a copy of the paper 'Leaving My Fingerprints: Motivations and
Challenges of Contributing to OSS for Social Good'. In this copy the threats to validity section has been redacted. Please read this paper and think about the threats to validity of this work.

Text

Please draft a threats to validity section for the paper. Structure this section using the categories that were discussed in this course.

Please note, you might not have to discuss all categories.

10.0 points · Open · 4/5 Page

+2.5 points

It is clearly indicated to what category a threat belongs, and this category is correct for all listed threats.

+7.5 points

At least two threats are discussed, and all of the discussed threats are accurate, realistic and are completely described.

+5 points

At least one threat is discussed, and most of the discussed threats are accurate, realistic, and are completely described. Only minor mistakes are made in the discussion of the threats.

+2.5 points

At least one threat is discussed and some of the discussed threats are accurate, realistic and are completely described. Mistakes are made in the discussion of the threats.