CS 591/791C Empirical Methods in Software Engineering & Computer Science
Midterm Exam – Spring 2010

This is a take home exam. By writing your name and student ID and signing in the box below you acknowledge that you will not seek help nor discuss the exam with anyone until after the exam due date on March 9, 2010 (hard copy, at the beginning of the class). Late exams will not be accepted. All answers must be either typed or written clearly and neatly. Include your name and page number (e.g., Page 1 of 5) on each page.

Student name	Ekrem Kocaguneli
Student ID	701095814
Student signature	

1	/10
2	/10
3	/30
4	/25
5	/30
6	/15
Total MS	/120
7	/30
Total PhD	/150

1. **Blind or even double blind studies are often used in clinical trials. Give an example of using blinding to reduce bias in software engineering or computer science experiments. (10 points)**

   In medical research double blind experiments are used to prevent the expectations of participants (subject and the researcher) from influencing the experiment. In software engineering also, the blinding can be used in multiple ways:

   i. **Blind allocation of materials**: It means assigning procedures to subject groups separately from the process according to which subjects are given any materials that they will use during the experiment. The allocation of materials to subjects can be computerized to minimize the bias and the interaction between subjects and researchers during the experiment.

   ii. **Blind marking:** Some treatments include the experiments for which an outcome is to be assessed. For example if the experiment is about comparing testing methods and the testers are to identify defects, then the result would be the identified defects. The methods would be marked prior to testing but the format of the answers would not include these markings, so the experimenter would not know the marking in advance and in that situation blind marking would be useful.

   iii. **Blind analysis:** Treatments are coded and the analyst does not know which treatment is which. As Kitchenham puts it, some statisticians believe this method to be an effective counter to *fishing for results*.

2. **Assume that developer A took twice as much time than developer B to develop a module of comparable size (in LOC). Can we say that the developer B is (significantly) more skilled that developer A? Justify your answer. (10 points)**

One important step of experimental design is to justify the outcome of measures in terms of their relevance to the objectives of the empirical study. For example in our case the outcome is the skill of the developer and LOC as a measure for skill is a poor surrogate: Using LOC needs to be justified. Furthermore, the effect of other independent variables is ignored in this experiment. Independent factors such as difficulty of modules, the programming language in which the two modules were written would greatly influence the amount of LOC produced by two developers as well as the time required for development.

Another potential pitfall in that scenario is related to the data collection. One of the guidelines for data collection is to define all software measures fully and for empirical software studies that is a little problematic. In our case, LOC has to be defined fully (whether it is executable or written lines of code, whether comments are included etc.). Also all the other independent factors (difficulty of modules, language, development environment etc.) also have to be defined and stabilized.

3. **Testing process in a software development company is composed of unit testing, integration testing, system testing, and acceptance testing. The project manager wants to determine the effectiveness of each type of testing. Specifically address the following: (30 points)**

   a. **Type of empirical study you would use to assess effectiveness of software testing types.**
      - Depending on the nature and size of our investigation, we can choose among three types of methods: case study, formal experiment and survey. In that case, we are investigating a single company and multiple projects in that company. Therefore, our study has to be scalable and probably be big in scale. For this purpose we cannot use formal experimentation, because formal experimentation requires careful controls and it is limited in size. Furthermore, according to the definition of Kitchenham et. al. basing on Basili's work, the case study definition tells that case study is focused on a single project and if the assessment is made for multiple projects then they regard it as either a case study or a survey. In our case, the company would probably want to take a look at multiple projects across the company, therefore I would choose to use making a case study across the company for assessing the effectiveness of the software testing types in that company.

   b. **Considerations you would make designing your assessment approach.**
      - The considerations I would make when designing a case study for that company are:
         i. What is my objective? I would define the objective to assess the effectiveness of different case studies.
         ii. What is the baseline to compare the results against? With some sample projects in the company I would try to form a baseline, which would tell me the baseline for the effectiveness of testing strategies.
         iii. What are my external project constraints? How big is my budget, what is the scale of the study etc.
         iv. What is my evaluation hypothesis? I would form different hypothesis to evaluate the effectiveness of each testing strategy. One example

hypothesis could be that all testing types are equally effective. Then I would see whether I reject or fail to reject this hypothesis after statistical tests.

v. What are my response variables and how I will define them? Example response variables could be the number of bugs found in each testing type, and this would depend on the explanatory variables such as amount of time spent for testing and the LOC or number of modules that were tested.

vi. What are the experimental subjects and objects of the case study? Experimental objects would be the modules in a program and testing methodologies whereas experimental subjects would be the testers.

vii. When will I use the method in the development lifecycle and when will I measure my response and explanatory variables? I would measure them at the end of each testing method.

viii. Are the data that is required for my study collectable in a reasonable time for a reasonable budget?

ix. What is going to be my confidence level in my evaluations?

x. What is my data analysis method going to be?

xi. Is this study going to yield the aimed confidence level from the data and statistical methods of choice?

- All the above considerations in fact would map to the 7 steps for a case study. The 7 steps would be as follows:
    i. Hypothesis definition: More than one hypothesis can be defined. One of them would be: All testing methods have equal effectiveness.

    ii. Select pilot projects: Discussions with domain experts from the company can tell which projects are the representative of their projects or we can choose a random sampling approach to select project or we can choose to use all projects up on a certain time.

    iii. Identify methods of comparison: We would need to define a baseline and compare the effectiveness of testing methods against this baseline.

    iv. Identify and minimize the effectiveness of confounding factors. In our case skill and capability of testers would be a confounding factor and we would try to eliminate that by trying to pick up project that were tested by similar testers.

    v. Plan case study: Planning would include defining our schedule and budget for everything that is within the concept of our case study.

    vi. Monitor case study: As the data is being collected, the study needs to be monitored whether it is going in accordance with the plan.

    vii. Analyze and report results: In this part we apply statistical methods for analysis and report our conclusions. While reporting the results we should also identify possible threats and mention their threats to the validity of our results.

c. Data & information you would capture.

- I would firstly form two sets for the data: Dependent variables and independent variables. The dependent variables are the output of our model, which are explained by the independent variables. Independent variables on the other hand could be directly measurable or could be calculate by some formulation of other independent variables. Below is a list of likely dependent and independent variables:

i. Independent Variables:
    - Number of bugs found via each testing method
    - Number of hours spent in each testing method
    - Number of people working in each testing method
    - Skill-set and experience of people working in each testing method
    - Amount of documentation generated for each testing method
    - Number of bugs missed (bugs that are found after testing and that are mapped to a particular testing method)

ii. Dependent Variables:
    - Effectiveness of each testing method: Although this is our goal, it is by definition too broad and it needs to use some combination of afore mentioned independent variables. To define the effectiveness of testing method, I would consult to domain experts and try to learn which factors are most influential for their company in terms of defining the effectiveness. For now a quick and easy effectiveness measure could be calculated as follows:
        - Effectiveness = (number of bugs found / number of hours spent) – (number of bugs missed / number of hours spent for fixing them)

d. Any constraints/limitations your approach may have.

- One limitation to this approach is that forming the company baseline could be difficult. Because if the company has no prior information regarding how much time was spent in each testing method or how many bugs were found in each method, then it would be difficult to form a baseline.
- Another limitation could be in terms of money and time. If there are no data from previous projects, then a lot of time and effort of the case study project would go to deriving data out of past projects and if the company has many past projects, then this would even take longer time.
- Furthermore, if the company has a lot of projects, then project selection methodology may be very critical. Not to introduce any bias into project selection, a random selection strategy may be adopted. However, over time the company might have changed its technologies as well as processes, which would make old projects obsolete in terms of information coming from testing methods.
- One more constraint would be the identification of confounding factors. Since the effectiveness concept is very large in terms of its definition, the confounding factors have to be defined very well and important ones should not be missed. Otherwise, the results and evaluations could be misleading.

e. Feasibility and cost effectiveness of your approach.

- If the data collection process can be automated and the selected projects could be easily decided by random selection, then the data could be quickly collected, which makes the method cost effective.
- Furthermore, once we automate the data collection, then we would be able to get data from more and more projects. Therefore our approach would scale well to the

needs of the company, which is basically an advantage of the choice of case study as our method.

- However, collecting too much data does not always mean something good. The collected data should be analyzed and cleaned of noise, which may require time. In terms of analysis and data cleansing, this approach may need too much time and too much money, depending on the content of our dataset.

f. **If you have available unlimited resources (time and money) what assessment approach you would choose?**

- If I had unlimited resources then I would choose to go with the controlled experiments. Because they allow defining and controlling all the variables, which in turn would give me the ability to identify the effect of each factor one by one. Therefore, my effectiveness definition as well as my explanation for the contributing factors to the effectiveness would be more comprehensive.
- Furthermore, my choice of controlled experiments would enable me to get the results of self-standing variables as soon as the experiment was done, so I would not have to wait for a long period.
- Since my resources are unlimited, I can experiment for a very long time; thus, eliminating the scalability issues.

4. After deployment of a software product the company is collecting the following data:

a. Details about configuration at each installation (operating system, device drivers, etc)
b. Number of users at each installation.
c. Number of failures recorded at each installation.

Describe at least two goals that could sensibly be addressed by this set of metrics. For each case (1) describe how the metrics may enable you to understand and meet your goal and (2) formulate a formal hypothesis to be tested. State clearly any limitations or reservations you might have about using this particular set of metrics for the stated purposes. (25 points)

- **Goal 1:** Company is trying to understand whether the workload due to different number of users does have an impact on the occurrence of failures.
  - **How to use metrics:** Company has the number of users related to each installation. Furthermore, company also has the number of failures that comes from each installation. In that case the by using the number of failures as the dependent variable and the number of users at each installation as the independent variable company may decide whether there is a correlation or not. To decide the correlation, company may use an r-square analysis, which is basically tells how well a regression line explains the real data points. If r-square value is above 75% or 80%, then there is a high correlation between the number of failures and the number of users at each configuration, and we will reject the null hypothesis that is given in the next bullet. However, if the r-square value is rather low, then we cannot claim that there is a high correlation between failure number and the user number, therefore in that case we would fail to reject the null hypothesis.

- **Formal Hypothesis to be tested and limitations:**
  - ***Hypothesis:*** Since the null hypothesis is the one which we want to reject, a possible null hypothesis could be as follows: **H_0:** Number of users does not affect the number of failures coming from a particular installation.
  - ***Limitations:*** As I have mentioned, r-square analysis could be a way to validate our hypothesis. Furthermore, some statistical test could also be used. For example since we have one dependent and one independent variable here, we could use a t-test at a significance level of 95% as well. However, this hypothesis also has some weak points with the available metrics. I am basing my hypothesis on the correlation between the user number and the number of failures. However, each installation has a lot of different confounding factors that may also contribute to the number of failures, such as different operating systems, different drivers installed on these operating systems as well as the purpose for which these machines are used. Therefore, before using these metrics and the hypothesis that I provided above, company will need to define and minimize the effect of such confounding factors.

- **Goal 2:** Company may also be interested in learning whether particular configurations such as driver x installed on operating system y is a source of failures or not.
  - **How to use metrics:** This is trickier than the previous goal, because the number of operating system and driver configurations is very high and company has to treat each one separately and has to keep the number of failures for each configuration separately. Once company has the configurations and the failure numbers ready, then the company can prepare a table as below:

Operating system	Driver	Failures from center 1	Failures from center 2	Failures from center 3	Failures from center 4	….	Failures from center n-1	Failures from center n
A	1	12	34	…				
A	2	22	12	…				
A	3	:	:					
B	1							
B	2							
B	3							

With the table above, we have the independent variable of center&driver (such as A&1 or A&2) and we have failures coming from different centers the driver and operating system was installed. For different combinations, we can perform a statistical analysis and decide whether the driver and operating system combination has an effect on the number of failures. Since we are interested in comparing multiple combinations at the same time Analysis of Variance (ANOVA) test could be of choice here. In the simplistic sense ANOVA gives a test of whether the means of several groups are equal, therefore we can think of it as a general form of student t-test. If the statistical test tells us that different combinations of operating system and drivers are the same, then the company will decide that driver-operating system combinations does not affect failures. However, if the statistical test tells us that driver-operating system combination errors coming from different sources are different, then the company will conclude that some driver-operating system combinations are more failure prone that other. However, before doing that we need a good hypothesis, which is given in the following bullet.

- **Formal Hypothesis to be tested and limitations:**
  - ***Hypothesis:*** The null hypothesis is the one which we want to reject, therefore we will state our null hypothesis as follows for this setting: **H_0:** Different combinations of operating systems and the drivers installed on them does not have any relation with the number of failures.
  - ***Limitations:*** This analysis is trickier than the previous one and is prone to many limitations. Some of the limitations are as follows:
    - For example while considering the operating system and driver combinations, we do not consider any user related information. A skilled person may foresee and overcome many failures, whereas a less skilled or a novice person who does not know how to use the operating system or the driver may cause a failure to occur.
    - Another limitation is that we do not consider the other drivers installed on the operating system, we only consider one operating system and one driver combination. It may be the case that the interaction of two different drivers causes the failure. For instance operating system A may work perfectly well with driver 1, whereas a combination with another driver (A&1&7) may cause a failure. If such failure prone combinations are very likely to happen, then we may conclude this analysis with misleading remarks.
    - Furthermore, the choice of statistical analysis also very important and may be a limitation to this analysis. I think ANOVA is the right choice here. However, some other statistical test that is similar to ANOVA but that characterizes the data better may give more sound results.

5. Describe the design of an experiment whose goal would be to study the reasons behind software projects running over budget. (30 points)

   a. What type of experimental design you would choose and why?

   - **The Why Part:** I will try to answer the why part first and after my explanations, I will pick up the appropriate design.
     - Firstly we need to define our objective. Our objective is to first define the factors that are influential to budget overruns and then screen out the most important ones. Therefore, we have a "*screening objective"*.
     - The other important thing we need to consider before settling with an experimental design is the number of factors that we will deal with. Although the exact numbers may change, we will have at least 5 or more

    factors that have influence on the budget overruns. Therefore, we need an efficient method to investigate all these combinations.
    - The characteristics of a factorial design are: Using resources efficiently, specifying the interactions clearly, and reducing the error through replication.
- **Selected Method:** Although we may have many levels for each factor we define, for this particular experiment I am just interested in finding out which factors contribute to budget overruns. Therefore I will let each factor have two levels. When we combine the reasons in previous bullet with the 2 level factors choice in this bullet, my choice for experimental design is going to be ***2^k factorial design***.

b. List and briefly justify the choice of independent and dependent (response) variables.

- **Independent factors:**
    - **Project manager capability:** Project manager is one of the most important factors in the success of a project. If the project is lead by a capable manager, then possible dangers can be foreseen and hindered. This factor will have two levels: 1 meaning a capable project manager and 0 meaning an incapable project manager.
    - **Analyst capability:** The project starts with the specifications and analysis of an analyst. Therefore, if the analyst is incapable the project will be built on misjudged analysis and will most likely overrun its budget. This factor will have two levels: Capable analyst represented by 1 and incapable analyst represented by 0.
    - **Programmer capability:** The programmers are key players in a software project, because capable programmers may work both faster and more reliable, thus reducing the test effort and also reducing the end product defects. Capable programmers are represented by 1 and incapable ones are represented by 0.
    - **Size of the project:** I will represent this factor with two levels: 1 for big and complex projects and 0 for relatively small projects. Although the difference of big and small for project size is not very clear, the more complex and bigger projects are harder to estimate because they entail the interaction of various complex factors. Therefore, bigger projects may be more prone to budget overruns.
    - **Existence of optimized processes:** The processes according to which a software product is built are very important. If the processes are well defined and optimized, then the estimations are more likely to be accurate and the projects are less likely to face a budget overrun. Existence of optimized processes will be represented with a 1 and their non-existence will be represented by a 0.
    - **Existence of measurement practices:** If the company has already established various measurement practices for their software practices, then they are aware of facts such as how many LOC they have developed in the past, what sort of errors they have received and how much time they have spent for each project under which conditions. Such a measurement

experience puts the company in a better position in terms of estimating their projects and this reduces the budget overruns. The existence of measurement will be represented by 1 and non-existence will be represented by 0.

- **Dependent factors:** These factors are explained by the independent factors that are given above.
    - **Whether there was a budget overrun in project:** The actual cost of the project depends on the independent factors. A model built on the independent factors (like a linear regression model) can give us the project cost. By comparing the actual cost and the estimated cost, we can decide whether there was a budget overrun or not. Therefore, the existence of a budget overrun will be represented by 1 and non-existence of it will be represented by a 0.

c. How you would select subjects and objects for your experiment? Explain your choice.

- Subjects and objects need to be representative of the population, because the lack of this would make the conclusions of the study invalid. In my design, I would choose the projects randomly (a random selection), because this would eliminate most of the bias inherent in the selection methodology. Furthermore, if the objects and subjects were not selected by random, then we would need to justify that our results still represent the whole population and are still valid.
- Another beneficial thing to do while selecting the objects and subjects would be to define some inclusion and exclusion criteria. The two criteria for selection that I would have for this study are given below:
    - **Inclusion criteria:** All the projects whose output was a software product and whose attributes can be mapped and collected in accordance with the dependent and the independent variables that were defined before can be included.
    - **Exclusion criteria:** If in any part of the selected projects are majorly handled by non-professionals (students etc.), then this project is not likely to represent industry standards and therefore will be excluded.

d. Discuss the possible threads to validity of your experiment.

I will address the threats to validity under 4 categories:

- **Construct Validity:** Construct validity is asks whether we are testing what we intend to test. This is a particularly important, since the lack of construct validity makes the complete study invalid. In this study, we follow the guidelines of design of experiment principles and indicate each factor clearly. However, for a comprehensive study, the afore mentioned 2 level factors may fall short of explaining all aspects of the factors and that may be a construct validity to this homework study.
- **Internal Validity:** Internal validity tries to ask whether the results elicited in a study are merely due to experiment manipulations. Since we are selecting our

projects by random selection and since we are covering all possible combinations of available factors via 2k factorial design, I think the study has internal validity.

- **Conclusion Validity:** The conclusion validity tries to understand whether the conclusions made in a particular study are justified. If we employ statistical tests in our study and base our conclusions on the combination of statistical analysis and the 2^k factorial design observations, then the conclusion validity would hold. However, if we suffice to merely comment on our observations without any statistical tests, then our results would lack conclusion validity.
- **External Validity:** The aim of external validity is to ask how and in what circumstances some results are still applicable The external validity of this study would depend on the selection pool of our projects and randomization would not help in that situation. For making our results externally valid, we would want our pool to cover projects from different business domains, coded in different languages under different programming practices and if possible we would like our software projects to come from diverse centers (both national and multinational). Otherwise, selecting projects from single business domain, from single city and coded in single programming language with the same type of programming methodology would completely remove the external validity of our results and would make it valid just for that particular setting.

e. How would using a case study instead of an experiment be different? Why you would chose a case study over an experiment?

- The characteristics of a case study are:
  - Being more of an observation
  - No control on the field
  - Being easier to plan
  - Allowing for complexity
  - Allowing unpredictability
  - Being scalable to big settings
  - Being easy to plan and harder to interpret
- On the other hand formal experiments has rather different characteristics:
  - Requiring appropriate level of replication
  - Randomly chosen experimental subjects and objects
  - Being carefully controlled (not allowing for non-predictability)
  - Often times being small in scale (not scalable)
  - Difficult to control when degree of control is limited (hence they tend to be small in scale)

Therefore, when I think of the above mentioned facts about the case studies and experiments I would choose case study over an experiment:

- When I need a scalable study which can explain big industrial project settings (experiments are small in size)
- When I need an observation type of result rather than very precise analysis
- When I need to represent a typical situation instead of a whole population (experiments have random sampling to represent the whole population)

- When I do not need to plan in too much detail and do not need to generalize my conclusions but just need to explain the effects of a phenomenon in a particular setting
- When I can wait and see the actual results of events (in experiments we can isolate self standing tasks and experiment to see the effect of this task without actually waiting for an actual project to finish).

6. Determine which of the following statements are meaningful. In each case justify your answer. In cases when the statement is not meaningful provide a meaningful and correct statement. (15 points)

   a. The length of program A is 50.

      - For experiments, we need to collect data and to collect data we need to define our measures fully in accordance with a unit of measure. This sentence is not meaningful because although it states the entity (program) and its feature (length) it does not state the type of the measurement and it does not indicate which scale it uses for this measurement.

   b. Program A is twice as long as program B.

      - This sentence is not meaningful either, because it tries to define a ratio scale on the length entity of the attributes (program). However, the measurement (whether its lines of code, whether it is number of functions) and measurement's unit as well as the measurement type (whether the length was elicited via a direct measurement or a calculated measurement) is not given. Therefore saying program A is twice longer as program B has no meaning with that amount of information.

   c. The design of program A is twice better than the design of program B.

      - This sentence has no meaning due to the reasons that are similar to those that were given in the previous part. In other words, we know that the entity is program design. However, we do not know the attributes and saying goodness of design does not tell anything, we need to define attributes that can describe the goodness of a design. Furthermore, like the attributes themselves no measure and no units for these measures are given in the sentence. By saying twice better, the sentence tries to use an ordinal scale and a ratio scale; however, due to lack of any unit or measure for comparison such usage of ordinal scale is not appropriate and due to the lack of measure for goodness the ratio scale is not appropriate either. Due to all these reasons, this sentence does not tell anything and does not have a particular meaning.

   d. The average size of Windows application program is about four times that of similar DOS program.

      - This sentence tries to use a ratio scale on the size attribute of the entity application program. However, there is no information about the measure that used for measuring the size of a program. Furthermore, since there is no measure regarding the entity of interest, there is also no unit of this measure. Therefore, when this sentence speaks about size, it does not tell anything what the size refers to and when it says four times, it does not tell how such a ratio was calculated. Hence, this sentence has no meaning.

e. Tool A achieved higher mean usability rating than tool B.

Note: Assume that program usability is rated on the following four-point scale:

4: Can be used by a non-programmer
3: Requires some knowledge of Java
2: Usable only by someone with three years Java programming experience
1: Totally unusable

- We have a 4 point ordinal scale for the attribute of usability, which is used for the entity of tool. There is no problem with having an ordinal scale in this study, however when we speak of mean, then we may have some problems, because summing the ordinal scale and taking the mean may not always give the overall picture of the usability of a tool. Furthermore, mean is a measure that is biased towards extreme values (minimum and maximum), therefore by using mean we may end up with wrong conclusions. Therefore, although this sentence makes more sense than the previous sentences, it is still not hundred percent true in its way of making use of ordinal scale via mean.

7. **PhD students only.** Suggest one paper that you would like to add to the list of papers assigned as homework readings.

   a. What is your motivation for choosing this particular paper?

I chose to evaluate the paper Cross versus Within-Company Cost Estimation Studies: A Systematic Review by Kitchenham et. al.

- The full citation of the paper is: B. A. Kitchenham, E. Mendes, and G. H. Travassos. Cross versus within-company cost estimation studies: A systematic review. IEEE Trans. Softw. Eng., 33(5):316–329, 2007.
- **Reasons related to topic, authors and journal:** My first motivation to choose this paper was that it's related to my research area (software effort estimation). Secondly, the paper presents a systematic review of the previous studies in cross-company and within-company effort estimation studies and I would like to get an overview of the topic as well as how a thorough systematic review is conducted. Thirdly, the paper's first two authors are B. A. Kitchenham and E. Mendes, who have published very respectable research work in very decent journals and they are particularly good in terms of their methodological explanations as well as details in their papers. Finally the paper was published in IEEE Transactions of Software Engineering that has a very strict evaluation procedure.
- **Reasons related to content:**
  - I particularly like the methodology part of this paper. It first starts by defining what systematic literature review means and lists the lists the reasons for a systematic literature review as well as all the steps that shall be followed while conducting a systematic literature review.

    - Furthermore, they propose 3 research questions which are very precise in their topic. However, what I liked about the research questions is that before defining research questions, they state the properties of good research questions, which are: Population, study factor, comparison intervention, and outcome.

    - They have also performed a very thorough data analysis (in that case data are previous studies). They have defined data in terms of quality questions and have weighted each question. Then they have given a quality note for each study they have worked with.

b. What you expect others to learn from the paper? Discuss both the positive and negative aspects (e.g., presentation style, novelty of the approach, significance of the results).

- **What I liked about the study:**
    - The best thing to me about the study was its rigorous methodology and rigorous data analysis. They have defined the scope of their study very well by explaining the properties of a literature review and they have also defined their research questions in accordance with research question properties. Furthermore, in their methodology part they have defined properties to identify the cross vs. within company studies and they have given them weights. I think these features as well as weights can help future studies in terms of what aspects of the study shall be considered more important than the other aspects.
    - The results part was very well organized. They have worked on ten different projects and it is difficult both to present and interpret ten different studies that were evaluated according to various criteria. However, they have followed the research questions while presenting their results and have made extensive study related tables that include defined factors versus ten different studies.
    - There are also general notes for effort estimation studies in the results section. They have presented tables that list the procedure factors in ten different studies and have listed the advantages and disadvantages of these factors. I regard these tables as very helpful for future studies.
- **What I did not like about the study:**
    - They have made their analysis on 10 papers only. That is also a validity issue that they have recognized. Because 10 papers is just a small part of the whole literature. Although they have gone through an extensive search strategy for picking up these ten studies (they have gone thorough manual analysis of 1344 retrieved projects from their search queries), I still think that they could have limited the factors they question and increased the number of projects in their study.
    - One very important threat to the validity of their results is that they have not considered the model that was utilized in the studies. However, I think the utilized model can have very different effects in terms within vs. cross-company results. They could have at least recognized the used models in

terms of parametric or non-parametric methods. Because each model may have very different assumptions, and comparing results coming from models with totally different assumptions can be misleading.

Notes: Please restrict your choice to peer reviewed papers published in IEEE & ACM journals and conference proceedings. Provide a full reference of the paper & hard copy stapled to your exam. (30 points)