An Experimental Study of Individual Subjective Effort Estimations and Combinations of the Estimates

Martin Höst
Dept. of Communication Systems
Lund University
P.O. Box 118
S-221 00 LUND, Sweden
+46 46 222 90 16
martinh@tts.lth.se

Claes Wohlin
Dept. of Communication Systems
Lund University
P.O. Box 118
S-221 00 LUND, Sweden
+46 46 222 33 29
claesw@tts.lth.se

ABSTRACT

The required effort of a task can be estimated subjectively in interviews with experts in an organization in different ways. Interview techniques dealing with which type of questions to ask are evaluated and techniques for combining estimates from individuals into one estimate are compared in an experiment. The result shows that the interview technique is not as important as the combination technique. The estimate which is best with respect to mean value and standard deviation of the effort is based on an equal weighting of all individual estimates. The experiment is performed within the Personal Software Process (PSP).

Keywords

Cost estimation, experimentation, measurement, metrics, project management.

1 INTRODUCTION

An important aspect of software engineering is the ability to do estimations of process and product attributes. Bad estimates result in that costs and deadlines are overrun. Some important attributes to estimate for projects are the cost of the project, the lead time of the project and the reliability of the delivered products. These attributes are of course not independent, but in most cases estimations are performed for the attributes one at a time, and this paper focuses on estimation with respect to cost, and specifically with respect to required effort. It has been decided to focus on required effort since this is one of the most important parts of the cost of a project.

Effort estimation is hard in general, but a number of methods for estimation exist in the literature, e.g., Boehm's COCOMO model [3] and methods based on Albrecht's function point approach [4]. These, and a number of other methods, rely on historical experience to predict required effort through a prediction relationship. This relationship can, for example, be derived by linear regression. This means that in order to predict the required effort for a certain task, the same or similar types of tasks must have been performed before. It also means that the required effort must have been measured every time the tasks have been performed.

This paper deals with effort estimation techniques that are not based on quantitatively measured experience from former projects. Instead the methods rely on subjective estimations performed by experts in the organization. This can in many cases be a good complement to quantitative measurement. The advantage of this type of method is that it can be applied even if an extensive set of measurement has not been performed before. This is, for example, the case when the software development process recently has been changed or when the process should be changed for the project where the effort should be estimated. In the literature, only a few methods have been proposed for doing subjective estimations. Probably, the most well known method is the method described by Putnam [10] for subjective estimation of length of a program. In [6] six techniques for subjective estimation of effort are evaluated together with Putnam's method used for effort estimation.

In this paper, the above mentioned techniques are presented and compared in a controlled experiment. The experiment is focused towards the area of software engineering, and in particular process improvement when the gathered experience for the new process is limited. The same type of experiment can however be performed in other areas whenever a number of methods for estimation should be evaluated. The same applies to the usability of the proposed estimation techniques. The techniques could be used in any area where subjective estimations would be appropriate.

The experiment presented in this paper is the continuation of a pre-study experiment. The pre-study experiment was performed in retrospect with respect to the task for which the effort was estimated, but it did anyhow indicate that the proposed estimation techniques could be used for subjective estimations. The design of the pre-study experiment has been used as a basis for the design described in this paper. Since the pre-study experiment involved five participants and the experiment described in this paper involves 26 participants, the experiment described in this paper is substantially larger than the pre-study experiment.

In this paper a number of different alternative techniques

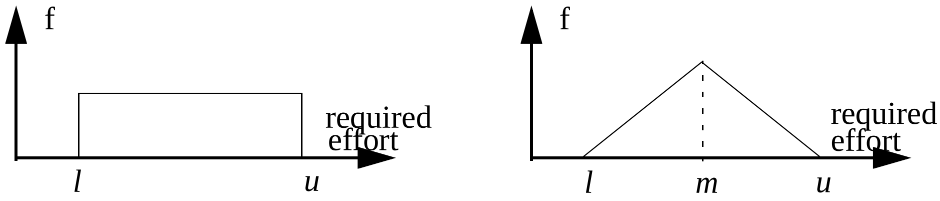

FIGURE 1. The probability density function of the rectangular distribution and the triangular distribution.

for subjective estimation of effort are presented in Section 2. In Section 3, a design of an experiment [1], [8] is presented, which can be used to determine which alternative technique to use in a certain environment. In Section 4 the operation of the experiment is outlined and the results of the experiment are presented in Section 5. In Section 6 the paper is summarized.

2 EVALUATION OF SUBJECTIVE EFFORT ESTIMATION TECHNIQUES

2.1 Introduction

Subjective effort estimation can be done in a number of different ways. The objective of the experiment presented in this paper is to evaluate some different possible techniques and determine if there is a statistically significant difference between them. The evaluated techniques are described in this section.

2.2 Methods for Subjective Estimation

2.2.1 Subjective Estimation

Subjective estimations can be performed by interviewing people for their opinions of the required effort. This type of interview is typically based on that people provide intervals describing the required effort according to their opinion. If only one point has to be given by an expert this could result in wrong estimates, for example, in order to ‘be on the safe side’ and it would not describe the certainty or uncertainty of the expert. An interval estimate, on the other hand, also takes into account the certainty of an estimate. A large interval corresponds to a large uncertainty and a small interval corresponds to a small uncertainty.

There are, however, no commonly used and accepted method for performing this type of interviews. Only some work has been carried out in this area, and the most famous is probably described in [10] where subjective estimation of the program length is addressed. One of the methods (alternative 7) for subjective estimation of effort which is evaluated in the experiment is based on the method described in [10].

In this paper the interviewed persons are denoted experts. In the experiment presented subsequently, the interviewed persons are not software engineering experts with long industrial experience. The people are considered to be experts from the point of view that they have the best knowledge of the task for which the effort is estimated.

2.2.2 Subjective Estimation Process

Subjective estimation can be performed by letting experts individually estimate parameters such as highest possible value, most likely value, etc., and then based on individual values from a number of different experts synthesize a single estimate of a factor of interest. A typical factor of interest is the most likely value of the required effort, but also the standard deviation of the required effort is important, because it can be seen as a measure of the certainty of the experts. A high standard deviation corresponds to a high uncertainty and a low standard deviation corresponds to a low uncertainty. Subjective estimation of the required effort can be performed by carrying out the following three major steps:

1. Presentation of prerequisites to experts.
2. Individual estimation.
3. Synthesization and overall estimation.

These steps are described subsequently.

2.2.3 Presentation of Prerequisites to Experts

Before the experts can estimate the required effort according to the specified procedure, they must be presented with the prerequisites of the estimation process. They must be presented with the task for which the effort should be estimated. The experts can, for example, be given a requirements specification and be asked to estimate the time it takes to derive a design of the system, or they can be given a design and be asked to estimate the time it will take to do the coding of the system, or the testing, etc.

The experts must also be presented with how to do the actual estimations. This may be some kind of instructions presented, and they may be combined with some kind of estimation form. Examples of different types of questions to ask the experts are: What is the lowest possible required effort for the task in question? What is the most likely required effort for the task in question? What is the highest possible required effort for the task in question? This type of questions can be combined with further instructions on how to answer them. This can, for example, be that the lowest, most likely and highest values are points from a triangular distribution. This means that the experts are asked to consider the actual required effort as drawn from a triangular distribution which they themselves decide the parameters of.

In the experiment, the following three prerequisites have been considered:

- A rectangular distribution: The experts are asked to estimate a lowest possible value and a highest possible value of the required effort. All values between these two extremes should be equally likely to occur in reality.

Basic distribution presented as prerequisite to experts		
Rectangular	**Triangular**	**No distribution**
alt 1: $f = \frac{1}{n}\sum_{i=1}^{n} f_i$	**alt 4:** $f = \frac{1}{n}\sum_{i=1}^{n} f_i$	**alt 7:** Beta distribution
alt 2: Rectangular distribution with $l = \min l_i$ and $u = \max u_i$	**alt 5:** Triangular distribution with $l = \min l_i$, $m = \bar{m}$ and $u = \max u_i$	
alt 3: Rectangular distribution with $l = \bar{l}$ and $u = \bar{u}$	**alt 6:** Triangular distribution with $l = \bar{l}$, $m = \bar{m}$ and $u = \bar{u}$	

TABLE 1. Summary of distributions presented to experts and overall synthesized distributions.

- A triangular distribution: The experts are asked to estimate a lowest possible value, a most likely value, and a highest possible value of the required effort according to a triangular distribution.
- No distribution at all: The experts are simply asked to estimate a lowest possible value, a most likely value, and a highest possible value of the required effort, without considering any distribution at all.

Using the distributions described above, requires that the experts have some knowledge of the concepts that they are based on. In the experiment presented in this paper, graphical representations, such as the ones in Figure 1 have been used to describe the rectangular and the triangular distribution.

2.2.4 Individual Estimation

In the second phase of the subjective estimation process, the experts individually estimate the parameters according to the prerequisites. This will in most cases involve interviews or filling out some kind of forms specialized for the purpose. This means that one subjective estimation is collected from each expert for a particular prerequisite.

In the experiment, the experts filled out forms individually. This may also be done with some kind of Delphi-method [3], where the experts first reach some kind of consensus concerning the required effort. The effect of this kind of method has, however, not been analysed in the experiment.

2.2.5 Synthesization and Overall Estimation

The objective of this step is to synthesize the individual estimates from the previous step into an overall estimate and then to do the overall estimations based on the synthesized overall distribution. This can be done in a number of different ways, and in the experiment three different alternatives have been evaluated for the two first prerequisites (rectangular distribution and triangular distribution). For the third prerequisite (no distribution at all) only one alternative has been evaluated. How this has been done is described in Section 2.4. The seven alternatives are described in more detail in Table 1, where f_i denotes the distribution (rectangular or triangular) estimated by expert i, while l_i, m_i, and u_i denote the lowest possible value, most likely value and highest possible value estimated by expert i, and n denotes the number of experts.

Alternatives 1 and 4 are based on an average distribution, i.e. a distribution which is formed by considering all the individually estimated distributions and giving equally much attention to all individual distributions. This way of forming a synthesized distribution is displayed graphically for two experts and the rectangular distribution as prerequisite in Figure 2, alternative 1. Alternatives 2 and 5 are based on the widest distribution that can be formed based on the individual estimates from the experts. This is graphically displayed in Figure 2, alternative 2. Alternatives 3 and 6 are based on a distribution formed by the average of the lowest possible values and the average of the highest possible values. This is graphically shown in Figure 2, alternative 3. Alternative 7 is based on the Beta distribution.

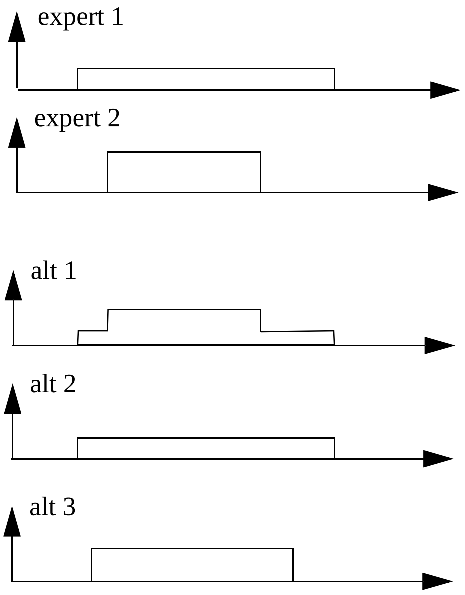

FIGURE 2. Alternatives 1 - 3 displayed graphically for the case with two experts.

Two different factors are important to determine based on the synthesized overall distribution. One is the mean value, which can be used as a prediction of the most likely value or mean value of the required effort. It is also important to estimate the accuracy of that prediction, which is done via the standard deviation of the overall synthesized distribution. The seven alternative methods for subjective effort estimation are further described in [6], where it is also described how to calculate the mean and the standard deviation of the overall synthesized distributions related to the seven different alternatives.

2.3 Objective of Evaluation

The objective of the experiment presented in this paper is to determine how to perform the estimation process described in the above sections. That is, which prerequisites should be used and how should an overall synthesized distribution be derived. This means that it is important to decide what we mean with a good subjective effort estimation process and with a bad subjective effort estimation process. This is probably not the same in every organization and for every occasion. It is not only important to determine the predictability with respect to the value of the effort itself (the mean value). It is also important to look at the predictability with respect to the accuracy of the required effort (the standard deviation). In the experiment presented, measures of prediction accuracy for the mean value and for the standard deviation have been used. That is, the response variable of the experiment is the prediction accuracy of the different alternative methods. This is further described below.

2.4 Pre-study

The seven estimation alternatives have been compared in a pre-study [6] performed during the spring of 1996 at Lund University in Sweden. The pre-study was, as this experiment, performed in the context of the Personal Software Process (see below). The major differences in comparison with this study are that the pre-study was performed in retrospect and with fewer participants. The results of the pre-study which are valid for this study can be summarized as:

- There could not be found any real difference between the two prerequisites rectangular distribution and triangular distribution. For alternative 7, it could be seen that the results indicate that the prerequisites rectangular distribution and triangular distribution are better than the prerequisite no distribution at all. The difference is, however, small and the significance not certain, and therefore all seven alternatives should be further considered.
- Nearly all participants perform the same estimations for the two prerequisites triangular distribution and no distribution. This means that the prerequisite no distribution is not used in the experiment. For alternative 7, the prerequisite triangular distribution is used instead.

3 THE EXPERIMENT

3.1 Introduction

The experiment can be used to evaluate the subjective estimation process. The presentation in this section describes a specific experiment which has been carried out and the result of the experiment is presented in Section 5. The description is intended to be general enough to provide information enough to carry out replications in other organizations and situations. It should also be straightforward to adjust the experiment in order to evaluate other prerequisites and other choices of overall synthesized distributions.

It should be possible to perform similar and related experiments in any type of organization, not only for software development tasks, but for any task in any area where the required effort should be estimated based on subjective data.

3.2 Experimental Environment (PSP)

The experiment has been performed in a student environment and more specifically in the PSP course. The course has been followed by Ph.D. students at Linköping University in Sweden during the spring of 1997. About half of the participants conduct research in computer science and software engineering. The other participants work in non computer science areas and perform research in, for example, control theory. These persons do not perform research in software engineering, but they use and develop software in order to solve problems in their research.

The PSP course involves developing ten different programs (#1A-#10A), and each participant develops the programs independently. The programs deal with list handling, counting the number of lines of code using a coding standard, and statistical analysis. For further information concerning the PSP programs, refer to [5]. All development is done according to a defined process that is the same for every software engineer. The process is enhanced during the assignments, from a basic process in the first assignment to a more advanced process in the later assignments. The assignments each require about 100 - 500 minutes to perform. Assignment #10A requires more time than the other assignments, while assignments 1-9 require about the same time.

3.3 Experimental Variables

3.3.1 Independent and Dependent Variables

As the objective is to evaluate the effect of different alternative methods for estimation of the predictability, the independent variable (see for example [4]) is the choice of alternative. This means that there is one independent variable which can take seven different values (treatments) in the experiment.

As the predictability should be determined both for the estimation of the mean value and for the standard deviation (std), there are two dependent variables in the experiment. The relative prediction error has been used as a measure of predictability for both the mean value and the standard deviation. This means that the following two dependent variables have been measured for every alternative:

- | estimated mean - experienced mean | / experienced mean

- | estimated std - experienced std | / experienced std

Since these measures involve both the experienced mean and the experienced standard deviation, they cannot be measured until after the task has been completed. It means also that a task for which the measure is determined must be carried out independently a number of times by different people. It is not necessary that a task is performed by the experts that provide the subjective estimates, but it is possible to do so.

3.3.2 Block

The objective of the experiment is to determine the effect of the independent variable on the dependent variables. There are, however, in many experiments a number of nuisance sources that will interfere with the effect of the independent variables. In this experiment the effect of the different assignments (PSP assignments #1A-#10A) has been identified as one such effect. It is easier to estimate the effort for some assignments than for other.

Actions have been taken to distinguish between the effects of the assignments and the independent variable. That is, the design involves the assignment number as blocking variable.

3.4 The Design

The design is a completely randomized block design. For every assignment (block, #1A-#10A), the required effort is estimated according to the different alternatives (treatment, alt 1 - alt 7). This has been done with the same persons for all alternatives and all assignments. Different people could be used, and that would result in one additional blocking variable. This is, however, not the case in this experiment.

The relative prediction error is determined for both the mean value and for the standard deviation.

The randomization imposes that the estimation of effort should be done in random order with respect to the different alternatives for every program. If this is not done it is not possible to distinguish the effect of the order (alt 1, alt 2, alt 3...) from the actual effect of the different alternatives that actually is of interest. This means that the experts should do the estimations with respect to the different prerequisites in random order for every assignment. This is further discussed in Section 4.

3.5 Analysis

The analysis of the described design is based on the relationship $y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$ where y_{ij} is the relative prediction error for prediction alternative *i* and assignment *j*, μ is an overall mean, τ_i is the effect if the *i*:th treatment (estimation alternative, *i* = 1, 2,... 7), β_j is the effect of the *j*:th block (PSP assignment, *j* = 1, 2,... 10) and ε_{ij} are independent random error terms equally normally distributed with mean 0. The sum of all treatment effects, as the sum of all block effects, is 0. This model can be analysed with standard analysis of variance (ANOVA) procedures with the null hypothesis

$$H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_7$$

where $\mu_i = \mu + \tau_i$. That is, the null hypothesis states that there is no effect of the estimation technique on the prediction error. Notice that rejecting this null hypothesis will not point out which estimation technique is the best, merely that all methods are not equally good. To distinguish between the different alternative estimation techniques, a number of different techniques can be used, such as the Least Significant Method and Duncan’s Multiple Range Test [7]. In the experiment the former has been used.

3.6 Threats to Validity of Experiment

There are two different kinds of threats to the validity of the experiment (see for example [2], [9]), internal and external. Internal threats deal with the validity of the result of the experiment and indicates if the result of the specific experiment is valid. External validity deals with the generalizability of the result, that is, if the result is general and valid for other settings than the specific experiment. The following internal threat has been identified for the experiment:

- One common threat is due to maturation, which is concerned with the effect that the participants’ attitude to the experiment changes during the experiment. Since the experiment is run during one whole semester this is an important threat. During the PSP course other estimation techniques are included in more advanced processes. This may mean that the participants do not see the benefit of doing the subjective estimations during the later assignments, since they anyway deal with estimations according to the PSP process. No such effects have, however, been noticed. This means that even if the threat is considered important, this type of error is not considered to be disturbing the result.

As the external threats are concerned with generalization of the results of the experiment, it is important to identify the areas in which the result should be consider general. When this is determined, the external threats can be indicated by the difference between the area of the validity of the experiment and the required area for validity. The type of experiment presented here can never be general for every organization or every occasion, but the sought for validity would be that it is valid in an development organization and for tasks performed by more than one person. Thus, the following external threats have been identified:

- The assignments for which the effort is estimated is small and performed by only one person. The result of the experiment is only valid for applications such as the ones used in the experiment. This is further discussed in Section 5.2.
- The participants are not a sample drawn from industrial practitioners. The participants are Ph.D. students, but some of them have industrial experience. The participants are further discussed in Section 5.2.

4 OPERATION OF EXPERIMENT

The experiment has, as mentioned above, been performed at Linköping University, involving Ph.D. students as participants. The experiment was performed with 26 participants who individually estimated the required effort for every assignment according to every prerequisite. This means that every estimation alternative (1-7) could be used for every assignment involving every participant's individual subjective estimations. An advantage of letting every participant estimate according to every prerequisite is that the participants will elaborate their estimates more thoroughly, and if the participants estimate for every prerequisite in the same order every participant will have equally much help in their estimates in the PSP course. A disadvantage of letting every participant estimate according to every prerequisite in the same order is that the experiment cannot be randomized appropriately (see Section 3.4). In the experiment it was, however, decided that every expert should estimate according to the two prerequisites in the same order (rectangular, triangular).

5 ANALYSIS AND INTERPRETATION

5.1 Analysis of Variance

The collected data have been analysed according to Section 3.5, and the result of an analysis of variance (ANOVA) for the mean value and for the standard deviation can be seen in Table 2 and Table 3.

TABLE 2. Summary of analysis of variance with respect to relative prediction error of mean value.

Source of variation	SS	df	MS	F_0	p-value
Treatment (alternative)	1.13	6	0.19	5.89	<0.01
Block (assignment)	0.30	9	0.03		
Error	1.72	54	0.03		
Total	3.15	69			

TABLE 3. Summary of analysis of variance with respect to relative prediction error of standard deviation.

Source of variation	SS	df	MS	F_0	p-value
Treatment (alternative)	4.49	6	0.75	23.1	<0.01
Block (assignment)	0.64	9	0.07		
Error	1.75	54	0.03		
Total	6.89	69			

As it can be seen in Table 2 and Table 3, the estimation technique significantly affects the relative prediction error. F_0 is larger for the standard deviation than for the mean value, which means that the effect is larger for the standard deviation than for the mean value.

The analysis described above shows that there is a significant effect of the estimation technique on the relative prediction error of both mean value and standard deviation. Nothing has, however, been said concerning which techniques that are better than the others etc. To determine the difference between the different techniques, the mean treatment effects ($\mu_i = \mu + \tau_i$ according to the model described in Section 3.5) for the different techniques are studied, see Figure 3.

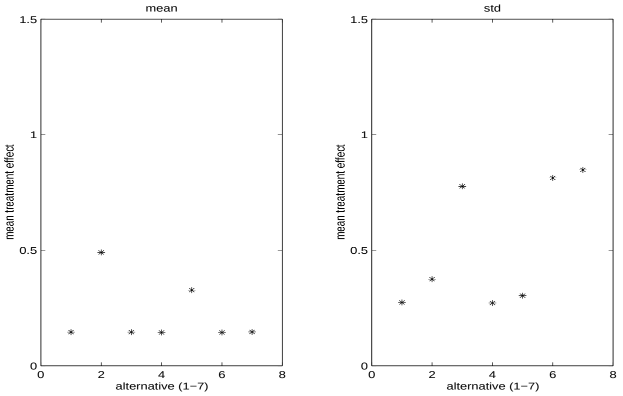

FIGURE 3. Treatment effects for relative prediction error with respect to mean value and standard deviation for the seven alternatives.

It can be seen that some of the alternatives perform better than the others. As the treatment effect in the experiment is the relative prediction error, a low treatment effect is good and a high treatment effect is not good.

Concerning the estimation of mean value, alternatives 1, 3, 4, 6, and 7 are significantly (significance level 0.05) better than alternative 5, which in turn, is significantly better than alternative 2. This means that the alternatives can be divided into three groups with respect to predictability of mean value: good (alternatives 1, 3, 4, 6, and 7), medium (alternative 5) and bad (alternative 2). Concerning estimation of standard deviation, alternatives 1, 2, 4, and 5 are significantly better than alternatives 3, 6, and 7. This means that the alternatives can be divided into two groups with respect to predictability of standard deviation: good (alternatives 1, 2, 4, and 5) and bad (alternatives 3, 6, and 7). This is summarized in Table 4.

TABLE 4. Comparison of the different alternatives.

		standard deviation	
		good	bad
mean	good	alternatives 1, 4	alternatives 3, 6, 7
	medium	alternative 5	
	bad	alternative 2	

The alternatives that perform best with respect to mean value, are the alternatives that are based on distributions that take into account the results from all experts. The alternatives that are not performing so well are alternative 2 and 5. These alternatives are based on the opinion from the expert who estimated the lowest value of all. The other experts' opinions concerning the lowest possible value is not considered. The same thing is true for the highest possible value. Only the opinion of the expert who estimated the highest possible value is considered. This means that the opinion of one single expert can impact the estimate of the mean value very much, and this does not seem to be appropriate when the mean value should be estimated. Concerning alternative 2 and alternative 5 it can be seen that alternative 5 is better than alternative 2. This means that if this way of deriving the synthesized distribution is used, even if it is not a good way of deriving it, it is better to base the interview questions on a triangular distribution than on a rectangular distribution.

The alternatives that perform best with respect to estimation of standard deviation are the alternatives that are based on the widest synthesized distributions. Alternatives 2 and 5 are based on the highest and lowest individual estimations, and alternatives 1 and 4 take all individual estimations into account. This means that the density functions of the synthesized distributions for alternative 1, 2, 4, and 5 are larger than zero for all values between the lowest value estimated by any expert and the highest value estimated by any expert. This is not the case for alternatives 3, 6, and 7. The parameters of these distributions are instead given by the mean value of the lowest possible value and the mean value of the highest possible value.

The alternatives that perform best with respect to estimation of both mean value and standard deviation are alternatives 1 and 4. These alternatives are based on synthesized distributions that are derived based on a wide distribution which considers all different distributions estimated by the experts, and which considers all individual distributions to the same extent. This means that this seems to be the best way to derive the synthesized distribution.

Concerning alternative 1 and alternative 4, which are significantly better than the others, there is no significant difference between alternative 1 and alternative 4. This means that it cannot be said that alternative 1 is better than alternative 4 or vice versa. Consequently, if the synthesized distribution is derived in the way that have been shown to be the best (alternative 1 and alternative 4) there is no significant difference between using the rectangular or the triangular distribution as prerequisite.

The result of the experiment is important to industrial projects, because subjective opinions are an important part of planning. The process for subjective estimation described in Section 2.2 can be used when projects are planned. Therefore it is important to know which type of questions to ask and how individual estimates should be combined into one estimate within the estimation process.

5.2 Replication of the Experiment

As it was said in Section 3.6 there are some threats to the external validity of the experiment. It is not certain that the assignments are representative to all assignments in all organizations, and it is not certain that the participants are representative compared to the participants in all organizations' projects. Before the results and the conclusions are used from the experiment, the experimental conditions must be compared to the conditions at hand. If the differences are too large, this may be a reason to replicate the experiment in the specific context with the specific participants.

For replication purposes it is important to consider how many participants that are necessary in order to reach a reliable result in an experiment as the one described. It is not possible to determine exactly how many participants that are necessary in a general experiment. It is, however, possible to investigate the effects of randomly excluding participants from the study and then investigating what the results of the experiment would be with a smaller number of participants. In Figure 4, the p-value is displayed for experiments with 100 randomly chosen constellations for every number of participants from 2 to 24. This is displayed for the result of the experiment with respect to the relative prediction error for both the mean value and the standard deviation.

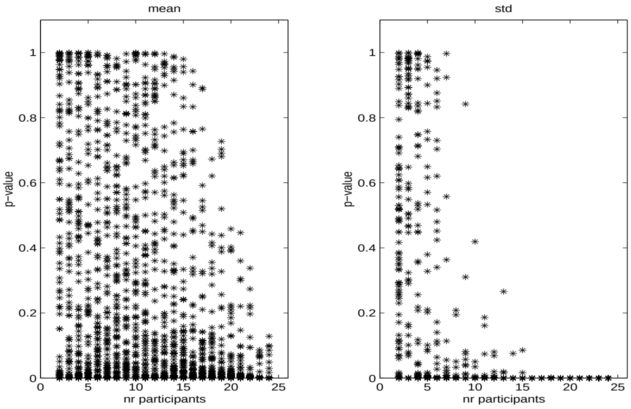

FIGURE 4. p-value for experiments with 100 randomly chosen constellations for every number of participants.

With respect to standard deviation, it can be seen that if there are less than about 12 participant the result of the experiment is very much depending the choice of participants. On the other hand, if there are more than about 12 participants, the result is not that dependent on which participants that are chosen. For example, all 100 constellations of 20 participants produce a result very close to 0 with respect to standard deviation. It seems like more people are necessary in the experiment to get a significant result with respect to mean value than to get a significant result with respect to standard deviation. Figure 4 shows that even a small number of participants can be removed from the experiment resulting in that the experiment can not show a significant difference between the alternatives. There are, for example, a number of different constellations of 20 persons that will not result in any significant difference between the alternatives.

In the experiment presented in this paper a large number of participants (26) have been used. This was easy to obtain because the experiment was performed within the PSP course, which was attended by a large number of participants. It is, however, not certain that this large number of participants can be obtained in any organization. The effect of using fewer participants is that it is harder to find a significant difference between the alternative methods.

6 SUMMARY

Subjective estimation of the required effort of a task can be carried out in a number of different ways. When the estimation is carried out based on interviews with experts, these interviews can be performed based on a number of different prerequisites and hence different questions. The answers from the interviews can also be combined in a number of different ways.

Seven different alternative techniques for estimation have been compared in the experiment. The comparison has been made with respect to the relative prediction error for both the mean value and the standard deviation of the required effort. The experiment has been conducted within the context of the Personal Software Process (PSP).

For the alternatives that have been found to be best, there is no difference between the performance of the different prerequisites. This means that it is not important if the experts are asked for the lowest possible required effort, the most likely required effort, and the highest possible required effort or if they are only asked for the lowest possible required effort and the highest possible required effort.

It is, however, important how the different experts' estimations are combined. Here, it seems like it is best to base the estimation on a density function derived as

$$f = \frac{1}{n}\sum_{i=1}^{n} f_i$$

where the f_i:s are the experts' individually estimated density functions and n is the number of experts. This distribution takes into account all the experts' individual estimations to the same extent.

The results indicate that it is necessary to involve a relatively large number of people in the experiment in order to get a significant result. Based on the data from this experiment, it is recommended that more than 20 persons should participate, if the experiment is to be replicated.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the Ph.D. students that participated in the PSP course in Linköping in 1997.

This work was partly funded by The Swedish National Board for Industrial and Technical Development (NUTEK), grant 1K1P-97-09673.

REFERENCES

[1] V. Basili, R. Selby, D. Hutchens. Experimentation in Software Engineering. *IEEE Transactions on Software Engineering*, Vol. 12, No. 7, 1986, pp. 733-743.

[2] V. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sørumgård, M. Zelkowitz. The Empirical Investigation of Perspective-Based Reading. *Empirical Software Engineering*, Vol. 1, No. 2, 1996, pp. 133-164.

[3] B. Boehm. *Software Engineering Economics*, Prentice-Hall, 1981.

[4] N. Fenton. *Software Metrics - A Rigorous Approach*, Chapman & Hall, 1991.

[5] W. Humphrey. *A Discipline for Software Engineering*, Addison Wesley, 1995.

[6] M. Höst, C. Wohlin. A Subjective Effort Estimation Experiment. *Information and Software Technology.* Vol. 39, No. 11, 1997, pp. 755-762.

[7] D. Montgomery. *Design and Analysis of Experiments*, John Wiley & Sons, third edition, 1991.

[8] S. Pfleeger. Experimental Design and Analysis in Software Engineering, Part 1-5. *ACM Sigsoft*, Software Engineering Notes, 1994-1995, Vol. 19, No. 4, pp. 16-20, Vol. 20, No. 1, pp. 22-26, Vol. 20, No. 2, pp. 14-16, Vol. 20, No. 3, pp. 13-15, Vol. 20, No. 4, pp. 14-17.

[9] A. Porter, L. Votta. An Experiment to Assess Different Defect Detection Methods for Software Requirements Inspections. *Proceedings 16th International Conference on Software Engineering*, 1994, pp. 103-112.

[10] L.H. Putnam, A. Fitzsimmons. Estimating Software Costs. *Datamation*, September, 1979.