A method for evaluation of the performance of requirements inspections

E. Kantorowitz, L. Arzi and A. Harel
Computer Science Dept
Technion – Israel Institute of technology
32000 Haifa
email: kantor@cs.technion.ac.il

Abstract
A new improved model is introduced for the performance of inspectors of user requirements documents. The model, which is based on a linearity assumption employs only two non-dimensional metrics in the [0;1] interval. They characterise the error detection ability of the inspectors and their domain knowledge level. The application of the model is demonstrated with an analysis of which kind of inspectors, low salaried students or experienced high salaried engineers, will perform a specified inspection most economically.

1. Introduction
The user requirement elicitation process results often in a user requirements document (URD). It has been observed that faults in this document can significantly increase the software development costs. Fairly [1], for example, reports cases in which it was 100-200 times more costly to correct faults detected during system operations than at the requirement elicitation phase. It is therefore useful to inspect the URD and correct the detected faults. The inspection method introduced by Fagan [2] with later variants [3] is considered to be an efficient way to detect URD faults. An overview of current inspection methods is found in [4] and[5]. Inspections involve considerable costs and one of the goal of this study was therefore to determine the smallest number of inspectors j that will produce a satisfactory result. Further savings may be achieved by employing the most economical kind of inspectors (lowest cost per detected fault) and the most economical inspection method. In order to find such inspectors and inspection method we need metrics for quantitative comparisons of different kinds of inspectors and different inspections methods. Metrics that meets these needs were developed in [6]. Unfortunately the method of [6] for measuring these metrics requires experiments with quite many inspectors. This paper introduces an improved method enabling measuring the metrics with a more modest number of inspectors.

2. The model
The efficiency of a particular inspection process is measured by the fault detection ratio (*FDR*), which is the ratio between the number of detected faults and the total number of faults in the document[6]. The total number of faults is unknown, but it can be estimated in a number of different ways. This study employed the traditional fault injection method. In this method faults designed to represent the real faults faithfully are injected into the URD before the inspection. *FDR* is estimated as the ratio between the number of detected injected faults and the known total number of injected faults. Our model may however be employed with other methods for estimating the total number of faults such as the capture-recapture [7] and defect-profile methods [8]. Comparisons of these methods are found in [9] and [10].

Our model employs the notation $P_{i,j}$ for the probability that *j* inspectors will detect fault number *i*. The given *n* faults are numbered from 0 to *n*-1 such that

$$P_{0,1} \geq P_{1,1} \geq P_{2,1} \geq ... \geq P_{i,1} \geq ... \geq P_{n-1,1} \qquad (1)$$

where $P_{i,1}$ is the probability that fault *i* is detected by one inspector. This probability may be estimated by the ratio between the number of inspectors that have detected this fault and the total number of inspectors. It has been observed in the experiments of [6] that some inspectors do not have sufficient domain knowledge to be able to detect all the *n* faults. The are only able to detect faults 0 to $n_{\max} - 1$, where $n_{\max} \leq n$. Faults $n_{\max}$ to *n*-1 can not be detected by them. The domain knowledge of the inspectors is thus characterized by the number:

$$FDR_{\max} = \frac{n_{\max}}{n} \qquad 0 \leq FDR_{\max} \leq 1 \qquad (2)$$

It was observed in all the experiments made in [6] that $P_{i,1}$ is roughly a linear function of *i*. We have no explanation for this linearity. It is therefore assumed that

$$P_{i,1} = P_{0,1} - \frac{P_{0,1}}{FDR_{\max}} \frac{i}{n} \qquad 0 \leq \frac{i}{n} \leq FDR_{\max} \qquad (3)$$

The correctness of this linear expression may be checked by inserting $i=0$ and $i=n_{max}$ which produce the correct values $P_{0,1}$ and 0, respectively. It is seen from equation (3) that the probability $P_{i,1}$ $0 \leq i < n_{\max}$ that one inspector detects fault *i* is proportional to $P_{0,1}$. A high $P_{0,1}$ value means therefore high $P_{i,1}$ values.

$P_{0,1}$ *is* <u>a measure of the abilities of the employed inspectors to detect faults in the given document with the employed inspection method</u>

FDR_{max} <u>is a measure of the inspectors' URD domain knowledge.</u> The inspectors do not have sufficient domain knowledge to detect more than this proportion of the faults.

The probability $P_{i,j}$ that error *i* will be detected by *j* different inspectors may be estimated by

$$P_{i,j} = 1 - (1 - P_{i,1})^j \qquad (4)$$

We insert expression (3) in (4) and introduce $x = i/n$:

$$P_{i,j} = 1 - \left(\frac{P_{0,1}}{FDR_{\max}}\right)^j \left(\frac{1 - P_{0,1}}{P_{0,1}} FDR_{\max} + x\right)^j \qquad (5)$$

FDR may be estimated by adding the probabilities for detecting faults 0 to $n_{\max} - 1$:

$$FDR(j) = \frac{\sum_{i=0}^{n_{\max}-1} P_{i,j}}{n} \qquad (6)$$

In order to achieve a more simple expression we assume $n \to \infty$ n whereby (6) is transformed to:

$$FDR(j) = \int_{x=0}^{FDR_{\max}} P_{i,j}(x)\, dx \qquad (7)$$

After some computations we get

$$FDR(j) = FDR_{\max}\left(1 - \frac{1-(1-P_{0,1})^{j+1}}{P_{0,1}(j+1)}\right) \qquad (8)$$

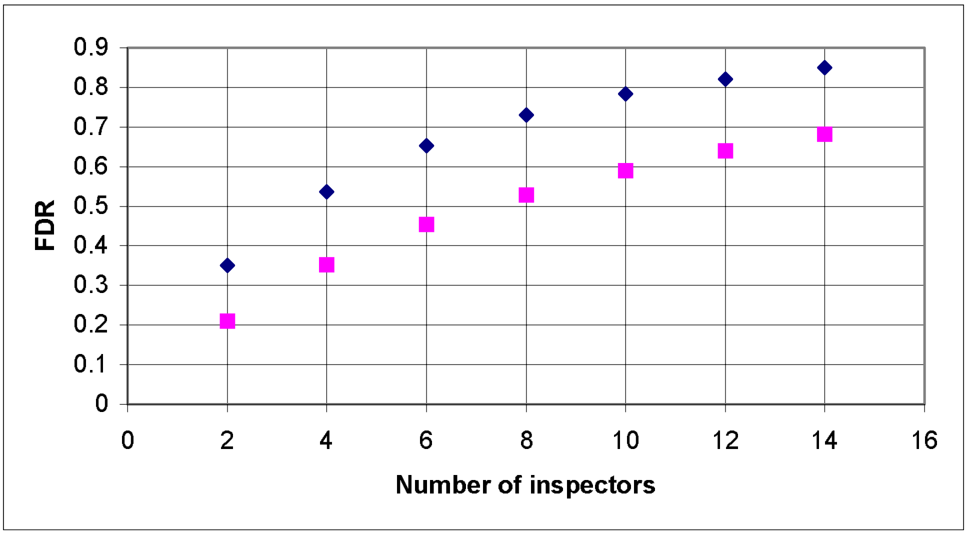
This $FDR(j)$ function is called the *performance function* of the inspection process. This equation estimates under the made assumptions the FDR value achieved by j inspectors. The function has two coefficients $FDR_{\max}$ and $P_{0,1}$ which must be measured. These two numbers are the metrics of the model, which are measured in an experiment with N different inspectors. The set of detected injected faults of each one of these N inspectors is recorded. These N different sets are employed for computing FDR_j for $j=1,...,N$. FDR_j is the FDR observed for j different inspectors. There are $n_{N,j} = \binom{N}{j}$ sets of j different inspectors among the N inspectors. For each one of these $n_{N,j}$ different sets of j inspectors we compute FDR_j by taking the number of faults in the union their detected injected faults. This number is divided by the total number of injected faults. Finally we estimate FDR_j as the average of these $n_{N,j}$ different FDR_j values. The unknown metrics $0 \le FDR_{\max} \le 1$ and $0 \le P_{0,1} \le 1$ are estimated from these N different observed FDR_j values by looking for the $FDR_{\max}$ and $P_{0,1}$ values that minimize:

$$D = \sum_{j=1}^{N} (FDR_j - FDR(j))^2 \qquad (9)$$

where $FDR(j)$ is computed by (8). When the values of these two metrics have been determined we can solve the problem that motivated the study, i.e. to determine the smallest number of inspectors j for which $FDR(j) \le FDR_{required}$. The following experiment illustrates how it is done.

3. Application of the model

The validity of the model has been tested practically with both students and industry engineers. The results of such an experiment are depicted in the figure below



Comparing the performance of engineers (marked by ♦) and first year computer science students (marked by ¦) as function of the number of inspectors. Both engineers and students worked in teams of two. Results are therefore shown for 2, 4,.. inspectors. The experiments employed the missile launching URD specified in an appendix in [6]. The inspection time was 40 minutes (20 minutes for the preparation phase and 20 minutes for the collection phase)

The values of the two metrics $FDR_{\max}$ and $P_{0,1}$ of our model were estimated by minimizing expression (9):

	$P_{0,1}$	$FDR_{\max}$
Students	0.40	0.99
Industry Engineers	0.74	0.99

It is seen that $FDR_{\max} = 0.99$ for both kinds of inspectors. This means that practically all faults may be detected by both kinds of inspectors. $P_{0,1}$ is however much higher for the engineers than for the first year students (0.74 versus 0.40). This means that the engineers are much better to detect the faults. This will be illustrated by the following example. Assume that an $FDR \geq 0.5$ is required. It is seen from the figure that this can be obtained by four teams of two students who achieve $FDR = 0.53$. With engineers only two teams of two are required, as they achieve $FDR = 0.54$. The required engineers' effort is $2 \cdot 2 \cdot 40/60 = 2.67$ hours while the student effort is $4 \cdot 2 \cdot 40/60 = 5.33$ hours, i.e. twice as many hours as the engineers. If the hourly salary of the students is less than half that of the engineers, then employing the students will be the most economical solution. This example compared the performance of two different kinds of inspectors. The model may also be employed to compare two different inspection methods for the same kind of inspectors.

Acknowledgement

The division of work among the authors of this paper is that L. Arzi and A. Harel managed the experiments and analyzed their results. The development of the mathematical model and the method for computing the coefficients of the model are the work of E. Kantorowitz. The authors wish to thank the persons who participated patiently in our experiments.

References

[1] Fairley, R ,"Software Engineering Concepts", McGraw-Hill, New York, 1985.

[2] Fagan, E. , "Design and Code Inspections to Reduce Errors in Program Development," IBM Sys. J.,15,3(1976), pp 182-211

[3] Gilb, T. and Graham, D. , "Software Inspection", Addison Wesley, Reading Mass.,1993

[4] Porter, A. , H. Siy and Votta, L. , "A Review of Software Inspections", in Advances in Computers, Vol. 42, Academic Press, NY, 1996, pages 39 - 76.

[5] Porter, A., Votta L.G. Jr., and Basili, V.R. ,"Comparing Detection Methods for Software Requirements Inspections: A Replicated Experiment", IEEE Transactions on Software Engineering, Vol. 21, No. 6, June, 1995, pages 563-575.

[6] Kantorowitz , E. , Arzi, L ., Gutmann, A. , "The Profile of N-fold Requirements Inspection Method", Requirement Engineering Journal,2,(1997), pp. 152-164 .

[7] Eick, S. G. , Loader, C. R. , Long, M. D. , Votta, G. and Wiel, S. V. , "Estimating Software Fault Content Before Coding", Proceeding 14th International Conference on Software Engineering, (1992),59-65

[8] Wohlin, C. and Runeson, P. , "Defect Content Estimation from Review Data", Proc. 20th International Conference on Software Engineering, (1998), 400-409

[9] Wohlin, C. , Runeson P. and Brantestam, Johan , "An Experimental Evaluation of Capture-Recapture in Software Inspections", Software Testing, verification and Reliability,5(1995),213-232

[10] Briand, L. , El Eman K. and Freimut B, "A Comparison and Integration of Capture-Recapture models and the Detection Profile Method", Proc. 9th International Symposium on Software Reliability Engineering,(1998), 32-41