Empirical Software Engineering, 5, 215–228 (2000)

Picking the Right Problem Frame—An Empirical Study

KEITH PHALP kphalp@bournemouth.ac.uk
Empirical Software Engineering Research Group, School of Design, Engineering & Computing, Bournemouth University, Poole House, Talbot Campus, Fern Barrow, Poole, Dorset, UK

KARL COX coxk@bournemouth.ac.uk
Empirical Software Engineering Research Group, School of Design, Engineering & Computing, Bournemouth University, Poole House, Talbot Campus, Fern Barrow, Poole, Dorset, UK

Abstract. Problem frames are a relatively new approach to requirements engineering, promising benefits not only in elicitation but also in subsequent design, by allowing their users to select methods and techniques appropriate to their given problem domain.

In order to be effective this approach relies upon the correct identification of relevant problem frames for a given problem or scenario. Hence, we examine whether people are able to identify the correct (relevant) frames for a given set of problem descriptions, and whether they can correctly gauge the relative contribution of each identified frame to the given problem.

We note the Euclidean distance of (individual and group) answers from an expert solution, considering each problem frame as a separate dimension. Examination of this distance (or magnitude of error) allows us to gauge the accuracy with which people can assign problem frames. We compare the performance of individuals within groups, and the performance where groups work together to provide a collective solution, comparing both of these with a fair-distribution strategy.

We found that people can choose the relevant frames with a reasonable degree of accuracy, but that this is improved where they work to provide a collective solution. We also note differences among groups, for example, that experience appears to improve the accuracy with which groups can collectively choose relevant frames.

Keywords: requirements, problem frame, problem domain

1. Introduction

Requirements Engineering is often considered to be one of the most important aspects of software development (Boehm, 1982), and yet paradoxically one where there is little emphasis either within the education of software engineers, or within the software engineering community (Bray, 2000). Therefore, it is perhaps not surprising that many software-engineering disasters are often attributed to failings within the requirements process (Glass, 1998).

This problem is exacerbated by the inappropriateness of analysis methods for describing problem domains (Bray, 2000). Furthermore, where the attempts at improving elicitation (Jacobson, 1994) have become heavily associated with analysis and design notations (Brooch et al., 1999) this has led to greater confusion over semantics (Cox and Phalp, 1999). Although there has been considerable research recently in describing requirements using scenarios (Weidenhaupt et al., 1998), this research has tended to focus on the production of heuristics and tools to guide the writing of, rather than the discovery of, functional requirements. CREWS (CREWS) suggest that their L'Ecritoire tool aids the discovery of goals (Rolland et al., 1998), since the attachment of scenarios to goals helps the discovery of further goals, but this is still essentially a post-elicitation exercise (arguments about the iterative nature of the software process excepted).

Hence, Jackson (1995) and subsequently Kovitz (1999) have argued for a new approach to requirements elicitation, which sees 'framing' the problem, and analysis of goals as a central activity, and which is separate from the subsequent specification of the system. The concept of these 'problem frames' was originally described by Jackson (1995) and has been extended by Kovitz (1999) to provide guidance about which notations are most appropriate for a given 'problem frame.' In essence, if one can correctly identify which type of problem (frame) one is dealing with, one knows how best to model it. This seems intuitively sensible. For example, one might suspect that the approach most suitable for building a word-processor would not necessarily be the best choice for building an Air Traffic Management System. Bray (2000) has suggested that this 'Problem Domain Oriented Analysis' PDOA (his term) is a significant methodological departure, which may be more appropriate to requirements elicitation and description than either structured (Yourdon, 1989) or object-oriented analysis (Coad and Yourdon, 1991). However, though promising, PDOA is largely untested, and is clearly reliant on whether analysts can correctly identify what kind of problem frame is being encountered. If not, then the approach is flawed, since an inappropriate frame would be used. Therefore, as a first step in ascertaining the utility of a 'problem frame' approach to requirements engineering we sought to investigate whether people could accurately determine the correct problem frame.

The rest of this paper is as follows. Section Two describes the background to problem frames whilst Section Three outlines the approach taken by our investigation. Section Four contains some of our results, section Five summarises the implications of these results, and finally section Six offers some conclusions and suggestions for further work.

2. Problem Frames

Historically the differences between requirements and specification activities have been blurred (Bray, 2000), distinctions often being made on the basis of their intended audience; requirements documents being seen as primarily customer focussed, with specification intended for designers.

Jackson departs from this approach by describing specification as the interface between the system and the application domain, and thus equivalent to external design (Jackson, 1995). Conversely, requirements analysis is viewed as a wholly different activity, which is primarily concerned with a description of the problem domain. In fact, Jackson further states that ideally there should be two specifications, that is, a natural language specification for the customer and a more formal specification for the designer (Jackson, 1999). Hence, the distinction between requirements and specification is about what is being modelled, and not, as if often supposed, the intended audience.

With this distinction clarified, recent initiatives in requirements engineering have been able to concentrate upon the problem domain, and specifically in knowing and categorising (framing) the type of problem.

> ‘*The point of this is that knowing the **type of problem** helps frame the analysis and specification (or external design), guiding what questions to ask, what aspects to model and, ultimately, the internal design*.’ (Bray, 2000)

The idea of classifying application domain types is far from new, and many would admit differences between say, information systems and real-time applications, systems software and applications software, or desktop systems versus embedded systems (Budgen, 1994). Indeed, even the formal education of those studying information systems differs from those studying software engineering; students learning different programming languages, design methods and so on. For example, traditional information systems design (Connolly et al., 1999) is typically data dominant, whilst real-time systems design (Yourdon, 1989) could be considered process dominant. These distinctions tend to be clearer in design and implementation than in the upstream requirements and specification activities.

Michael Jackson has suggested that this classification idea may be extended to the framing of problem domain, in order to aid requirements analysis. He suggests an initial set of five application/problem frames (recognising that there may be more to come). Aside from the advantages to requirements analysis already noted, problem frames also give possibilities for requirements reuse. Furthermore, since the same kind of problems may suit the same kind of solutions, and solution methods, there are possible advantages to be gained in subsequent specification and design.

We note that to utilise problem frames one would need (at a minimum) to correctly recognise that a problem fitted a relevant, standard frame or frames and then fit the frame to the problem. It it this need for correct framing that is the focus of our study, since the utility of the problem frame approach depends upon it.

3. The Pilot Study

Our intention was to investigate whether individuals (or groups) could accurately assign a particular problem to the correct (or relevant) problem frame or frames. A pilot study was carried out, using two distinct groups. Hence, we not only wished to investigate whether any individuals or groups could choose frames accurately but also whether there might be differences among groups. Group A comprised eight part-time students, all currently working within the software industry, but with differing levels of industry experience. Group B comprised fourteen full-time post-graduate students. Most students in Group B had little or no recent industrial experience, though some did have past experience of the software industry, in one case where an individual worked as a programmer for 15 years. Hence our expectations were that group A would perform better than group B, and all of our tests of significance between these groups (section Four) are single-tailed.

In brief, the study required us to:

- Produce a list of problems, or descriptions of problems, which covered the existing frames.

- Conduct a simple test where individuals or groups choose the frame or frames relevant to each problem description.

- Devise a measure of how far away the individual (or group) was from a correct or expected (expert) answer. This distance measure was to indicate the magnitude of error, for each question.

These steps will now be described in detail.

3.1. Study Requirements

Problem Descriptions: Altogether there were 22 problem descriptions, intended to cover the entire range of problem frames. These included examples of our own and some taken from Bray (2000) and Kovitz (1999). These descriptions were used as the basis for a simple problem frame test.

The Test: Before the test, each group was given the same lecture, introducing problem frames and the exercise. Both groups were given the same descriptions, and initially had to assign the frames to each description individually (under test conditions). For each problem (question) they had to list the relevant frames in order of importance, and assign a percentage score to each listed frame, denoting the percentage of the problem within that frame. Hence, each problem must have (a total of) 100 percent across 5 fames, irrespective of the number of frames relevant to it. Once the individual part of the test was completed, the groups had to come to a collective agreement on each problem, again noting the frames and the percentage contribution of each. Each test (exercise) took two hours for each group.

Distance: In deciding how far the answer to a question was from the expert view, we chose to take a simple Euclidean distance. We treated each frame as a dimension, and considered each answer as a point in the five dimensions. Distance is thus the square root of the sum of squares of the difference in each axis (frame); which conveniently means that distance equates to the magnitude of relative error for each question. Rather than use percentages we considered each axis, or frame, as being bounded by 0 and 1. Hence, the Euclidean distance, or relative error, for any question lies between 0 and the square root of the number of axes (in our case 5). Again, in order to aid analysis, we therefore divided each relative error by root 5, so that all corrected distances lie between 0 and 1.

The Expert Solution: The derivation of an expert solution was also a multi-person exercise, loosely based on the Delphi estimation technique. Before carrying out the tests, staff (all teaching problem frames), each produced their own answers. In doing so we used our own judgement, and other sources, such as existing problem frame examples. We met to discuss our answers and went away to produce a second draft. Having carried out the exercises, we then came together with our final draft answers and argued until we came to a consensus view for each answer.

3.2. Questions for the Study

The exploratory study concentrated upon finding out whether it was possible to correctly assign problem frames to given problem descriptions. This initial goal also raised the following questions.

1. What is the magnitude of error for individuals or for collective (group) answers?

2. Are individuals able to recognise problem frames with an 'acceptable' degree of error?

3. Are there significant differences between the two groups?

4. Do they perform better than by chance or some fair strategy by which an individual with no knowledge could select problem frames?

The first question simply requires the derivation of some measure of error, one of the goals for our study. The acceptability of this error (question Two) is largely subjective, in that one typically compares the error of some technique with a competitor. For example, one might suggest that case based reasoning is a more accurate predictor of effort than rule induction (Mair and Shepperd, 1999). However, there is no such competitor for problem frames. Therefore, any comparison we make is not like with like, and we must make some (subjective) judgement as to whether the magnitude of error seems reasonable. Hence, the study will provide a measure, to answer question One, and data for question Two, but does not attempt to test any specific hypotheses relating to these questions. Questions Three and Four are more standard, in that one could describe them in terms of some hypotheses, as follows:

H1: Group A (with greater experience) perform better at choosing problem frames than group B, when working individually.

H2: Group A (with greater experience) perform better at choosing problem frames than group B, when working collectively.

H3: Both groups (A & B) perform better than a 'Fair Distribution Strategy' (FDS) that evenly allocate all frames to each problem, when working individually.

H4: Both groups (A & B) perform better than a 'Fair Distribution Strategy' (FDS) that evenly allocate all frames to each problem, when working collectively.

Therefore, the following section will attempt to examine our result, and specifically to answer these hypotheses.

4. Results

Our results section is structured as follows. Section 4.1 gives an overview of the derivation of our baseline, which we refer to as a fair distribution strategy (FDS), and which we consider as being equivalent to chance. Section 4.2 examines group performance when subjects produced answers individually, noting differences between the groups, their relative accuracies and the performance against FDS. Finally Section 4.3 examines the performance of groups when they were allowed to come to a collective answer.

4.1. A Fair Strategy to Represent Chance

In the absence of any competing techniques, we compared groups A and B with an artificial baseline, which represents an even-handed distribution of problem frames to each question. A simple strategy would have been to randomly choose a single frame for each question (weighting at 100%). However, the chance result is then very dependent on the random allocation, and this seems unlikely to give a representative view across only two groups. Instead we chose a 'Fair Distribution Strategy' (FDS), which assumes that all frames are equally relevant to all questions, and gives a 20% weighting to each frame. This FDS and our expert solution would have led to a mean magnitude of error of 0.32, across both question and subject analyses.[1] Hence, this figure serves as our baseline, both in comparison of individual (section 4.2) and collective performance (section 4.3).

4.2. Individual Performance

In comparing the performance of groups working individually we examined differences in two ways. Firstly, we examined the groups' mean error for individual questions, showing how (on average) each group responded differently to the question set. Secondly, we examined differences between subject means showing differences in accuracy across questions.

Table 1 (below) shows mean distances[2] from the expert solution for each question for both groups (and additionally for the fair distribution strategy). Note that there was a substantial range of accuracy across questions. First, consider group A. In one case (question 15), all subjects within group A had exactly the same answer as the expert solution. Contrast this with question 21, where there was a mean corrected error (distance) of 0.38. It would appear that some questions, or problem situations were harder to correctly gauge than others. Of course one might expect this; questions 15 and 16 (the two lowest errors for group A) were both single frame (perhaps more obvious) scenarios, whereas questions 11 and 21 (the highest) each had three relevant frames. Group B also appear to have found some questions much harder than others. However, it does not appear that both groups agreed about which questions were most difficult. Although question 15 was the most accurately answered by both the similarities ended here. For example, by far the largest error for group B was on question 14, which appeared to cause no particular problem for group A. Hence, there appeared to be inter-group differences in terms of which questions were most problematic (an issue worthy of further study).

Since we examined the performance of these groups on questions, across students, it seemed appropriate to examine the statistical significance of these results by considering their respective mean responses to each question as our variables. This allowed us to carry paired t-tests, comparing group A with group B, and each group with the fair distribution strategy (see Table 2).

Our first result (a comparison of question means for groups A and B) is not significant. Hence we reject H1; we cannot say that group A's individual performance across questions was significantly better than group B's. However, the comparison of both groups with the fair distribution strategy does yield significant results, with probabilities of less than

Table 1. Distances from solution for each question.

Question	Question Means (chance)	Question Means (Group A)	Question Means (Group B)
1	0.31	0.15	0.18
2	0.35	0.10	0.13
3	0.35	0.10	0.15
4	0.31	*0.46*	0.27
5	0.40	0.34	0.22
6	0.40	0.23	0.14
7	0.35	0.25	0.19
8	0.30	*0.39*	0.26
9	0.26	0.16	0.20
10	0.35	0.16	0.28
11	0.26	*0.37*	*0.34*
12	0.35	0.14	0.14
13	0.15	*0.28*	*0.29*
14	0.40	0.19	*0.51*
15	0.40	0.00	0.11
16	0.40	0.07	0.18
17	0.33	0.09	0.23
18	0.40	0.17	*0.21*
19	0.17	*0.26*	*0.32*
20	0.24	*0.32*	*0.24*
21	0.22	*0.38*	*0.39*
22	0.40	0.15	0.23
Means	0.32	0.22	0.24

Table 2. Paired t-tests of group performance on questions.

Paired (one tail) T-tests	
Group A & B question means.	0.1805
Group A question means & FDS	0.0038
Group B question means & FDS	0.0051

0.01. Hence, we support H3; it does appear that both groups performed better than a 'Fair Distribution Strategy.'

We also examined the performance of individuals across questions, giving a picture of the accuracy of an individual subject in each group. Table 3 shows combined subject means, with those within group A being ranked.

Since (for subject) we had different sized groups, we adopted a single tailed Mann-Whitney U test to test for significance. The U test statistics can also be seen in Table 3. Our P value of 0.097 suggests that, taken across questions, subjects in group A were better at estimating problem frames than subjects in group B. However, we note the relative weakness of this result, and the fact that it contradicts our previous analysis by question. Hence, overall our examination of H1 appears inconclusive, and we cannot claim that group A performed significantly better, when working individually, than group B.

Table 3. Examination of subject means.

A & B combined	A & B combined (& sorted)	Ranks for y (group A)	Mann-Whitney	
0.2645	0.0747	1	$n1$ (cases in y)	8
0.1557	0.1557	2	$n2$ (cases in x)	14
0.2049	0.1698		$n1 + n2$	22
0.2258	0.1801		$n1(n1 + 1)/2$	36
0.1991	0.1823		$n1n2$	112
0.199	0.1836		$R1$	73
0.2614	0.199	7		
0.0747	0.1991	8	U	37
0.2487	0.2049	9	U prime	75
0.2054	0.2054			
0.2196	0.2141		Variance	214.66667
0.2691	0.2196			
0.1698	0.2258	13	Expected value	56
0.2603	0.2487		Z	**1.2967949**
0.2841	0.2603		P	**0.0973509**
0.3043	0.2614	16		
0.2141	0.2645	17		
0.1801	0.2691			
0.3034	0.2841			
0.3031	0.3031			
0.1836	0.3034			
0.1823	0.3043			
Sum of ranks		73		

4.3. The Performance of Groups Working Collectively

In our previous exercise individuals had worked alone to come to their own answers. Having done this, they then argued the merits of each question in order, before coming to a collective (agreed) answer. This seemed an appropriate test, since it is a variant of a strategy that is well used in cost-estimation, and which is suggested to increase the accuracy of such estimations. Table 4 shows the difference between the two collective answers (groups A and B) and the expert answer, and additionally the difference between the groups.

The first thing to note is the huge difference in the accuracy of the two groups. This is largely attributable to the very marked increase in accuracy of group A. (Group A now have a corrected mean magnitude of question error of 0.043). Note that this collective score is also more accurate than the most accurate subject within group A. This suggests that although that individual may have been influential, the opinions of others have also been heeded. Indeed, in nine cases group A's collective answer completely agreed with the expert answer. We also note an increase in overall accuracy when group B tackled the question collaboratively, (with a corrected mean magnitude of question error of 0.179). Note that four of the group B subjects were this accurate individually, with one scoring 0.17 and three 0.18. Hence, group B seems to have collectively become about as accurate as its best individuals, but no more. Finally, note that the largest distance is between the two groups,

Table 4. Distances from expert and distances between groups.

Question or Scenario	Euclidean Distance between A Collective and expert	B Collective and expert	A Collective and B Collective
1	0.00	0.17	0.17
2	0.06	0.06	0.00
3	0.00	0.00	0.00
4	0.06	0.28	0.35
5	0.00	0.06	0.06
6	0.00	0.00	0.00
7	0.00	0.06	0.06
8	0.06	0.06	0.06
9	0.06	0.19	0.25
10	0.06	0.29	0.28
11	0.06	0.41	0.36
12	0.05	0.06	0.03
13	0.04	0.45	0.47
14	0.00	0.63	0.63
15	0.00	0.00	0.00
16	0.00	0.00	0.00
17	0.09	0.09	0.19
18	0.00	0.00	0.00
19	0.13	0.33	0.45
20	0.13	0.00	0.13
21	0.13	0.46	0.51
22	0.00	0.32	0.32
Means	0.04	0.18	0.20

indicating that the expert solution falls somewhere between both groups' version, though 'nearer' to group A. In fact this is somewhat surprising, since one would normally expect both groups to fall to some side of the expert answer. Again this suggests that for many questions the groups had a very different point of view.

We admit that since there is no definitive solution it may just be that group A's view happens to coincide with the expert view, both of which may be inaccurate. Pragmatically, though, it does seem likely that collectively group A are better at choosing appropriate frames than group B.

Table 5 summarises the performance of groups A and B as individuals, A and B as groups, and the FDS for each question. As for our examination of individuals we again use single-tail paired t-tests to compare accuracy (Table 6).

We previously noted (section 4.2) that working individually, both groups performed significantly better than the FDS. However, the collective performance of both groups when compared to the FDS provides even stronger results. We found a mean magnitude of question error of 0.04 for group A, and 0.18 for group B, compared with 0.32 for FDS. In addition, paired t-tests show significant results for both groups versus the FDS. Furthermore, we note that group A outperform the FDS on every question.

An even more striking result is the collective performance of group A compared to that of group B, with a probability of 0.0011. This confirms that group A working collectively were

Table 5. Comparison of collective, individual and fair distribution.

	FDS	Individuals		Collective	
Question	Question Means (chance)	Question Means (Group A)	Question Means (Group B)	Question Means (Group A)	Question Means (Group B)
1	0.31	0.15	0.18	0	0.17
2	0.35	0.1	0.13	0.06	0.06
3	0.35	0.1	0.15	0	0
4	0.31	*0.46*	0.27	0.06	0.28
5	0.4	0.34	0.22	0	0.06
6	0.4	0.23	0.14	0	0
7	0.35	0.25	0.19	0	0.06
8	0.3	*0.39*	0.26	0.06	0.06
9	0.26	0.16	0.2	0.06	0.19
10	0.35	0.16	0.28	0.06	0.29
11	0.26	*0.37*	*0.34*	0.06	*0.41*
12	0.35	0.14	0.14	0.05	0.06
13	0.15	*0.28*	*0.29*	0.04	*0.45*
14	0.4	0.19	*0.51*	0	*0.63*
15	0.4	0	0.11	0	0
16	0.4	0.07	0.18	0	0
17	0.33	0.09	0.23	0.09	0.09
18	0.4	0.17	*0.21*	0	0
19	0.17	*0.26*	*0.32*	0.13	*0.33*
20	0.24	*0.32*	*0.24*	0.13	0
21	0.22	*0.38*	*0.39*	0.13	*0.46*
22	0.4	0.15	0.23	0	0.32
Means	0.32	0.22	0.24	0.04	0.18

Table 6. Results of paired t-tests (individual and collective).

Paired (one tail) t-tests	
Group A & B question means.	0.1805
Group A question means & FDS	0.0038
Group B question means & FDS	0.0051
Collective A & B question means	0.0011
Collective A question means & FDS	0.0011
Collective B question means & FDS	0.0040

far more accurate than group B. Hence, these highly significant results support acceptance of both of our collective hypotheses H2 and H4.

5. Summary of Results and Implications

In order to reduce threats to validity, one group in this study consisted of experienced professionals, the other inexperienced post-graduates. In addition, the questions used offered a selection of single and multi-frame scenarios across all problem frames. Having derived a measure for the distance of an answer from an expert solution the study sought to test the following hypotheses.

H1: Group A (with greater experience) perform better at choosing problem frames than group B, when working individually.

H2: Group A (with greater experience) perform better at choosing problem frames than group B, when working collectively.

H3: Both groups (A & B) perform better than a 'Fair Distribution Strategy' (FDS) that evenly allocates all frames to each problem, when working individually.

H4: Both groups (A & B) perform better than a 'Fair Distribution Strategy' (FDS) that evenly allocates all frames to each problem, when working collectively.

Though we cannot support H1, our results being inconclusive, we believe that the study provides supporting evidence for H2, H3 and H4. Hence, though the results described are of limited statistical power, we now discuss the implications of our results, and note further patterns observed within the study.

- Based upon individual performance it is not clear that experienced subjects are significantly better at selecting relevant problem frames than inexperienced subjects (H1). Though we believed that experience would count our results here are inconclusive.

- Based on collective performance it does seem that experienced subjects are significantly better at selecting problem frames than inexperienced subjects (H2). When people work collectively it appears that experience counts. Group A showed a mean distance from the correct solution of only 0.04 when allowed to work in this way, and totally agreed with the expert view approximately for 9 of the 22 questions.

- Individuals can assign problem frames with a reasonable degree of accuracy. The lowest error rate we found was one person, from group A (experienced professionals), who had a mean magnitude of error across questions of 0.07. Compare this to the FDS (our baseline), which gives an equivalent magnitude of error of 0.32. Both groups performed better individually than the FDS (H3). However, there is some range in accuracy even within groups, the least accurate in group A had an error rate across questions of 0.27, and the group mean magnitude of error across questions was 0.198.

- All questions or scenarios are not equal, in that it is harder to correctly assign the relevant problem frames to some than others. The lowest question error rate for individuals was 0, that is, some individuals exactly agreed with the expert solution for some questions. Furthermore, when allowed to work collectively, group A frequently agreed with the expert solution. In contrast, the highest mean question errors were 0.51 (q14) for individuals in group A and 0.46 (q4) for individuals in group B, both much higher than the FDS error (0.32).

- Question difficulty does not appear uniform, in that different questions proved problematic to each group. Interestingly, the highest overall error (0.63) was for group B's collective answer to q14, a question where group A was totally correct. A similar lack of unanimity is observed when looking at individual performance within the groups. The only definite exception being that both groups had the least mean magnitude of question error for question 15, (0 for group A and 0.11 for B).

- A collective estimation or assignment process pays dividends. Indeed, one general result of this study is that it appears to vindicate the use of the collective estimation approaches.[3] Group A, with industrial experience, showed a vast improvement as the result of discussion and collective assignment of problem frames (moving from mean question error of 0.216, and mean subject error of 0.198, to a magnitude of error of only 0.04). Group B, with little or no industry experience, managed to perform as well on their collective assignments as their best individual. Both groups performed better than FDS (H4). Hence, we suggest that problem frames can be assigned with reasonable accuracy where groups work together to provide a collective solution.

6. Conclusions

This study sought to find out whether people, working both individually and in groups, can correctly identify the relevant problem frames for a given problem domain. If not, then the problem frame approach would be fundamentally flawed. Hence, we devised a test where subjects were given problem descriptions, and had to identify the relevant frames and their relative contribution to each problem. Euclidean distance was used to measure the magnitude of error, that is, how far subjects were from an agreed 'expert' solution.

We found that our subjects could choose problem frames with a reasonable degree of accuracy, and better than a baseline based on a Fair Distribution Strategy, but that they performed much better when a collective choice was made.

We also note some group differences, in that for our study software engineering experience seems to count, those with experience performing much better where a consensus approach is used. In addition, we found inconsistency between groups as to which questions were problematic. Hence, though we might suppose that some (typically multi-frame) problems are much harder to get right than others, without further study we cannot be certain about what kind of situations will prove most difficult.

Given the limited power of our study, we would be cautious in arguing for the utility of the problem frame approach. However, magnitudes of error seem acceptable, particularly if one allows individuals to work together in choosing relevant frames.

We now intend to conduct further studies, starting with larger groups, using the same question sets, and then examining the impact of different question sets. Of particular interest is what kinds of problems are most difficult to correctly assign and whether this is dependent on experience specific to certain domains. Since a collective approach appears promising we also intend to investigate the impact of anonymity, using a Delphi technique, as opposed to our approach of open discussion.

Notes

1. A more radical expert solution, increasing single frame scenarios would increase this error, but then it might also have a similar impact on the group scores.
2. An argument for why this is in effect a (corrected) magnitude of error was given in Section Three.
3. An anomaly (shown by Table 5) is that for questions where the expert answer is balanced across frames, the FDS performs well and conversely for questions where there is a single, or dominant, frame. Hence, for some questions, the FDS yields less error than the group.

References

Boehm, B. 1981. *Software Engineering Economics*. Prentice Hall.

Booch, G., Rumbaugh, J. and Jacobson, I. 1999. *The Unified Modeling Language User Guide*. Addison Wesley.

Bray, I. K. 2000. *Introduction to Requirements Engineering*. In draft form.

Budgen, D. 1994. *Software Design*. Addison-Wesley.

Coad, P. and Yourdon, E. 1991. *Object Oriented Analysis*. Prentice Hall.

Connolly, T., Begg, C. and Strachan, A. 1999. *Database Systems: A Practical Approach to Design, Implementation and Management—Second Edition*. Addison-Wesley.

Cox, K. and Phalp, K. 1999. Semantic and Structural Difficulties with the Unified Modeling Language Use-Case Notation version 1.3, OOPSLA workshop on Rigorous Modeling and Analysis of the UML: Challenges and Limitations. Denver, Colorado, USA, 2 November 1999.

CREWS. 2000. Co-operative Requirements Engineering with Scenarios. Homepage at: http://sunsite.informatik.rwth-aachen.de/CREWS/

Glass, R. 1998. *Software Runaways.* Prentice Hall.

Jacobson, I. 1994. Basic use-case modelling (continued). *Report on Object Analysis and Design*, volume 1.

Jackson, M. A. 1995. *Software Requirements and Specifications: A Lexicon of Practice, Principles and Prejudices*. Addison-Wesley.

Jackson, M. A. 1999. Re: Predicate logic. Personal communication.

Kovitz, B. L. 1995. *Practical Software Requirements: A Manual of Content and Style*. Manning.

Mair, C. and Shepperd, M. 1999. An investigation of rule induction based prediction systems. *ICSE '99 Workshop on Empirical Studies of Software Development and Evolution*. Los Angeles, USA, 18 May 1999.

Rolland, C., Souveyet, C. and Ben Achour, C. 1998. Guiding goal modelling using scenarios. *IEEE Transactions on Software Engineering: Special Issue on Scenario Management* 24(12).

Weidenhaupt, K., Pohl, K., Jarke, M. and Haumer, P. 1998. Scenarios in system development: Current practice. *IEEE Software* March/April: 34–45.

Yourdon, E. 1989. *Modern Structured Analysis*. Prentice Hall.

Keith Phalp graduated from the University of Kent with a BSc in Mathematics in 1986. Having taught mathematics, he returned to further study, gaining a Masters in Software Engineering (1992), and a PhD in Computer Science (1995), both from Bournemouth University. He is a Chartered Mathematician, and a Member of the Institute of Mathematics and its Applications (IMA). Dr Phalp's research concentrates on the Empirical Study of the Software Development process, and specifically upon the efficacy of requirements techniques. He currently teaches software implementation, software design and process modelling at Bournemouth University, where he is Course Leader for the Masters in Computing.

Karl Cox is a doctoral student at Bournemouth University, UK, where he is a member of the Empirical Software Engineering Research Group. He graduated with a History degree from the University of Wales Swansea in 1990. He then obtained a Masters with Distinction in Software Engineering from Bournemouth University in 1998. His research interests are in the uses of scenarios and use cases in requirements engineering.