Software Qual J
DOI 10.1007/s11219-007-9023-6

Improving the quality of use case descriptions: empirical assessment of writing guidelines

Keith Thomas Phalp · Jonathan Vincent · Karl Cox

Abstract Use cases are the main requirements vehicle of the UML and are used widely to specify systems. Hence, the need to write clear and accurate use case descriptions has a significant impact for many practitioners. However, many have pointed to weaknesses in the support offered to those writing use cases, and a number of authors advocate the use of rules in the composition and structuring of use case descriptions. These rules constrain the user, by only allowing certain grammatical constructions, typically guiding the structure or the style of the description For example, the CREWS research project pioneered Use Case Authoring Guidelines, suggesting that the adoption of such guidelines improved resulting use case descriptions. Replication of CREWS studies appeared to confirm the view that use case descriptions were improved through the application of guideline sets, but also noted that learning such rules presented a significant overhead. Hence, a leaner set of guidelines (the CP rules) was developed.

This paper describes empirical work to assess the utility of these two sets of writing guidelines (CREWS and CP). In particular, descriptions are assessed against a set of established criteria—a use case quality description checklist, which the authors described in a previous paper.

Our findings suggest that the leaner set of guidelines performs at least as well in terms of their ability to produce clear and accurate (comprehensible) descriptions. Hence, that a tractable set of rules may prove applicable to the industrial context, which could lead to effective validation of use cases.

K. T. Phalp (✉) · J. Vincent
Software Systems Modelling Group, Bournemouth University,
Poole House, Fern Barrow, Poole, Dorset BH12 5BB, UK
e-mail: kphalp@bmth.ac.uk

J. Vincent
e-mail: jvincent@bmth.ac.uk

K. Cox
Empirical Software Engineering Group, National ICT Australia,
Australian Technology Park Eveleigh, Sydney, NSW 1430, Australia
e-mail: karl.cox@nitca.com.au

1 Introduction

Use cases (Booch, Rumbaugh, & Jacobson, 1999) are now a well-established and popular method of specification (Graham, 1998; Kulak & Guiney, 2000). Indeed, the intuitiveness of the notation is typically cited as a major reason for their large-scale adoption (Jacobson, Christerson, Jonsson, & Overgaard, 1992). The freedom to write descriptions that are accessible to a breadth of audiences allows greater contribution from a variety of stakeholders, which can significantly improve the effectiveness of validation (Maiden & Corral, 2000; Sutcliffe, 1998).

Despite such widespread usage, there are still many suggestions that the application of use cases (Jackson, 2001), and particularly use case descriptions, is problematic (Phalp & Cox, 2001). Reservations about their utility tend to fall into one of two categories. There are those who consider that the notation itself does not provide enough power or expressiveness to describe the nuances of specification (Ratcliffe & Budgen, 2005). Typically, these authors suggest that the notation should either be augmented (such as with pre and post conditions) (Phalp & Cox, 2003a; Some, 2006) or supplemented (with other diagram types) (Ratcliffe & Budgen, 2005). Similarly, there are those who suggest that lack of prescription in the application of the use case is the problem (Alexander & Stevens, 2002; Cockburn, 2001). That the very freedom and expressiveness allowed by, what is, in essence, a structured form of natural language, leads to problems in structure and comprehension (Alexander, 2003; Anda, Sjoberg, & Jorgensen, 2001; Ben Achour, Rolland, Maiden, & Souveyet, 1999). The authors feel that both categories have valid arguments. Augmentations to the original use case description may help to solve specific issues, for example, addition of pre and post conditions can highlight dependencies amongst events (Kanyaru & Phalp, 2005; Phalp & Cox, 2003b). However, a focus on improving the quality of 'standard' use case descriptions may be more in keeping with the ethos of providing an accessible notation for the requirements phases (Adolph, Bramble, Cockburn, & Pols, 2003; Cockburn, 2001). Further, that in order to be broadly applicable, such guidelines should not be onerous, but should be as simple as possible, whilst still providing tangible benefits (Cockburn, 2001).

Hence, this paper describes a set of simple use case description guidelines (the CP rules) (Cox, 2002) and examines the impact of writing guidelines on improving the quality of use case descriptions (Cox, Phalp, & Shepperd, 2001). In doing so, we compare these rules to another (larger set) developed by CREWS, which have already been shown to provide improvements in use case quality (Cox & Phalp, 2000). Therefore, our goal was not to provide more powerful guidance, but rather to attempt to take the principal factors that had a positive impact on use case quality (Cox, Aurum, & Jeffery, 2004) and to distil these into a smaller set of rules, suitable for practical application.

The remainder of this paper is structured as follows. Section 2 reviews previous work on description guidelines, and the rationale for our studies. Section 3 describes our approach to measuring the utility of structure guidelines, and how to measure the extent to which those guidelines are applied. Section 4 provides empirical results for the structure guidelines. Section 5 considers how to gauge the quality (or communicability) of descriptions, which requires the introduction of a qualitative assessment framework. Section 6 provides an overview of the results on communicability, discusses our findings overall, and ties together rule usage and communicability. Section 7 considers the validity of our work and issues of experimental design, and Sect. 8 concludes.

2 Use case guidelines and studies

Many authors have suggested using guidelines for use case descriptions (e.g. Ben Achour et al., 1999; Cockburn, 2001; Cox, 2002; Rolland & Ben Achour, 1998) and such guidance is often entirely plausible. For example, Cockburn's (2001) recommendation of 'subject... verb... direct object... prepositional phrase', appears to be particularly straightforward and intuitive. Similarly, Graham (1998) suggests structure for task events, and Alexander and Stevens (2002) suggest that requirements statements should also be similarly straightforward. Although these structure guidelines are meant to aid composition, the ultimate goal is to improve the resulting description. Indeed, by examination of descriptions, it should be possible to gauge the effectiveness of guidance, even if this is only to measure compliance with the given suggestions.

Despite this interest, few groups have conducted systematic evaluations of use case writing guidelines (exceptions include CREWS (e.g. Ben Achour et al., 1999; Rolland & Ben Achour, 1998), ESERG (e.g. Cox & Phalp, 2000; Cox et al., 2001), Simula (e.g. Anda & Sjoberg, 2005; Anda et al., 2001) and NICTA (Cox et al., 2004)). This lack of evaluation is somewhat surprising considering the impact that the use case has had over the last few years, and, given their prevalence, the potential cost benefits of improvements (Hofmann & Lehner, 2001). Furthermore, to some extent, previous evaluations are self-fulfilling, in that if one suggests that a particular structure is advantageous it is perhaps not surprising if one finds the evidence for its adoption in empirical study. Therefore, one might, as suggested within this paper, require guidelines to produce both adherence and some other measurable quality improvements. Hence, we now briefly review the empirical work conducted to date, in order to provide a context and rationale for the experiments described within this paper.

2.1 Previous empirical studies

Ben Achour et al. (1999) describe an experiment to investigate the effectiveness of their own writing guidelines. Their results suggest that use of the CREWS Guidelines produces more complete and better-structured descriptions than those produced without guidelines. However, they note that there were too many guidelines within their set, and that the guidelines should, perhaps, come with examples of their application.

In replication, Cox and Phalp (2000) suggested that the CREWS Content Guidelines seemed useful in aiding completeness in descriptions, whilst the Style Guidelines appeared to enhance structure. However, it was also noted (Cox et al., 2001) that many of the Content Guidelines were rarely used, such as CG1, 2, 3, and 8. This suggested that a reduced set of Content Guidelines might be equally effective. Similarly, Anda et al. (2001) described a variation on this single use case description experiment, where they explored the understandability of the wider use case model (diagram as well as descriptions). They provided three sets of guidelines: minimal (general advice on writing use cases), template (a framework for the use case model) and style (a slightly reduced version of the CREWS Guidelines based on the findings of Cox, 2002). Descriptions were evaluated in terms of understandability, based on a suggested marking scheme (Cox & Phalp, 2000); which considered the overall comprehension of the description (in terms of plausibility, readability, consistent structure and consideration of alternative flows). Anda et al. (2001) showed that style guidelines were important (for the descriptions) in terms of level of detail (no internal design nor user interface details were allowed), realism (a logical and complete

sequence) and consistency (correct use of terminology). The overriding result was that the style and template guidelines were significantly better than the minimal guidelines, for constructing, and subsequently comprehending, use cases.

The experiments described above provide a platform for a continued exploration of how guidelines improve the writing of use case descriptions. The overriding impact of the few experiments outlined was that writing guidelines do seem to make a difference to the quality of use case descriptions. They seem to make them more complete, consistent and understandable.

This paper thus explores this claim further by building upon the work already conducted. The CREWS Use Case Authoring Guidelines are the starting point for bringing more structure to use case descriptions. Although there are some potential problems with them, as identified in (Ben Achour et al., 1999; Cox & Phalp, 2000), it is worth noting that, since the CREWS Guidelines are established and well regarded, the set is taken as a baseline for experimentation in this paper. Since understandability is an important aim of the use case description, work is required to develop writing guidelines that help construct a more understandable use case description.

Whereas previous studies have tended to simply measure adherence to their suggested guidelines as an indication of quality, this paper also attempts to measure quality by reference to a comprehensive evaluation scheme derived from the qualities of good use case descriptions. This reference scheme, which we presented in Phalp, Vincent, and Cox (2006), provides an effective (and independent) measure, allowing us to examine the impact of rule usage on the quality of the resulting description.

3 Use case structure guidelines

If one assumes that the structures recommended by guideline sets are desirable then the effectiveness of guideline sets can be measured simply by ascertaining which set of rules performs best in producing such structures (though an independent measure is introduced in a later section). Since guidelines differ, and even use different terminology, we first need to find which guidelines are directly comparable from both sets of rules. The complete set of CP rules can be found in Cox (2002). However, for our purposes we briefly describe those rules that relate directly to structure. It is also worth noting that whilst we use the term structure to refer to those rules which attempt to impose structure upon the description, the CREWS group denote these as Content Guidelines.

One reason for restricting this treatment to structure guidelines is that earlier studies (Cox, 2002; Cox et al., 2001) had shown that these structural guidelines had the most impact upon use case quality. In addition, whilst both groups (CP & CREWS) provide style guidelines, previous studies found no significant difference in their usage (Cox et al., 2001; Phalp & Cox, 2002), or impact. Hence, this study attempts first to judge the ease of usage of such structure guidelines and (by reference to an assessment framework) consider independently their impact on quality.

3.1 The guideline sets

The following sections outline the guidelines considered in this study.

3.1.1 CP structure rules

The CP Structure Rules have only three structures, less than half the number of the comparable CREWS Content Guidelines ('verb' refers to present simple tense).

Structure 1: Subject verb object.
For example: *The operator presses the button.*
Structure 2: Subject verb object prepositional phrase.
For example: *The system reminds the operator to save all the open files.*
Structure 3: Underline other use case names.
For example: *The user makes a new equipment request.*

3.1.2 CREWS content guidelines

CG1: <agent> <'move' action> <object> from <source> to <destination>.
For example: *The clerk sends the report from the store to the office.*
CG2: <source agent> <'put' action> <object> to <destination agent>.
For example: *The clerk gives the report to the manager.*
CG3: <destination agent> <'takes' action> <object> from <source agent>.
For example: *The manager gets the report from the clerk.*
CG4: <agent> <action> <agent>.
For example: *The clerk informs the customer.*
CG5: <agent> <action> <object>.
For example: *The operator presses the button.*
CG6: 'If' <alternative assumption> 'then' <action>.
For example: *'If' the record is blank 'then' search for customer ID number.*
CG7: 'Loop' <repetition condition> 'do' <action>.
For example:
1. *'Loop' records available*
2. *'Do' fill in records.*

CG8: <action 1> 'meanwhile' <action 2>.
For example: *Enter consultation notes 'meanwhile' search for X-ray record.*

3.1.3 Comparing structure guidelines

An examination of the rules reveals that only two sets of structure guidelines appear to be candidates for direct comparison. That is, CP1 can be compared against CG5 and CP2 against all of CG1–3. Fortunately, these are both rules that previous studies have shown to be prevalent in subsequent descriptions (Cox & Phalp, 2001; Cox et al., 2001), and, hence, structures resulting from such rules should manifest themselves in sufficient numbers to warrant an empirical treatment. The following section outlines a study to find the extent to which these particular structure guidelines are manifested in resulting use case descriptions.

4 Measuring the impact of structure guidelines

The CP rules were intended to be smaller and easier to apply than the CREW guidelines. Hence, this goal implied that ease of use would be shown by more of these rules being seen in use case descriptions than for CREWS. Therefore, our hypothesis is:

H1 The constructs suggested by the CP rules are found in significantly higher numbers than the equivalent CREWS guideline constructs when both guideline sets are applied to the same problems.

4.1 Background to the experiment

Both measures (guideline usage and communicability, Sect. 5) were taken from a single experiment using 60 subjects from a software design unit shared by students in Computing and Software Engineering. These were split into four experimental groups, each of 15 students, of comparable ability (Table 1). All groups had been taught about standard use case descriptions in previous classes, and were introduced to the appropriate guideline set under experimental conditions.

One of the problems of using students is that they may not be considered a representative reflection of the general software engineering population. However, their behaviour is more akin to software engineering professionals than the general population (Adolph et al., 2003). For example, Höst, Regnell, and Wohlin (2000) show that there is no significant difference between students and practitioners in performing lead-time impact assessment. Furthermore, even if absolute scores from students would be different to those from practitioners the general trend in applicability of guideline sets may still be gauged from such studies, and thus the goal of comparing guidelines still valid.

Another possible objection is that the tasks given were somewhat artificial, standard, examples, and were chosen to be achievable within a single session (the students had 1 h to read guidelines, read the tasks descriptions and produce the use case descriptions). However, these tasks also replicated those carried out both by earlier CREWS studies, and our own pilot studies, and this allows for a richer analysis of a body of evidence (Pickard, Kitchenham, & Jones, 1998) in the longer term.

4.2 Findings for rule usage

Although hypothesis H1 is a general statement about guideline usage our analysis of structure guidelines shows that there are just two possible sets which merit direct comparison. That is, CP1 can be compared against CG5, and CP2 against all of CG1–3. Since our hypothesis was intended to indicate whether CP rules perform better than CREWS we took the step of performing a single tailed test. We note that there was a risk with a single tailed test, such that if CREWS had performed significantly better than CP then this could not be stated as a significant finding and the null hypothesis would still have to be accepted (Miller, Daly, Wood, Roper, & Brooks, 1997). However, it is, subsequently, clear from our findings that CP always performed as well or better, and hence the additional power of the single tailed test is justified.

Table 1 Experimental groups, guidelines, and tasks

Group	Guidelines	Use case task
A	CP rules	ATM
B	CP rules	Retail
C	CREWS guidelines	ATM
D	CREWS guidelines	Retail

4.2.1 CP1 versus CG5

Table 2 summarises the results for CP1 versus CG5.

Since we are simply examining the counts of usage it is appropriate to test the significance of the differences in means for each group pairing. The ordering of students is simply a convenient representation and not a reflection of the data. We test for significance with a simple *t*-test, with $\alpha = 0.05$. In brief, we find that there is no significant difference between groups A and C, but that there is a significant difference between groups B and D. That is, group B used the subject verb object construction (CP rule 1) significantly more than group D used the equivalent CG5.

4.2.2 CP2 versus CG1–3

Table 3 summarises the results across the four groups for CP2 versus CG1–3. In this case there is significant difference between both CP groups versus both CREWS groups, that is, AB versus CD. However, given that there is no significant difference between groups A and C; this can be explained by the effect of the differences in usage between groups B and D.

4.3 CP structure usage versus CREWS usage

Overall, we find significant differences between one set of groups (the retail scenario) and no significant difference between the other (ATM). That is, for both CP1 and CP2 there are more instances of usage than the equivalent CREWS rules for one of the given scenarios. Specifically, CP was found to outperform CREWS for one of the groupings, but that for the other problem there was no significant difference.

Table 2 CP structure results for CPI versus CGS

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	33	33	0	4	29	50	47	16	83	61	20	33	62	70	31
C	35	25	48	10	65	50	45	33	69	62	18	46	79	20	33
B	6	33	45	30	6	57	24	14	22	29	36	30	42	12	22
D	4	0	23	0	25	38	35	5	15	0	18	45	18	0	13
$\alpha = 0.05$	A, C: $p = 0.30$					B, D: $p = 0.02$					AB, CD: $p = 0.27$				

Table 3 CP structure 2 versus CREWS equivalent

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0	0	0	0	0	33	5	37	0	0	0	5	0	23	8
C	8	13	5	3	4	23	0	7	6	15	0	4	0	0	0
B	50	33	36	17	22	29	29	14	33	14	18	40	16	29	17
D	21	15	38	23	15	0	6	35	16	25	18	0	0	0	0
$\alpha = 0.05$	A, C:$p = 0.34$					B, D: $p = 0.004$					AB, CD: $p = 0.02$				

A positive interpretation of these results in isolation might be that the leaner CP rules perform as well or better than CREWS, in producing the desired structure constructs within the resulting use case description. In some ways these results are at odds with the pilot study (Cox et al., 2001), which reported that CP fared better with the ATM problem and that results with retail were not significantly different. Again, if we take a positive view, we might suggest that CP appears (over both studies) to fare as well or better in producing the desired use case description structure. However, such variation in the data might also suggest that the effects are relatively small, and that variation could be noise in the data. This danger in coming to conclusions across studies is one reason that Pickard et al. (1998) suggest only combining results where experiments are truly homogenous.

Despite such reservations it is clear that the smaller CP rules perform at least as well, and possibly better, in guiding the structure of use case descriptions. However, so far, such a result only implies an improvement in quality. In order to be able to have a more independent assessment of use case quality, and thus the real impact of guidelines, it is necessary to gauge the descriptions against a set of qualitative criteria.

5 Use case communicability

It is relatively easy to perform quantitative analyses to examine the extent of usage of suggested structural constructs. To attempt to measure the communicability of the use case requires either that: the descriptions are graded by some other independent means, or a further 'comprehension' experiment is carried out to test the ability of subjects to extract information from the descriptions.

Comprehension experiments to judge use cases have been carried out by the authors (Phalp & Cox, 2003a). However, this approach also includes pitfalls for the unwary. For example, further subjectivity is introduced in producing question sets, and one must also be careful about what kind of information is required, e.g. design versus specification. Hence, this paper will concentrate on the former approach, that is, to judge the quality of use cases against agreed criteria.

In order to judge use case quality we adopt a set of quality factors, or use case facets, referred to as the '7Cs of communicability'. These facets are derived primarily from discourse process research, and from other research in use case description. A full description of the factors and their derivation can be found in (Phalp et al., 2006), and hence is omitted here. Indeed, it is important to realise that the use of this approach is simply to give an independent marking criteria, and is typical of the approach that one would apply in grading student work. However, in reality the factors will not be totally independent of any rule sets, since both would typically be founded on the same underlying theories of text comprehension.

6 Measuring communicability

Whereas hypothesis H1 focused on the number of structure constructs found in the descriptions, we now attempt to gauge the proposition:

Use case descriptions produced with the CP rules are significantly more comprehensible than the equivalent CREWS use case descriptions.

Since we cannot assess this quality directly we adopt the 7Cs, as 'independent' marking criteria. Hence, hypothesis 2 becomes:

H2 Use case descriptions produced with the CP rules score significantly better than the equivalent CREWS use case descriptions, when marked against the 7Cs use case quality facets.

6.1 Findings for communicability

Despite having the 7Cs as a given marking criteria, the experimenters were aware that marking would still be subjective to some extent. Indeed, this is the every day experience of academics. Hence, as is normal practice, two experienced markers blind marked and then averaged marks for the descriptions. Each student gained a final mark, out of a possible 100, and we then compared the performance of students in different groups.

However we also wanted, partly by way of confirmation, to analyse the significance of individual facets, when marks were averaged across a group of students. Hence, Sect. 6.3 describes this second form of analysis, which aims to understand, and confirm, which facets are being influenced by the use of guidelines.

6.2 Analysis of marks for communicability

6.2.1 Groups A and C

Figure 1 illustrates the marks (see Table 4) for groups A and C (tail calculation is taken from Fenton & Pfleeger, 1996; Wohlin et al., 2000). There appear to be two outliers in group A (A4, scoring 35 and A12 scoring 90) and three in group C (C11 (22), C4 (29) and C1 (36)), and some overlap in the scores.

A paired (single tailed) *t*-test, gives $p = 0.096$, suggesting only weak significance, and it would be unwise to suggest that there is any substantial difference between the performance of groups A and C based upon these results.

6.2.2 Groups B and D

Figure 2 illustrates the marks (see Table 4) for groups B and D. About 12 out of 15 of group B scored higher marks than group D. In this case, a paired *t*-test (single tailed)

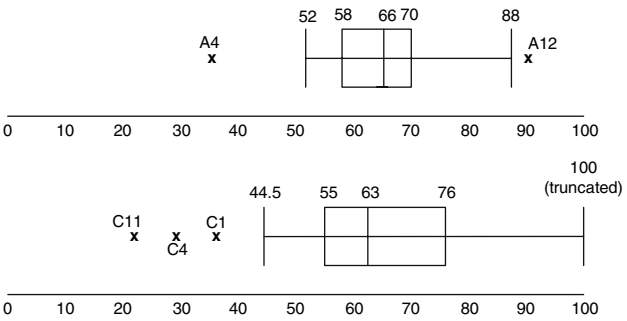

Fig. 1 Communicability marks: A and C

Table 4 Marks for communicability (A, B, C & D)

Group A	35	53	56	58	60	62	64	66	68	69	69	70	73	73	90
Group C	22	29	36	55	57	61	62	63	69	70	75	76	77	82	82
Group B	32	56	62	65	67	69	71	72	76	78	79	87	89	92	95
Group D	44	58	60	61	61	65	69	70	71	71	76	76	76	82	97
Means	Group A: 64.4, Group C: 61.07							Group B: 72.67, Group D: 69.13							
Standard deviation	Group A: 12, Group C: 18.77							Group B: 15.99, Group D: 12.19							

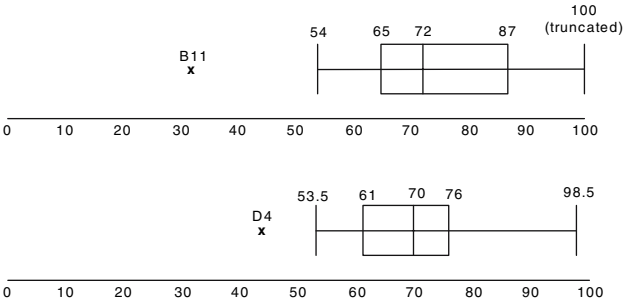

Fig. 2 Communicability marks: B and D

reveals a highly significant difference between the scores for each group ($p = 0.0001$). This result was also confirmed by analysis of rankings (omitted here for brevity), which is again highly significant.

6.3 Analysis of marks for each facet across groups

In addition to the examination of the student scores across groups, a further analysis was conducted to ascertain whether there were significant differences among facets (taking average scores for students for each group).

Since we did not require a total weighted score, we simply show marks out of 10 for each of the 7 facets. Table 5 shows the mean grades for all four groups. Given that we have means across subjects, and that there was still the potential for subjectivity, a non-parametric approach was deemed appropriate and Mann–Whitney *U* tests were conducted on the data.

6.3.1 Groups A and C

Table 6 compares groups A and C. Overall there is no significant difference between groups A and C ($p = 0.65$). The only significant differences among the facets are for Consistent Structure, where Group A's structure is significantly better than group C's ($p = 0.008$).

Table 5 Scores for communicability facets

7C	A	C	B	D
Coverage	8.4	8.3	9	8.8
Cogent	7	6.9	9.1	8.5
Coherent	7.3	8.3	8.5	8.5
Consistent abstraction	7.9	8.6	7.1	8.2
Consistent structure	8.9	7	8.3	8.1
Consistent grammar	5.3	4.5	5.4	5.3
Consideration of alternatives	2.7	3.3	4	3.1
Totals (out of 70)	47.3	46.9	51.4	50.5

Table 6 Communicability A versus C

7C	p (=<0.05)
Coverage	0.42
Cogent	0.44
Coherent	0.76
Consistent abstraction	0.51
Consistent structure	0.008
Consistent grammar	0.25
Consideration of alternatives	0.66
Total	0.65

Given that the only significant difference is in Consistent Structure it seems appropriate to examine this facet further. Raw scores show that the majority of group A score 8 or more (13 out of 15), whereas just under half of C do so (7 out of 15). Hence, examination of the data confirms that group A only performed better than group C in terms of consistent structure, and shows general agreement with the analysis of marks, where there was little difference between groups A and C.

6.3.2 Groups B and D

Table 7 compares groups B and D. The significant difference among facets is the score for Cogent ($p = 0.04$). This score can be explained by the observation that group D had 12 Text Order problems, with 9 out of 15 descriptions having at least one order problem, whereas Group B had only five ordering problems, committed by just three subjects.

However, it is interesting to note that this 'Cogent' factor contributes more to scores than any other, and, hence, this significant difference accounts for the highly significant differences reported in the analysis of student scores.

6.4 Findings for communicability overall

In considering communicability, we attempted to find whether use case descriptions produced with the CP rules scored significantly better than the equivalent CREWS use case

Table 7 Communicability B versus D

7C	p (=<0.05)
Coverage	0.18
Cogent	0.04
Coherent	0.62
Consistent abstraction	0.87
Consistent structure	0.33
Consistent grammar	0.49
Consideration of alternatives	0.35
Total	0.31

descriptions, when marked against the 7Cs use case quality facets. As with H1, for H2 we have mixed results. For groups A and C, the CP rules appear to perform slightly better—though not significantly so. However, we do find a highly significant difference in the performance of groups B and D, in favour of the CP rules.

A further statistical analysis, of the differences among facets, explained these results, since it showed a significant difference for Consistent Structure between A and C, but no other significant differences. For groups B and D, the 'Cogent' facet showed significant difference. Given that this facet contributes more strongly to scores, this explains the highly significant difference in the marks between groups B and D.

Once more, a positive interpretation of results would be that the CP rules perform as well, or better, than the CREWS guidelines. However, more importantly, the results for H1 and H2 also suggest that the independent assessment confirms the view that guideline usage and overall quality are related. That is, for marks at least, where CP rules led to an increase in the structures found within the use case (H1), the quality assessment also confirmed that these appear to be better descriptions (H2). Similarly, we find no significant improvement in communicability where the rules are applied no better.

Perhaps most gratifying is to note that in all cases (both CREWS and CP) increased usage (of both rule sets) does appear to improve communicability. In other words, that for H1 and H2 we have agreement. That is, when more rules are applied the resulting use case descriptions are improved. Equally, where H1 revealed no significant differences between the structures found in descriptions groups, H2 also found no significant differences in quality.

7 Validity

This section discusses the perceived validity threats to this experiment only. It is not intended as an exhaustive treatment (see Wohlin et al., 2000), rather as an illustration of how such threats were considered within the design and conduct of the study.

7.1 Conclusion validity

Conclusion validity considers whether the conclusions drawn from the statistical results are valid or whether they might be biased by issues affecting the treatment and the outcome. For this experiment we have a random heterogeneity of subjects, in that all subjects were

undergraduate students on a computing course, where there is no streaming by ability. As might be expected, some subjects performed better or worse than others. Four subjects from group C scored low marks (below 40 marks). A post-hoc covariance analysis of student ability on their degree course to compare against the results here was considered to see whether the subjects performed unusually badly in the experiment; however, this was not possible because subject names were not recorded, primarily to avoid bias in marking.

7.2 Internal validity

Internal validity is threatened by unknown influences on the causal relationship between treatment and outcome. If these are not accounted for, they can invalidate the results.

7.2.1 History

In separate sessions that spanned a week after the introduction to use cases (due to timetabling restrictions), the subjects completed the experimental tasks. There is a risk that subjects, having participated in the experiment, would pass on any knowledge to those yet to take the experiment (causing diffusion of treatments). The only way to control this was to make sure no experimental material was taken from the location of the experiment. There is no indication of an overall significant improvement in the results from group A to B and from group C to D except with regards to the CP Structures and CREWS Content Guidelines, which must in part be due to the nature of the different tasks. CP Structure 1 is used more in group A than B. CREWS CG5 is used more in group C than D. The opposite is true for CP Structure 2 and CREWS CG1–3, indicating that the Retail task has more complex interactions than the ATM.

7.2.2 Maturation

The experiment lasted an hour (the writing part 45 min). Concerns over boredom or over-enthusiasm were not considered significant. As such, no subjects dropped out of the experiment. In terms of the experiment itself, 1 h is not much time to learn a set of guidelines and write a use case description and in the wider industrial context would not be particularly realistic. However, due to timetabling pressures it was impossible to obtain more time with the subjects. It is unknown whether any subjects studied use case descriptions between the lecture and the experiment itself. This is a factor that could not be controlled. The subjects were not informed of CREWS or CP throughout the duration of the experiment.

7.3 Construct validity

This considers inadequate preoperational explication of constructs (Wohlin et al., 2000). The measures of structure for H1 are simply counts and are clearly defined. For H2, the hypothesis measures relate to a published set of facets, which are themselves based upon established discourse process theory (coherence), software engineering (scope, span) and grammar. Furthermore, these results (H2) are explained by an analysis of the significance of each facet, which appears to confirm the findings. Finally, the clear agreement of H1 and H2 gives credence to the validity of the measures used.

7.4 External validity

It is clear that the results of the experiment cannot be generalised to every development employing use case descriptions. However, they might be considered representative of undergraduate computing and software engineering students. One cannot generalise to software houses because of the time constraints on the experiment and, importantly, the prior knowledge brought to the task by experienced practitioners.

7.4.1 Nature of the problem (Höst et al., 2005)

The tasks are similar to those of the pilot study, though slightly shortened because of reduced available time with the subjects. It is also the case that CREWS have used the ATM example to explain their guidelines (Cox, 2002) and the Retail task in experimentation (Jackson, 2001). The ATM is a ubiquitous software engineering problem, e.g. (Cox & Phalp, 2000; Cox et al., 2004); it is therefore reasonable to use because of this apparent general acceptance through example in the literature. Both tasks should also be familiar to the subjects.

7.4.2 Setting (Robson, 1993)

Use case descriptions should really be determined with the aid of stakeholders. It has been suggested that requirements engineers, systems analysts, etc., tend to write descriptions without the direct involvement of the clients and that this is potentially risky (Ying, 2001). As shown in Hofman and Lehner (2001), the most successful projects have close customer contact throughout the whole of the requirements process and that such customer input is vital to the success of projects. In this way the experiment is artificial. However, given all relevant facts and process (from elicitation already conducted), one would expect the analyst to write the first description and then validate this with clients. As such, although the experiment can be considered artificial for contextual reasons, the goal of the task is to write a description based on the information provided and the subject's pre-existing knowledge of the ATM or Retail domains.

8 Conclusions

A number of previous studies have suggested that use case guidelines can improve the quality of use case descriptions. However, some studies have also revealed that, though effective, application of existing guideline sets is problematic. Hence, the authors derived a small set of guidelines, the CP rules, which were intended as a possible replacement for the existing approaches.

This paper describes an experiment to compare these CP rules with the (ambitious baseline) of the already proven CREWS Use Case Authoring Guidelines. Given that both sets of guidelines suggest similar structure rules our experiment first attempted (H1) to ascertain whether the CP rules would lead to more desirable structures than the CREWS guidelines. The two guideline sets were used to write use case descriptions for two different scenarios (requiring four experimental groups). In brief, our findings were that the smaller CP rules did produce a significantly greater number of such structures for one scenario (retail), but that the effect for the other scenario was not significant.

We then assessed the use case descriptions against a set of quality criteria (H2). This analysis allowed us to discover whether the incidence of suggested use case structures also tallied with better marks for quality. Again we found a significant difference in quality (CP over CREWS) for the retail scenario, but not for the other. That is, where we had noted significant differences in the number of structures (application of guidelines) use case quality was also significantly different (in both cases showing an improvement with CP rules). Similarly, where there was no significant difference in rule usage, there was no significant difference in quality (as assessed against criteria). This, in itself, is an important result. Not only can we confirm that where suggested structures are found, use case quality is improved (as previous studies have shown), but also this study suggests that even differences in the number of times such structures are found may account for differences in the quality of the use case descriptions.

In terms of our original intention, however, we find that, despite the effects described, there may be little difference in the performance of the CP Rules and the CREWS guidelines overall. In some senses this further vindicates the authors contention for a simpler rule set. Furthermore, since the CREWS guidelines are (as stated earlier) already well regarded, this is a satisfactory achievement.

Finally, it is worth remembering that the experiment further demonstrates that application of rules (measured by usage) results in better descriptions (in terms of communicability). Hence, we contend that many users could benefit from the adoption of a minimal (CP) set of rules to improve the quality of use case descriptions.

References

Adolph, S., Bramble, P., Cockburn, A., & Pols, A. (2003). *Patterns for effective use cases*. Addison Wesley.

Alexander, I., & Stevens, R. (2002). *Writing better requirements*. Harlow: Addison-Wesley.

Alexander, I. (2003). Misuse cases: Use cases with hostile intent. *IEEE Software*, Jan/Feb, 58–66.

Anda, B., Sjoberg, D., & Jorgensen, M. (2001). Quality and understanding of use case models. In J. Lindskov Knudsen (Ed.), *15th European conference on object-oriented programming*, LNCS, Springer, Budapest, June 2001, pp. 402–428.

Anda, B., & Sjoberg, D. (2005). Investigating the role of use cases in the construction of class diagrams. *Empirical Software Engineering Journal, 10*(3), 85–309, July 2005.

Ben Achour, C., Rolland, C., Maiden, N., & Souveyet, C. (1999). Guiding use case authoring: Results of an empirical study. In *4th IEEE international symposium on requirements engineering, RE'99*, Limerick, Ireland, June 1999, pp. 36–43.

Booch, G., Rumbaugh, J., & Jacobson, I. (1999). *The UML user guide*. Harlow: Addison-Wesley.

Cockburn, A. (2001). *Writing effective use cases*. Harlow: Addison-Wesley.

Cox, K., & Phalp, K. (2000). Replicating the CREWS use case authoring guidelines experiment. *Empirical Software Engineering Journal, 5*(3), 245–268.

Cox, K., Phalp, K., & Shepperd, M. (2001). Comparing use case writing guidelines. In *7th international workshop on requirements engineering: Foundation for software quality, REFSQ'01*, Interlaken, Switzerland, June 2001, pp. 101–112.

Cox, K. (2002). Heuristics for use case descriptions. PhD Thesis, Bournemouth University, UK.

Cox, K., Aurum, A., & Jeffery, R. (2004). An experiment in inspecting the quality of use case descriptions. *Journal of Research and Practice in Information Technology, 36*(4), 211–229.

Fenton, N., & Pfleeger, L. (1996). *Software metrics—A rigorous and practical approach* (2nd ed.). Thomson Computer Press.

Graham, I. (1998). *Requirements engineering and rapid development*. Harlow: Addison-Wesley.

Hofmann, H., & Lehner, F. (2001). Requirements engineering as a success factor in software projects. *IEEE Software*, July and August issue, 58–66.

Höst, M., Regnell, B., & Wohlin, C. (2000). Using students as subjects—A comparative study of students and professionals in lead-time impact assessment. *Empirical Software Engineering Journal, 5*(3), 201–214.

Jackson, M. (2001). *Problem frames*. Harlow: Addison-Wesley.

Jacobson, I., Christerson, M., Jonsson, P., & Overgaard, G. (1992). *Object-oriented software engineering: A use case driven approach*. Wokingham: Addison-Wesley.
Kanyaru, J., & Phalp, K. (2005). Supporting the consideration of dependencies in use case specifications. In *11th International workshop on requirements engineering: Foundation for software quality—REFSQ'05*, Porto, Portugal, 13–14 June 2005.
Kulak, D., & Guiney, E. (2000). *Use cases: Requirements in context*. Addison-Wesley.
Maiden, N., & Corrall, D. (2000). *Scenario-driven systems engineering*. London: IEE Seminar.
Miller, J., Daly, J., Wood, M., Roper, M., & Brooks, A. (1997). Statistical power and its subcomponents—Missing and misunderstood concepts in empirical software engineering. *Information and Software Technology, 39*, 285–295.
Phalp, K., & Cox, K. (2001). Guiding use case driven requirements elicitation and analysis. In Y. Wang, S. Patel, & R. Johnston (Eds.), *7th International conference on object-oriented information systems*, OOIS'01, LNCS, Springer, Calgary, Canada, August 2001, pp. 329–332.
Phalp, K., & Cox, K. (2002). Supporting communicability with use case guidelines: An empirical study. In *6th International conference on empirical assessment in software engineering*, Keele University, 8–10 April 2002.
Phalp, K., & Cox, K. (2003a). Exploiting use case descriptions for specification and design. In *7th International conference on empirical assessment and evaluation in software engineering*, Keele University, Staffordshire, UK, 8–10 April.
Phalp, K., & Cox, K. (2003b). Using enactable models to enhance use case descriptions. In *Proceedings of the ProSim'03, International workshop on software process simulation modelling (in conjunction with ICSE)*, Portland, USA.
Phalp, K., Vincent, J., & Cox, K. (2006). Assessing the quality of use case descriptions. *Software Quality Journal, 15*(1), 69–97, March 2007.
Pickard, L., Kitchenham, B., & Jones, P. (1998). Combining empirical results in software engineering. *Information and Software Technology, 40*, 811–821.
Ratcliffe, M., & Budgen, D. (2005). The application of use cases in system analysis and design specification. *Information and Software Technology, 47*(9).
Robson, C. (1993). *Real world research*. Oxford: Blackwell.
Rolland, C., & Ben Achour, C. (1998). Guiding the construction of textual use case specifications. *Data and Knowledge Engineering Journal, 25*(1–2), 125–160.
Some, S. (2006). Supporting use case based requirements engineering. *Information and Software Technology, 48*(1).
Sutcliffe, A. (1998). Scenario-based requirements analysis. *Requirements Engineering Journal, 3*, 48–65.
Wohlin, C., Runeson, P., Höst, M., Ohlsson, M., Regnell, B., & Wesslen, A. (2000). *Experimentation in software engineering: An introduction*. Kluwer Academic.
Ying, L. (2001). Actor-led object modelling for requirements and systems analysis. In Y. Wang, S. Patel & R. Johnston (Eds.), *7th International conference on object-oriented information systems*, LNCS, Springer, Calgary, August 2001, pp. 37–46.

Author Biographies

Keith Phalp originally read for a first degree in Mathematics, which he then taught for a few years, before completing a Masters in Software Engineering in 1991, followed by Ph.D. in process Modelling in 1994. He then spent three years as a post-doctoral research fellow at the University of Southampton, again in the area of process modelling. In 1997, Dr. Phalp took up a lectureship at Bournemouth, and became course leader for the Masters in Software Engineering. He has been there ever since, and during that time has taught units covering the majority of the software development process. He is currently Reader in Software Engineering within the Software Systems Modelling Group at Bournemouth University, which he co-founded with Dr Jonathan Vincent. His research focuses on requirements engineering, and its relationship to both business needs and to the specification and design of systems.

Jonathan Vincent has a B.Eng. in Electrical and Electronic Engineering, an M.Sc. in Computing (Software Engineering) and a Ph.D. in Computer Science. He is currently a Reader in Natural Computing within the School of Design, Engineering and Computing, at Bournemouth University, UK, where he directs the Software Systems Modelling Group (www.sosym.co.uk). He has wide ranging research interests and has published in a variety of areas within software engineering and computer science, including component based software engineering, software quality, modelling, evolutionary computation and neural networks.

Karl Cox has a Masters degree in Software Engineering and a Ph.D. in Computer Science, both from Bournemouth University. Dr. Cox's research interests are centred on requirements engineering, specifically: the Problem Frames approach as a means of providing a framework for understanding the problem context of business needs; goal modelling, combined with problem frames, as a means of describing business goals, strategies, and objectives that are aligned to software; process modelling, which captures the details of processes that businesses implement to carry out their daily work; and use cases, which is concerned with ways to improve the comprehensibility of use case descriptions, and the misunderstanding and misuse of use cases that often occurs. Prior to joining NICTA, Dr. Cox was a Research Fellow in the School of Computer Science and Engineering at the University of New South Wales (UNSW), Sydney, Australia, and Lecturer at Bournemouth University in the UK.