

## Presenting software engineering results using structured abstracts: a randomised experiment

David Budgen · Barbara A. Kitchenham ·  
Stuart M. Charters · Mark Turner ·  
Pearl Brereton · Stephen G. Linkman

Published online: 17 July 2008  
© Springer Science + Business Media, LLC 2008  
**Editor:** Dag Sjöberg

**Abstract** When conducting a systematic literature review, researchers usually determine the relevance of primary studies on the basis of the title and abstract. However, experience indicates that the abstracts for many software engineering papers are of too poor a quality to be used for this purpose. A solution adopted in other domains is to employ structured abstracts to improve the quality of information provided. This study consists of a formal experiment to investigate whether structured abstracts are more complete and easier to understand than non-structured abstracts for papers that describe software engineering experiments. We constructed structured versions of the abstracts for a random selection of 25 papers describing software engineering experiments. The 64 participants were each presented with one abstract in its original unstructured form and one in a structured form, and for each one were asked to assess its *clarity* (measured on a scale of 1 to 10) and *completeness* (measured with a questionnaire that used 18 items). Based on a regression analysis that adjusted for participant, abstract, type of abstract seen first, knowledge of structured abstracts,

---

D. Budgen (✉)  
Department of Computer Science, Durham University, Durham, UK  
e-mail: david.budgen@durham.ac.uk

B. A. Kitchenham · M. Turner · P. Brereton · S. G. Linkman  
School of Computing and Mathematics, Keele University, Staffordshire, UK

B. A. Kitchenham  
e-mail: b.a.kitchenham@cs.keele.ac.uk

M. Turner  
e-mail: m.turner@cs.keele.ac.uk

P. Brereton  
e-mail: o.p.brereton@cs.keele.ac.uk

S. G. Linkman  
e-mail: s.g.linkman@cs.keele.ac.uk

S. M. Charters  
School of Applied Computing, Lincoln University, Lincoln, New Zealand  
e-mail: charters@lincoln.ac.nz

software engineering role, and preference for conventional or structured abstracts, the use of structured abstracts increased the completeness score by 6.65 (SE 0.37,  $p < 0.001$ ) and the clarity score by 2.98 (SE 0.23,  $p < 0.001$ ). 57 participants reported their preferences regarding structured abstracts: 13 (23%) had no preference; 40 (70%) preferred structured abstracts; four preferred conventional abstracts. Many conventional software engineering abstracts omit important information. Our study is consistent with studies from other disciplines and confirms that structured abstracts can improve both information content and readability. Although care must be taken to develop appropriate structures for different types of article, we recommend that Software Engineering journals and conferences adopt structured abstracts.

**Keywords** Structured abstract · Randomised controlled laboratory experiment

## 1 Introduction

The adoption of the *evidence-based* paradigm for use in software engineering has made considerable progress in the period following the seminal paper published in 2004 (Kitchenham et al. 2004). Since then, over 20 systematic reviews have been published on software engineering issues (Kitchenham et al. 2007).

A key requirement for the continuing development of evidence-based software engineering (EBSE) is the ability to find, evaluate and aggregate all of the appropriate sources of evidence. In particular, the evidence-based paradigm is one that relies heavily upon the use of *systematic literature reviews* as the means of aggregating the (empirical) evidence that is needed to address a given research question (Kitchenham 2004; Webster and Watson 2002; Petticrew and Roberts 2006). A *secondary* study such as a systematic literature review requires exhaustive searches of the literature in order to identify potentially relevant *primary* studies. Such searches involve two stages: firstly researchers need to perform a wide search to identify as many candidate primary studies as possible; secondly they must undertake a more detailed review of these candidates against specific inclusion and exclusion criteria. Indeed, the first step of the search process is very likely to identify many studies that will actually be irrelevant.

Current procedures, based on experience from other domains such as clinical medicine, education, psychology and the social sciences, suggest that a review of the title and abstract of a primary study should be sufficient to enable the researcher to determine whether or not it is relevant to the study being undertaken (Kitchenham 2004). However, when conducting systematic literature reviews in the domain of software engineering, researchers have reported difficulties with identifying whether or not primary studies are relevant to a topic of interest (Brereton et al. 2007; Jedlitschka and Pfahl 2005). This is because the information provided in abstracts is often incomplete, with the effect that it may be necessary to read other parts of the paper to determine whether or not it is relevant to the particular study.

Consulting the full paper not only involves additional time and effort, but also, since not all papers may be available on-line, there may be costs and delays involved in obtaining these. So lack of information in abstracts may significantly increase both the cost and the time required to perform a systematic literature review.

While our motivation for this study is that of systematic reviewing, as described above, there are other occasions where decisions may need to be based upon

information that has been retrieved from abstracts. These include practitioners who want to know about the effectiveness of particular tools or techniques; researchers planning to perform additional studies on a topic; and researchers who may be intending to replicate a primary study.

One approach to improving the standard of abstracts is to adopt the use of *structured abstracts* (Booth 2003; Hartley 2004). A structured abstract employs a standard set of headings through which the authors précis the key aspects of a study, such as its context, aim, method, results and conclusions. The result of empirical studies conducted in Educational Psychology suggests that structured abstracts are a potentially valuable approach to improving the readability and value of abstracts (Hartley 2003), and in Bayley and Eldredge (2003) the authors identify other benefits of adopting this form as both aiding with searching the literature as well as helping to improve the design of a study. A recent study of six dental journals, three of which had adopted structured abstracts in the late 1980s, also provides convincing empirical evidence in favour of improved relevant information content provided through the use of structured abstracts (Sharma and Harrison 2006) (a broader discussion of previous research related to structured abstracts can be found in Kitchenham et al. (2008)).

To investigate whether the same improvements in information content and readability could be obtained if structured abstracts were to be adopted for software engineering papers and reports, we have undertaken two studies. The first of these was an observational study, that used the conventional abstracts presented for the 2004 and 2006 EASE conferences (Evaluation and Assessment in Software Engineering), prepared structured versions of the abstracts, and analysed them using established readability measures. This confirmed that structured abstracts are longer than unstructured abstracts but that they scored more highly when analysed for readability (Kitchenham et al. 2008). In this paper we use the guidelines for reporting empirical studies provided in (Kitchenham et al. 2002) to describe the second study. This took the form of a randomised controlled laboratory experiment in which participants were asked to act as ‘judges’, assessing one abstract of each form for clarity and completeness.

The following sections first identify the detailed research question that the study set out to answer; describe our experimental method, including the preparation of the structured abstracts; summarise the results; discuss possible threats to validity and observe how our results compare with those from other domains. We finally draw our conclusions about the benefits likely to accrue from the adoption of this form for software engineering papers.

## 2 Research Question

Hartley’s work has suggested that not only is the distinction between structured and unstructured abstracts a significant issue, but also that the typographic layout employed provides an important factor that needs consideration (Hartley and Sydes 1996; Hartley 2000). The second issue leads to two further factors that need to be considered in the specific context of performing systematic reviews in software engineering.

1. The typographical layout used for many conferences (typically two columns, with the abstract being in the same format) differs quite markedly from the forms used

in many journals, which print the abstract in a single-column format, even when using two column format for the main body of the paper.

2. The difference between reading the abstract on the screen or on paper. Hartley's work was essentially paper-based, but in recent years the availability of electronic databases and of search engines has altered practice, with the result that (particularly in the context of systematic reviews), the abstract is highly likely to be read from a screen, at least when determining whether to include a particular primary study.

There are therefore several research questions that could have been addressed in this study.

1. Are structured abstracts easier to understand—which in our context relates to being able to extract the required information about the study?
2. Are structured abstracts easier to understand when presented in single-column than in double-column format?
3. Are structured abstracts easier to read when using a screen than when using paper?
4. How good are the unstructured abstracts currently provided with software engineering papers? In other words, regardless of the question of which form is better—how adequate are existing abstracts?

A comprehensive study should address all of these. However, given finite time and resources we decided to restrict ourselves to addressing the first of those particular questions within a screen-based context, where the abstracts are displayed using a web browser, since this is the format that is most likely to be employed when accessing on-line bibliographic information. This study therefore addressed the question:

When displayed on a screen using a web browser and basic HTML formatting, do structured abstracts contain more relevant information in a more readable format than unstructured ones?

Within this context, we interpret 'more readable format' as meaning that the reader can more readily obtain the necessary information about the study described in the paper.

So, for this paper we have set out to investigate the following two hypotheses:

- *Null Hypothesis 1:* Structured abstracts and unstructured abstracts are not significantly different with respect to completeness.
- *Alternative Hypothesis 1:* Structured abstracts are significantly more complete than unstructured abstracts. (Note that we are also interested in the possibility that structured abstracts are less complete than unstructured abstracts.)
- *Null Hypothesis 2:* Structured abstracts are not significantly different from unstructured abstracts with regard to clarity.
- *Alternative Hypothesis 2:* Structured abstracts are significantly clearer than unstructured abstracts.

Although we have presented our experiment in terms of testing two hypotheses, we must point out that having added information to unstructured versions of abstracts it is unlikely that a structured abstract could fail to score better than an unstructured abstract with respect to *completeness*. Furthermore, if we find that

*clarity* is significantly better for structured abstracts we will not be able to tell whether this is due to the restructuring or due to the additional information. Thus, if we formally reject both null hypotheses we will interpret this to mean that conventional abstracts have problems with respect to completeness and clarity and that structured abstracts are a potential means of addressing these problems. We cannot claim that restructuring abstracts alone will address clarity and completeness problems.

### 3 Experimental Method and Materials

The study was conducted as a controlled laboratory experiment, where participants were asked to read the abstracts of two different papers, one of which was structured, while the other was not. The participants were asked to assess the information content of each abstract. The order in which the abstracts were presented (i.e. whether they read the structured one first or second) was randomised across participants as was the allocation of abstracts. Before starting on the study, we first developed a comprehensive *research protocol* detailing our plans (Budgen et al. 2007b). We also updated this to record any 'divergences' from this that occurred during performance of the study.<sup>1</sup>

#### 3.1 Population

Participants were drawn from undergraduate students in their final year of study, postgraduate students, researchers and practitioners. Our rationale for this was that while for the immediate purpose of conducting systematic literature reviews the category of *practitioner* is less significant (on the assumption that such reviews will mainly be conducted by students and researchers), in the bigger picture of software engineering these constitute an important grouping, since we also need to consider the role of structured abstracts in the context of conveying information to them. Overall therefore, we tried to include all of these categories in our study. For the purpose of analysis, we included a question in the demographic element of the questionnaire that enabled us to separate students from more experienced workers.

Our aim was to enlist between 50 and 100 participants to act as judges and we were finally able to recruit 64 participants, divided into 20 students and 44 researchers and practitioners.

#### 3.2 Selection of Participants

For the categories of *researcher* and *practitioner* we constructed a list of possible participants which tried to avoid including people likely to be pre-disposed to favour structured abstracts (possibly those already involved in evidence-based research). As a primary source we used the EASE conference mail-list together with contacts suggested by the EASE programme committee. This process was not completely

---

<sup>1</sup>The protocol was reviewed by the team as it was developed, and was then subjected to a more detailed review at a meeting with James Hartley present as expert advisor. We also took advice from the information services (libraries) at Keele and Durham to the effect that using and restructuring the abstracts in this manner did not infringe copyright in any way.

unbiased, but since structured abstracts are only used very rarely in software engineering, we considered the likely effects to be small. Our aim was to have half of our participants drawn from these categories (taken together).

For *students* we used a more generalised recruitment process. Three groups considered to be suitable recruits for our study were:

- final year students—limited experience of reading papers, but should have the background to be able to read the abstracts;
- taught postgraduate students—again limited experience of reading papers, but with background broadly similar to the final year students—with one possible limitation that current cohorts on taught postgraduate degrees are heavily weighted to non-native English speakers;
- research postgraduate students—probably the most ‘representative’ group of those likely to need to read abstracts and to conduct systematic forms of literature review—but only available in limited numbers and possibly also containing a large proportion of non-native English speakers.

Again, our aim was to have half of our participants from these categories. Initial recruitment was by asking for volunteers from the appropriate cohorts at Durham and Keele Universities, but this was later extended to volunteers from a number of other universities.

In practice, the students proved to be the more difficult group to recruit—possibly because, unlike (say) psychology students, our students are not accustomed to volunteering to take part in such studies. We were also very careful not to put pressure on students to take part, so that the invitation was usually issued by the researchers rather than the academics—and as these may have been less well-known to the students this may also have been a negative factor.

### 3.3 The Abstracts

Here we address the questions of the choice of abstracts to use, how they were to be rewritten in a structured form, and what consultation was undertaken with the original authors.

#### 3.3.1 *Selecting the Abstracts*

A key question was the number of abstracts to use. Based upon the size of our research team and the likely number of participants, we decided to use a total of 25 papers, since this spread the re-writing task evenly among members of the team.

The 25 papers were taken from the set of 103 empirical papers previously identified and analysed by Sjøberg et al. (2005). This set of papers were taken from nine journals and three conference proceedings, published over the period 1993–2002. The task of selection was performed on a random basis by Dag Sjøberg, and maintained the same proportion of journal papers and conference articles as were present in the complete set. This resulted in a subset of 6 conference articles and 19 journal papers. None of the abstracts for these papers were in a structured form.

The basic structure we propose is intended to be suitable for most empirical studies, whether experiments, case studies, surveys or observational studies. However, all these papers describe controlled experiments (including both randomised

experiments and quasi-random experiments). This indicates some limitation to the generalisability of our results and also influenced our associated guidelines for constructing structured abstracts, which currently are oriented towards experiments rather than other types of empirical study.

### 3.3.2 *Rewriting Abstracts into a Structured Form*

Hartley's studies used two approaches:

- all abstracts being re-written by the principal investigator (James Hartley) as described in Hartley (2003);
- all abstracts being re-written by the original authors (Hartley and Benjamin 1998).

When using the first approach there is an additional option of sending the re-written abstract to the original authors to ask their permission to use and to invite their comments and suggestions about the revised form. This was the approach we adopted in our study.

To create the structured abstracts, each abstract was re-written by one member of the team and then checked and reviewed by another member of the team. Five of us acted as authors, and all six as reviewers, organised so that each of the five abstracts rewritten by an author was reviewed by a different reviewer, allocated randomly. We used the following headings and contents guidelines to construct the structured abstracts:

**Background:** Previous research or rationale for a study

**Aim:** Hypotheses to be tested or goal of the study

**Method:** Description of the type of study, treatments (including control), number and nature of experimental units (people, teams, algorithms, programs, tasks etc.), the experimental design, outcome being measured

**Results:** Treatment outcome values, standard deviation and/or level of significance

**Conclusions:** Future work, limitations of study

(These are largely those proposed in Jedlitschka et al. (2008), except that we included 'limitations' under 'Conclusions' rather than treating as a separate heading.)

Guidelines on the procedures to follow for the task of re-writing were prepared in advance, and a summary of these are attached as Appendix 1. To ensure consistency of style we also adopted the following conventions:

- Each heading should be set in boldface type and should end with a colon
- The sentence beginning after the heading should start with a capital letter
- References should be removed and acronyms expanded
- The word count should be kept below 300 if possible

Since not all of the information identified for the structured form was necessarily available from the existing abstract, where necessary this was obtained from the body of the paper.

One of our concerns was that the process of rewriting should not change the writing style. With this in mind, we instructed team members to rewrite the abstract as far as possible by using the original sentences from the abstract, and, in the event

of missing information, to take this from the body of the paper. However, adherence to this instruction was not formally monitored by the research team. As a final check, as described in the next sub-section, we also asked the original authors to check our revised abstract for both correctness and style.

### 3.3.3 Involvement of the Original Authors

After re-writing each abstract into a structured form, we then tried to contact the original authors, asking them to check our re-written abstract, and providing them with some information about structured abstracts. The authors of sixteen papers did respond, and for those papers where we were unable to contact any of the authors, a further member of the team was used to act as an additional checker. (A further author did respond after the review process—fortunately, they agreed with the changes!) Where the authors made any suggestions about the wording of the structured abstract we endeavoured to take these into account, on the basis that these would be closer to the abstract that they would have written. After all required changes had been incorporated, the original reviewer assessed the structured abstract for completeness and clarity using the same set of questions as the participants.

### 3.4 Scoring the Abstracts for Clarity and Completeness

The initial version of the questionnaire to be used in judging the abstracts was based upon the one used in Hartley and Benjamin (1998). We used a set of 18 questions to assess completeness (counting the number of ‘yes’ responses to measure completeness for subsets of information and overall). To measure clarity, we used a subjective value on a scale of 1 to 10 in order to be consistent with previous research (Hartley 2003). Participants were asked to provide this in response to the following question:

Please give an assessment of the clarity of this abstract by entering a number on a scale of 1–10, where a value of 1 represents *Very Obscure* and 10 represents *Extremely Clearly Written*.

To ensure that the set of questions was complete and consistent for our purpose, a perspective-based analysis of the questionnaire was conducted by one member of the team (DB), with the outcomes being checked by three others (BAK, SC, MT).

The arguments for adopting a perspective-based approach for validation are essentially the same as those used in Kitchenham et al. (2006) when using this approach for evaluating guidelines for reporting empirical studies in software engineering. These are that using this approach is valid; feasible for us to do; cost effective; and provides learning potential (although this last point was less immediately relevant for this study).

The set of perspectives adopted were a subset of those used in Kitchenham et al. (2006). These were chosen on the basis that not all of the original categories of user would be likely to seek information from the abstract. We therefore considered the following four perspectives:

1. *Researcher* who reads the abstract to determine whether the paper offers important new information on a topic of interest;



2. *Practitioner/Consultant* who wants to know if the paper will provide information that will be of use in industry or commerce, and whether the results are likely to be of direct use to their company/clients;
3. *Systematic Reviewer* seeking quantitative or qualitative information that can be integrated with the results of other studies;
4. *Replicator* considering whether it may be worthwhile or practical to repeat the study.

The category of *systematic reviewer* was essentially an extended form of the perspective of *meta-analyst* as used by Kitchenham et al. (2006), since we should be looking beyond aggregation of numerical results. The perspectives of *reviewer* and *author* were not considered relevant to this analysis, nor were any other perspectives identified.

Since the original set of perspective questions were intended for use in assessing full papers, our first step was to determine which ones were also appropriate to an abstract, and we then used these as a check upon the questionnaire. In a few cases we used variations upon the original questions. (Again, a fuller analysis is presented in the research protocol (Budgen et al. 2007b).)

To complete the exercise, each of the perspective questions was matched against the questions used in the questionnaire. There was no attempt at exact matches, and it was accepted that a perspective question might be mapped on to more than one question in the questionnaire. The outcome of this exercise is shown in Table 1. (The numbers in the columns relate to the original perspective questions, where letters have been used, these indicate questions that were adapted to meet the needs of an abstract rather than a paper.)

Inevitably some interpretation had to be made in mapping the perspectives, and so some questions from a particular perspective have been mapped on to more than one question from the questionnaire. This process then left two sets of elements to consider:

- *Unmapped elements in the questionnaire* There were no questions without a link to a perspective, so on that basis the questionnaire required no changes.
- *Unmapped perspective elements* There were two of these, discussed below.
  1. Discussion of implications for future research (researcher question 17; practitioner question 22).
  2. Description of the form of analysis employed (reviewer question d).

The first of these corresponded to the question 17 of the original questionnaire, which was one that had been removed when adapting it for this study. Replacing this with new wording was therefore considered an appropriate step. The new question 17 was worded as: “Is there any discussion of required future research?”. The second arose from one of the elements added by our own perspective-based analysis, extending the original set used in Kitchenham et al. (2006). The proposed action in response was to insert a new question 12 (moving all other questions down one place), worded as: “Is there any description of the form of analysis performed?”. The final version of the questionnaire as used in this study is listed as Appendix 2.

**Table 1** Checking the original questionnaire against the perspectives for completeness

No.	Original question	Researcher	Practitioner	Systematic reviewer	Replicator
1	Is the rationale for the study reported?	2, 5	2, 6		
2	Is the aim/purpose of the study reported?	2	2	3	
3	Is a hypothesis (or hypotheses) provided?	4		3	2
4	Is there any indication of where this study took place?		9		
5	Is the number of participants reported?			1	3
6	Are the types of participant reported?			c	3
7	Is any information about the experience of the participants provided?				3
8	Is the skill level of the participants reported?				3
9	Is there any description of how the study was performed?	10	16	6	5
10	Does it report how the participants were allocated to different tasks or conditions?				5
11	Is the way that the data was collected reported?			a	
12	Are the main results summarised in the abstract?	13	4		
13	Are actual numbers from the results presented in the abstract?	a			
14	Is any statistical information provided about the results?	a		b	
15	Are any conclusions drawn?	14	3, 4		
16	Are any limitations of the study identified?	16	8, 14		6

### 3.5 Allocation to Participants

We performed randomisation through the allocation of papers to participants, and so simply constructed lists of participants as these were recruited.

To allocate the papers, we used the same mechanism as was employed by Hartley. Given a set of abstracts (both structured and unstructured forms), which are indexed as 1.. $n$ , (where the index value is obtained by allocating random numbers to each paper and then ordering on the random number), then:

- Each participant was allocated unstructured paper  $i$  and structured paper  $i + 1$ ;
- The participant allocated unstructured paper  $n$  was then allocated structured paper 1 using a cyclic allocation strategy.

The value of  $i$  for each participant was allocated on a random basis by using the *rand()* function of PHP, and a manual check was also performed to ensure that no participant was allocated an abstract taken from a paper for which they were a co-author.

## 4 Data Collection and Analysis

### 4.1 On-Line Data Collection

Once the final versions of the structured abstracts were completed, these, together with the original unstructured versions were converted to HTML and stored in a database. We also performed some initial analysis at this stage, in that values of the Flesch Reading Ease measure and Gunning Fog Index were recorded for each document (Flesch 1948; Flesch-Kincaid Readability Test 2006; Automated Readability Index 2006). While in (Hartley and Sydes 1997) the authors note that these metrics ignore many factors, they do also observe that when applied to two versions of the same abstract then the results should provide some indication of whether or not one version might be easier to read than the other. The length of each index (in words) was also recorded.

The order of judging was deemed to be important and therefore the order of allocation was tightly controlled, with participants only being able to view the second abstract after the task of judging the first one had been completed. To achieve this, and to coordinate data collection, we used web-based form-driven data collection. Participants first registered their interest in taking part and were allocated a unique code. They then accessed the project server using a web browser and logged on to this by using their identifying code, after which they were presented with a series of pages containing (in the following order):

- the first abstract together with the list of questions, which could be completed using ‘radio button’ style selection
- the second abstract together with the same list of questions
- an invitation to provide up to three ‘likes’ and ‘dislikes’ about structured abstracts, followed by some questions of a demographic nature (the details of these questions are provided in Appendix 3).

Participants were provided with a ‘next’ button for proceeding through the pages and completed each task by pressing a ‘submit’ button to store their responses. They could only progress to a following task when all questions for the current one had been answered. The only data recorded was the set of responses and a log of the time taken (this provided some assessment of the effectiveness of the forms as well as how much time participants typically spent on the exercise). Where possible, participants were asked to use the Mozilla browser to ensure consistency of appearance on the screen.

### 4.2 Data Preparation

#### 4.2.1 Quantitative Data

The data collected for the assessment of each abstract included:

- *Structured* which indicates whether the assessment referred to the structured (1) or unstructured version of the abstract (0)

- *Seen1st* which indicates whether a specific participant saw a specific abstract first (1) or second (0)
- *Order* which indicates whether the order in which the participant saw abstracts was structured followed by unstructured (1) or vice versa (0)
- The participant identifier
- The abstract identifier
- The answer to each of the 18 completeness questions (yes/no/uncertain)
- The answer to the clarity question (a value between 1 and 10). This was used as our dependent variable in our analysis of clarity.

The data collected for each participant included the answer to each of the demographic questions.

In order to analyze the data a number of new variables were constructed for the data related to the assessment of each abstract:

- *Total Yes* = Total number of yes answers to the 18 completeness questions. This was used as a dependent variable in our tests of completeness.
- *Total U* = Total number of uncertain answers to the 18 completeness questions
- *Total Aim* = Total number of yes answers to the two questions about aims
- *Total Method Context* = Total number of yes answers to the four questions concerning the context of the study
- *Total Method Conduct* = Total number of yes answers to the four questions concerning the conduct of the study
- *Total Results* = Total number of yes answers to the three questions concerning results
- *Total Conclusions* = Total number of yes answers to the three questions concerning results
- A dummy variable identifying which participant provided the assessment (i.e. a set of 64 variables which take the value 1 for the participant who provided the specific assessment and 0 otherwise). Note participants were labeled in the order in which they enrolled for the study, so the dummy variables cover values *P1* to *P75* with 11 variables having no data attached to them.
- A dummy variable identifying which abstract was being assessed (i.e. a set of 25 variables *A1* to *A25* which takes the value 1 for the abstract being assessed and 0 otherwise).
- Two dummy variables were constructed to summarise whether the participant had prior knowledge of structured abstracts: *Known* had a value 1 if the participant knew anything about structured abstracts and 0 if the subject did not know anything or did not answer the question. *NotKnown* had a value 1 if the participant did not know anything about structured abstracts and 0 if the participant knew about structured abstracts or did not answer the question.
- Three dummy variables were constructed to summarise whether the participant preferred structured or unstructured abstracts or had no preference: *Prefs* took the value 1 if the participants preferred structured abstracts and 0 if they preferred unstructured abstracts, had no preference or did not answer the question; *Prefu* took the value 1 if the participants preferred unstructured abstracts and 0 if they preferred structured abstracts, had no preference or did not answer the question; *Nopref* took the value 1 if the participant had no preference and 0 if

the participant preferred structured abstracts, preferred unstructured abstracts or did not answer the question.

- Two dummy variables were used to summarise whether the participant was a student. *Student* took the value 1 if the participant was a student (i.e. confirmed that they were a post graduate research student, a graduate research student or an undergraduate), or 0 if the participant was not a student or did not answer the question; *NotStudent* took the value 1 if the participant was not a student and 0 if the participant was a student or did not answer the question.

The quantitative data was analysed to test the hypothesis that structured abstracts were more complete and clearer than unstructured abstracts using two regression-based analyses. A regression approach was used rather than an analysis of variance approach because the dataset was unbalanced (e.g. different number of participants viewed each abstracts and other factors of interest such as preference and role were not balanced across the participants).

First we performed a stepwise regression analysis using all the data points for each dependent variable including the following independent variables:

- Structured
- Order
- Seen1st
- Known
- NotKnown
- Student
- NotStudent
- Prefs
- Prefu
- NoPref
- Twenty four dummy variables identifying the abstract
- Sixty three dummy variables identifying the participant.

The analysis allowed us to cater for difference among participants and among abstracts and provides an unbalanced sequential cross-over analysis. The analysis was developed specifically to deal with the data collected in this experiment. It was based on concepts found in Senn (2002), who discusses how to analyse cross-over experiments, and Milliken and Johnson (1992), who discuss how to handle messy (i.e. unbalanced) data and explain the use of linear regression to analyse statistical designs.

Since use of a specialised analysis is potentially risky, we also analysed only the data points corresponding to the first abstract seen by each subject, including the following independent variables:

- Structured
- Known
- NotKnown
- Student
- NotStudent
- Prefs
- Prefu
- NoPref

This analysis corresponds to a simple parallel experiment and allowed us to check whether the results were consistent for both analyses.

#### 4.2.2 *Qualitative Data*

In addition to their preferences with respect to the form of abstract and their software engineering roles, the participants were asked to identify three things that they liked about structured abstracts and three things that they disliked. This data was analysed using a bottom-up form of content analysis. In a method similar to the first step in grounded theory (Glasser and Strauss 1967), we reviewed all the responses and identified a set of categories to describe the positive and negative answers. We ensured that the categories covered all of the comments. Finally, we counted the number of responses in each category and tabulated them against role and preference.

The categories and assignment of answers to categories was done by one researcher and checked by another. After the checking process, the checker had no disagreements concerning the categories (which were simply a summary term derived from the comments themselves) and two disagreements with respect to categorising the 168 comments. The two proposed changes were accepted by the coder.

In a separate question, the participants were also asked if they had any other comments. Twenty-two participants provided general comments, which were reviewed and categorised in a similar manner to the comments concerning structure.

## 5 Results

### 5.1 Preliminary Data Validation

The data were checked for consistency. In particular, values for each abstract in each format were checked to see whether participants provided broadly similar assessments. One participant provided a very low value (i.e. 3) for the clarity of the structured abstract compared with values of 11 and 10 for other participants who reviewed the structured version of the same abstract. The participant also commented that “I would give a higher assessment of clarity of the structured abstract if I saw the conventional abstract first”. Thus, it appeared that the participant was unhappy with his/her evaluation of the structured abstract. For this reason, the data related to the structured abstract for that participant was removed from the data set.

### 5.2 Analysis Results

#### 5.2.1 *Summary Statistics*

The average and standard deviation for responses for each abstract in its structured and original format are shown in Table 2 for completeness (TotalYes) and Table 3 for clarity.

**Table 2** Completeness (TotalYes) per abstract for structured and original abstracts

Abstract number	Structured			Unstructured		
	Number of participants	Mean	SD	Number of participants	Mean	SD
1	2	14	1.41	2	1	1.41
2	2	8.5	0.71	2	5	1.41
3	2	10.5	2.12	2	1.5	0.71
4	2	11	0	2	5.5	0.71
5	3	14.3	2.52	2	4.5	0.71
6	3	11	1.71	3	2.3	2.08
7	2 (3)	10.5 (8)	0.71 (4.36)	3	4.3	0.58
8	3	14	1	3	11	1
9	3	7.7	1.53	3	5.3	2.31
10	3	11	0	3	2.7	2.89
11	3	11	1	3	3.7	1.15
12	3	10.7	0.58	3	9.3	1.53
13	2	10.5	3.54	3	5	1
14	3	10.7	2.08	2	7.5	2.12
15	2	13	2.82	3	3	2.55
16	2	14	1.41	2	6	0
17	3	13	2.64	2	3	1.41
18	2	14.5	0.71	3	2.7	1.53
19	4	13	2.94	1	9	0
20	2	14.5	3.54	4	7.7	0.96
21	3	12.3	2.31	2	4	1.41
22	2	9.5	2.12	3	5.7	2.52
23	2	10	0	3	3	0
24	2	9.5	0.71	3	5	1
25	2	10.5	0.71	2	2	0
Total	62 (63)	11.61 (11.47)	2.42 (2.64)	64	4.83	2.80

The average completeness values for the structured abstracts in Table 2 are all greater than the averages for the unstructured abstracts. For the clarity values in Table 3, in one case the average clarity for structured and unstructured abstracts was the same; in all other cases the average clarity was greater for the structured abstracts than the conventional abstracts. The results in Table 2 and Table 3 make it fairly clear that the structured abstracts score better on completeness and clarity than the unstructured abstracts. Nonetheless, we performed a more detailed analysis of the results to assess whether any factors such as the order in which the abstracts were seen, knowledge of structured abstracts, or whether the participant was a student, might have biased our results.

Tables 2 and 3 show that the trend for the mean completeness and clarity to be greater for the structured than unstructured version of each abstract is stable, whether or not the unusual data point was removed (values in parentheses are those that include that data point). However, the combination of the large standard deviations and the comment of the participant suggested that the data point was not trustworthy. We therefore removed it from the more detailed analyses. (See next section.)

**Table 3** Clarity per abstract for structured and original abstracts

Abstract number	Structured			Unstructured		
	Number of participants	Mean	SD	Number of participants	Mean	SD
1	2	8	1.41	2	1	0
2	2	7	1.83	2	7	1.41
3	2	9	1.41	2	3	1.41
4	2	7	0	2	4	1.41
5	3	7.7	1.15	2	6.5	2.12
6	3	7.7	1.15	3	3.3	2.52
7	2 (3)	8 (6)	1.41 (3.61)	3	5	1
8	3	8	0	3	5.7	1.53
9	3	7.7	0.58	3	6.7	0.58
10	3	8	0	3	2.7	0.58
11	3	8.3	0.58	3	4.3	1.53
12	3	7.7	1.15	3	6.7	1.15
13	2	8	0	3	7	1
14	3	8	1	2	7.5	0.71
15	2	8	0	3	4	1.73
16	2	9	1.41	2	4	0
17	3	8	1	2	3	1.41
18	2	7	1.41	3	4.3	2.31
19	4	7.7	2.87	1	4	0
20	2	7.5	0.71	4	5.5	1.29
21	3	8	1	2	4.5	2.12
22	2	6.5	0.71	3	4.3	3.51
23	2	8	1.41	3	4.3	0.58
24	2	8	1.41	3	5	1
25	2	6.5	2.12	2	3.5	2.12
Total	62 (63)	7.79 (7.70)	1.19 (1.37)	64	4.75	1.97

### 5.2.2 Demographic Information

Fifty-seven participants answered questions about their knowledge of structured abstracts. Overall 41 (72%) said they had known about structured abstracts before taking part in the experiment. The type of knowledge that participants had is shown in Table 4.

Fifty-seven participants answered questions about their software engineering role (Table 5).

**Table 4** Participants' knowledge of structured abstracts

Knowledge of structured abstracts	Yes	No
Heard about them but not seen them before	17 (41%)	24 (59%)
Read papers about their use	19 (46%)	22 (54%)
Read papers with structured abstracts	32 (78%)	9 (22%)
Created structured abstracts for your own papers	18 (44%)	23 (56%)



**Table 5** Participants' software engineering roles

Participant role	Yes
Full time researchers	26 (45.5%)
Practitioner	2 (3.5%)
Postgraduate research student	13 (23%)
Postgraduate student	4 (7%)
Undergraduate student	6 (10.5%)
Other	6 (10.5%)

The people who specified “Other” included:

- Part-time researcher and part-time practitioner (2)
- Scientist and project leader
- Professor
- Academic (research and teaching)
- Technology transfer consultant

Fifty-six participants reported their experience in years. Responses ranged from 0 to 45 years with a median of 9 years. A box plot of the experience distribution of the participants is shown in Fig. 1. It shows the distribution of students and non-students separately. Twenty participants identified themselves as students. Their median experience was 2 years. The median experience of the other 36 participants was 15.

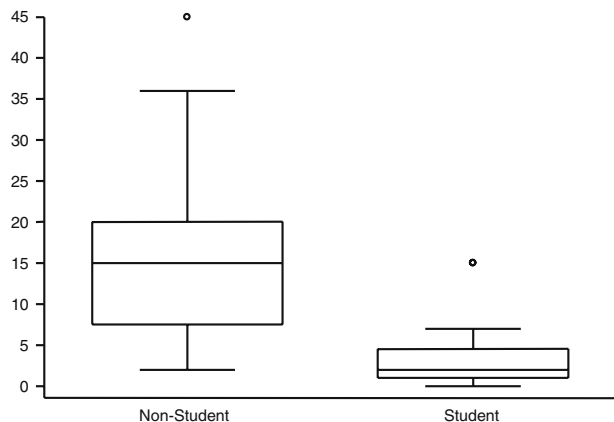
Overall it is clear that our participants included mainly academics, researchers and students, with few practitioners.

### 5.3 Hypothesis Testing

The impact of structure on completeness (TotalYes) and clarity was assessed using two separate regression models.

The first regression model used all the available observations (126 in total) but included dummy variables for the major experimental design components i.e. whether the abstract version was structured or not, whether the abstract version was the first

**Fig. 1** Distribution of experience in years for study participants



seen by the participant, whether the sequence for the participant was structured followed by unstructured (or not), and for each abstract for each participant. In addition, dummy variables were included for other factors that might have biased our results i.e. for role (student or not), knowledge of structured abstracts, and preference.

The observations were analysed by a stepwise regression algorithm using the SPSS statistical tool. The analysis produced the model shown in Table 6 for the completeness variable. This confirms that structure significantly increases the completeness of the abstracts. Furthermore, none of the uncontrolled factors (preference, role, knowledge of structured abstracts) had a significant effect on the dependent variable. In addition most of the design factors had no significant impact including the cross-over variables (sequence and 1stseen). Five of the abstracts had a significant impact on the dependent variable: in all cases they increased the dependent variable. Five participants had a significant impact on the dependent variable: three scored abstracts lower than other participants, two scored higher.

The final model was checked using the STATA statistical program. The model was identical but analysis of the goodness of fit statistics identified one point with a very high leverage. The full stepwise regression was re-run with the abnormal point removed and the same model was obtained with no changes to the selected variables or their coefficients.

The stepwise regression model obtained for the clarity dependent variable is shown in Table 7. The analysis confirms that structured abstracts score significantly higher on the clarity scale than the unstructured abstracts. Again some abstracts and participants had a significant impact on the dependent variable. The final model was checked using the STATA statistical program. The model was identical but again one data point had a high leverage (the same participant and the same abstract as before). Again re-running the stepwise analysis with the high leverage data point removed did not change the variables selected or their coefficients.

As a further check on the stability of the results, we fitted another regression model to the data points corresponding to the first abstract viewed by each participant (62 observations in total). This was equivalent to a parallel experiment. In this analysis only the dummy variables related to role, knowledge of structured

**Table 6** Stepwise regression model for completeness (TotalYes)

Variable	Coefficient	Standard error	<i>t</i>	Significance
Constant	4.20	0.282	14.92	<0.001
Structured	6.65	0.371	17.89	<0.001
A8	4.97	0.866	5.745	<0.001
A12	2.47	0.866	2.858	0.005
A16	2.47	1.049	2.358	0.020
A19	3.43	0.995	3.445	0.001
A20	3.58	0.868	4.129	<0.001
P10	2.97	1.468	2.026	0.045
P28	-4.20	2.073	-2.027	0.045
P40	-3.74	1.537	-2.434	0.016
P48	-3.53	1.468	-2.402	0.018
P51	3.74	1.468	2.366	0.020

**Table 7** Stepwise regression model for clarity

Variable	Coefficient	Standard error	<i>t</i>	Significance
Constant	4.44	0.180	24.618	<0.001
Structured	2.98	0.234	12.736	<0.001
A9	1.24	0.550	2.253	0.026
A12	1.24	0.550	2.253	0.026
A13	1.77	0.600	2.951	0.004
A14	1.57	0.600	2.624	0.010
P14	2.57	0.933	2.756	0.007
P27	2.57	0.933	2.756	0.007
P28	−3.44	1.317	−2.610	0.010
P40	−2.43	0.933	−2.601	0.011
P48	−2.93	0.933	−3.137	0.002
P49	3.073	0.933	3.294	0.001
P71	2.57	0.933	2.756	0.007
P74	2.07	0.933	2.222	0.028

abstracts and preference were included in addition to the main independent variable identifying whether the abstract was structured or not.

For completeness, the model produced by the stepwise regression is shown in Table 8. This confirms that the structured abstracts are more complete than the unstructured abstracts. However, they also indicate that student participants scored completeness higher than non-students for all abstracts.

For clarity, the model produced by the stepwise regression is shown in Table 9. This confirms that the structured abstracts score significantly higher on the clarity scale than the unstructured abstracts.

#### 5.4 Missing Information

Table 10 shows the number of abstracts that were judged by one or more participants to have missed out a specific topic area. The number in parenthesis identified the number of participants who judged an abstract to be missing that topic. If the value in parenthesis is greater than the other value, it means that several participants judged a particular abstract to be missing information on the specific topic. Overall it is clear that the structured abstracts are less likely to miss information completely. The topics that are most likely to be missing from unstructured abstracts are *Method* and *Conclusions*.

None of the unstructured abstracts provided relevant information in every section and five of the unstructured abstracts appeared to provide no relevant information in any section. It must be noted that the structured versions of the abstracts did not

**Table 8** Model for completeness (parallel experiment)

Variable	Coefficient	Standard error	<i>t</i>	Significance
Constant	4.24	0.444	9.54	<0.001
Structured	6.45	0.554	11.65	<0.001
Student	1.98	0.592	3.351	0.001

**Table 9** Model for clarity (parallel experiment)

Variable	Coefficient	Standard error	<i>t</i>	Significance
Constant	5.23	0.304	17.188	<0.001
Structured	2.39	0.424	5.643	<0.001

always improve the situation. Several abstracts were assessed as having information related to a specific section in the unstructured format but not in the structured format, particularly for the Background section.

### 5.5 The Relationship between Completeness and Clarity

To investigate the relationship between completeness and clarity, for each abstract we subtracted:

- The mean completeness score for the structured abstract from the mean completeness score of the unstructured abstract, to give us a measure of the increase in completeness.
- The mean clarity score for the structured abstract from the mean clarity score of the unstructured abstract, to give us a measure of the increase in clarity.

The relationship is shown graphically in Fig. 2. This shows a positive relationship between the improvement in clarity and the improvement in completeness. A linear regression showed a significant gradient, 0.397, with a 95% confidence interval of (0.212,0.582) and a non-significant intercept, 0.405 with a 95% confidence interval of (−0.967, 1.777). These results show that the more information is added to an abstract, the greater the increase in clarity, implying that the increase in clarity is due to the increase in information rather than the change in structure. However, the results also suggest that adding structure to an already reasonably complete abstract will not degrade clarity.

### 5.6 Qualitative Data Analysis Results

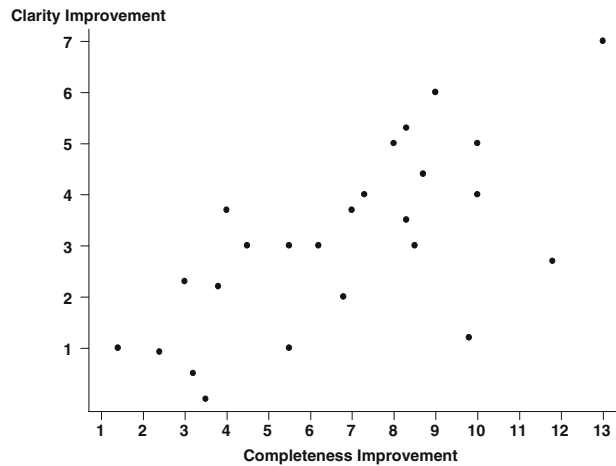
Fifty-seven participants expressed their preference for abstract format:

- 4 (7%) preferred conventional abstracts
- 13 (23%) had no preference
- 40 (70%) preferred structured abstracts.

**Table 10** The number of abstracts missing information on a specific topic

Topic	Structured	Unstructured
Background	9 (12)	13 (21)
Aim	1 (1)	13 (19)
Method (experimental method)	1 (1)	15 (35)
Method (experimental conduct)	3 (3)	18 (35)
Results	2 (2)	10 (15)
Conclusions	2 (2)	17 (27)

**Fig. 2** The relationship between the average improvement in clarity and completeness for the structured abstracts



The relationship between role and preference is shown in Table 11. It appears that the students are more enthusiastic about structured abstracts than other roles, but there is no statistically significant difference.

Participants from all preference groups provided comments describing both things they liked and things they disliked about structured abstracts. The researchers classified the responses into a set of positive categories and negative categories shown in Tables 12 and 13 respectively.

The number of positive comments by preference is shown in Table 14 and the number of positive comments per category is shown in Table 15. (Three comments had multiple classifications and in each case the comment was treated as being two separate comments.) Most participants were able to identify one or more positive aspects of structured abstracts no matter what their preference. Most positive comments were the associated methodology, followed by readability, structure and search. Generally participants seemed to appreciate the formality of structured abstracts and thought it would help both authors and readers.

The number of negative comments by preference is shown in Table 16 and the number of negative comments per category is shown in Table 17. (One participant said that structured abstracts were both boring and too long in one comment. Again, this comment was treated as two comments for classification and counting purposes.) Participants were able to identify negative comments irrespective of their personal preference. The most frequent negative comment was that the structure was inappropriate for other types of paper. This is a fair criticism. The structure we propose is only appropriate for empirical studies and may not be the best structure

**Table 11** The relationship between role and preference

Student	Preference			Total
	Conventional	No preference	Structured	
No	4	9	23	36
Yes	0	4	17	21
	4	13	40	57

**Table 12** Positive categories for structured abstracts and their meaning

Category	Meaning
Conciseness	Makes abstract more concise
Comparability	Makes it easier to compare results
Completeness	Good summary Missing parts are noticeable
Methodology	Provide guidelines for abstract writer Help write paper
Paper relevance	Easy to decide whether to read paper
Readability	Easier to read, clearer
Search	Easier to locate specific bits of information Help automated searching
Standardisation	
Structure	Structure itself is good Specific parts of the structure are welcome
Timesaving	Save time reading paper

for some specialized type of empirical study (e.g. systematic literature reviews) and is certainly inappropriate for theoretical studies.

The next most common negative comment concerned the issue of whether the format was too restrictive and risked leaving out important information. Again this is a fair criticism. As they gain experience of using structured abstracts, individual journals amend the basic structure to suit different types of paper. For example, the *Annals of Internal Medicine* added a limitations section to their structured abstract format (Editorial 2004) and Hartley and Benjamin report three different formats for experimental articles and three different formats for review articles used by four

**Table 13** Negative categories for structured abstracts and what they mean

Category	Meaning
Automatic rejections	Journal might unfairly reject papers not in the right format
Boring	Dull, boring
Difficult	Hard to construct
Duplication	Duplicated information
Inappropriate	Not be suitable for all types of abstract
Misuse	May stop people thinking
No graphics	No graphics
Poor English	Stilted
Poor readability	Hard to read
Over-restrictive	People may be forced to put in unnecessary information and leave out important information
Too detailed	Too much detail Too technical
Too long	Take too much space
Uninformative	Can convey less information
Useless headings	Heading are not helpful

**Table 14** Number of positive comments by preference

Preference	Total	Made at least one positive comment	Made at least two positive comments	Made 3 positive comments	Total comments
C	4	3	2	0	5
N	13	12	9	2	23
S	40	37	29	18	84
	57	52	40	20	112

psychology journals (Hartley and Benjamin 1998). For non-empirical papers, educational research proposes a structure with sections: background, purpose, sources of evidence, main argument, conclusions.

The general comments we received could be generally classified as follows:

- Six comments were related to criticisms of the questionnaire itself.
- Six comments were related to how the individual reacted to the experiment itself (and as a result, one set of responses were removed from the analysis).
- Six comments either expressed interest in, or were in favour of or querying how to use, structured abstracts.
- Three comments noted that conventional abstracts could be as good as structured abstracts.
- One comment expressed the view that the headings were unnecessary.

## 6 Discussion

Our study suggests that conventional abstracts are likely to omit substantial amounts of relevant information. The results also indicate that structured abstracts are significantly more complete and clearer than unstructured abstracts.

**Table 15** Number of positive comments in each category by preference

Category	Preference			Total
	C	N	S	
Methodology	1	6	16	23
Readability	0	4	15	19
Structure	2	1	15	19
Search	2	3	12	17
Completeness	0	2	10	12
Standardisation	0	4	5	9
Timesaving	0	1	6	7
Paper relevance	0	1	5	6
Comparability	0	0	2	2
Conciseness	0	1	1	2
Total	5	23	87	115

**Table 16** Number of negative comments by preference

Preference	Total	Made at least one negative comment	Made at least two negative comments	Made three negative comments	Total comments
C	4	4	2	1	7
N	13	11	4	2	17
S	40	20	9	2	31
	57	35	15	5	55

Participants felt that structured abstracts provided a useful standard methodology that would help authors both in the task of writing abstracts and also in the task of writing the research papers. They also pointed out that the structure would help them to find individual pieces of information and would support automated searches. In addition, six participants added comments expressing interest in structured abstracts and the results of the study.

However, participants also noted some disadvantages of structured abstracts in particular, that structured abstracts are not suitable for all types of study, may leave out important information, are long, and can be rather dull and boring. In addition, three participants added comments to the effect that there was no reason why conventional abstracts could not be as good as structured abstracts, and one suggested that the headings were unnecessary.

### 6.1 Comparison with Previous Studies

After generating the structured forms for the 25 abstracts, we performed an initial assessment of these by scoring them ourselves as well as gathering some other

**Table 17** Number of negative comments in each category by preference

Category	Preference			Total
	C	N	S	
Inappropriate	0	3	9	12
Over-restrictive	3	5	4	12
Too long	0	4	8	12
Boring	1	0	2	4
Poor English	1	2	0	3
Difficult	0	1	1	2
Duplication	0	1	0	2
Misuse	0	1	1	2
Too detailed	0	0	2	2
Poor readability	0	0	2	2
Automatic rejection	0	0	1	1
No graphics	0	0	1	1
Uninformative	1	0	0	1
Useless headings	1	0	0	1
Total	7	17	32	56



basic measures (Budgen et al. 2007a). The mean length of the abstracts (in words) increased from 161 (SD=59.15) to 267.4 (SD=47.09). Our own assessment of clarity showed a mean increase of 1.44 (from 5.64 to 7.08), while for completeness there was an increase of 6.16 in the mean value (from 5.64 to 11.8).

Hartley and Benjamin (1998) reported a study in which the authors themselves re-wrote their own abstracts into a structured format. The 30 abstracts were then judged for completeness based on 22 questions about the content of the abstract. The abstracts were judged by one senior academic (Hartley) and two other participants drawn from a pool of 15 first and second year psychology students. In this study, the average score for the unstructured abstracts was 6.4 (29% of the total possible score) with a standard deviation of 2.8. The average score for the structured abstracts was 9.1 (42%) with a standard deviation of 2.6.

In another study, Hartley (2003) reported a study where he re-wrote the abstracts himself. In this study he used a completeness checklist comprising 14 questions and a subjective clarity evaluation on score of one to 10. For traditional format abstracts he obtained:

- Checklist score mean = 5.5 (39% of the total) and sd = 1.0
- Clarity rating mean = 6.2 and sd = 2.0

For structured format abstracts he obtained:

- Checklist score mean = 9.7 (69%), sd = 1.4
- Clarity rating mean = 7.4, sd = 2.0

A recent field study (Sharma and Harrison 2006), although confirming that structured abstracts included more relevant information indicated that actual improvements were somewhat less than those found in laboratory studies. Using a questionnaire with 29 items, Sharman and Harrison found an increase of 12.2% for completeness in abstracts from three journals that adopted structured abstracts and no improvement (i.e. a change of  $-0.01\%$ ) in abstracts from three journals that did not adopt structured abstracts.

These results are summarized in Table 18. These demonstrate that all studies showed an improvement in completeness but the improvement was less when the authors themselves wrote the structured abstracts.

## 6.2 Limitations

### 6.2.1 Construct Validity

This study measured completeness in terms of the number of “yes” responses to a questionnaire comprising 18 items and clarity as a numerical value in the range 1–10. Both measures are based on subjective assessments by the participants. In both cases the measures were chosen because they used the same measurement approach as other related studies. Furthermore, the completeness questionnaire was subject to an extensive validation exercise as described in Section 3.4. Nonetheless, subjective measures must be treated with some caution. Indeed, three participants commented that the questions were rather vague and difficult to answer with a simple *Yes* or *No*. However, problems with the questionnaire should not have caused any major bias in the results since they affect both conventional and structured abstracts equally.

**Table 18** Comparisons with other results

Study	Study type	Person who produced the structured abstract	Completeness % increase for structured abstracts	Clarity increase for structured abstracts
Hartley and Benjamin (1998)	Laboratory experiment	Authors	12.3	n/a
Hartley (2003)	Laboratory experiment	Hartley	30.0	1.2
Sharma and Harrison (2006)	Field experiment	Authors	12.2	n/a
Kitchenham et al. (2008)	Laboratory experiment	Research team members	34.2	1.44
This study	Laboratory experiment	Research team members	37.3	3.04

A more serious criticism was made about the qualitative data. Two participants felt the results were biased because the default answers were favorable to structured abstracts. This limits the validity of the data collected about the personal preference for abstract form. However, it should not affect the individual comments about what was liked and what was disliked.

### 6.2.2 Internal Validity

The experimental design ensured that participants were randomly assigned to abstracts (i.e. the materials) and to whether they saw a structured and conventional abstract first (i.e. the treatment order). This design should protect against selection bias and learning effects. We were also able to confirm that a variety of uncontrolled factors, i.e. knowledge of structured abstracts, software engineering role, preference for structured abstracts, did not bias our results.

However, our design did require that participants make an assessment of each abstract as an artifact in its own right, whereas comments from two participants suggested that they had made relative assessments: One participant commented that he/she felt the abstracts were not really comparable due to length differences; another said he/she would have scored differently if he has seen the structured version first. However, we also analysed the data using only the data collected about the first abstract viewed by each participant. This corresponds to a standard parallel experiment and the results confirmed that structured abstracts were better than unstructured abstracts with respect to both the clarity and completeness metrics—which was also the view expressed in many of the comments made by the participants.

### 6.2.3 Generalisability

The abstracts used in this study were a stratified random sample from 103 experiments and quasi-experiments identified by Dag Sjøberg and his colleagues and found in a survey of nine journals and three conference proceedings (Sjøberg et al. 2005) during the period 1993–2002. The results should therefore generalize to the sample

frame. Furthermore, the selected journals and conferences are among the leading software engineering journals and conferences, so if anything we would expect the conventional abstracts used in this study to be of a better quality than abstracts in general.

Our study participants constitute a convenience sample not a random sample, so their opinions about structured and conventional abstracts do not automatically generalize to other software engineers and researchers. However, our demographic information indicates that we have a reasonable number of very experienced academics and researchers as well as a number of relatively inexperienced researchers. Thus we have a selection of informed opinion from the groups that would be expected to make most use of journal abstracts.

## 7 Conclusion

Our results demonstrate that conventional abstracts, even those in leading journals and conferences, often omit large amounts of relevant information. We believe this provides a strong argument that the quality of abstracts needs to be improved.

Our results also confirm that structured abstracts can provide an effective method for addressing the limitations of conventional abstracts. They include more relevant information and are easier to read than conventional abstracts. These results are consistent with the results of studies in other disciplines. However, we do not claim that improvements in completeness and clarity are due simply to restructuring the abstracts; our results clearly show that improvements are also due to including more information. Furthermore, results in other disciplines show that improvements are likely to be less when authors themselves construct abstracts. This raises the question as to whether authors should be advised to use structured abstracts or simply increase the amount of relevant information in their abstracts.

We agree with those participants who pointed out that conventional abstracts can be made as good as structured abstracts; however, we believe this is unlikely to occur without providing novice researchers with a better idea about what they need to include in abstracts. Therefore, since structured abstracts provide that basic discipline to help novice researchers and also provide a means to assist searching for information, we would argue that their use provides a viable approach to improving the quality of abstracts. Furthermore, it is evident that many of the experiment's participants also take this view. In addition, our previous study (Kitchenham et al. 2008) used abstracts rewritten by novices who were given only limited training. This provides evidence that non-expert researchers can successfully construct structured abstracts.

The price of adoption is also a small one, with the principal 'cost' associated with the use of structured abstracts being that they are longer than conventional ones. However, since this mainly arises because they contain more information, this is not an unreasonable consequence. Also, since the structure we have employed is intended for use with empirical papers, in particular software experiments, there is also the need to devise an appropriate set of variations to the headings in order to meet the needs of other forms of paper, such as those concerned with methodology, or software tools.

Nevertheless, since our results concur well with the experiences from other disciplines where the use of structured abstracts is now considered to be good practice,

we would strongly encourage their adoption by software engineering journals and conferences. By doing so, they will be able to provide better quality summaries of articles for their readers, as well as enhancing the quality of literature surveys, when performed by both researchers and students, by making it easier for them to identify those articles that contain relevant information.

**Acknowledgements** This work was supported by an award from the U.K.'s Engineering and Physical Sciences Research Council (EPSRC). The authors would also like to thank Dag Sjøberg for providing a random selection of papers to use in this study; all those who helped by participating in the study as 'judges'; John Bailey who organised the data collection; and Professor Jim Hartley of Keele University for his advice and guidance.

## Appendix 1 Procedures for Rewriting of Abstracts into Structured Form

The process to be followed was made as prescriptive as possible, so that all of the editors were following the same procedures. A full description is provided in the study protocol, here we give the basic outlines and describe some of the conventions employed.

### A.1 The Rewriting Process

The process was organised as the following sequence of steps (these have been left in 'directive' form).

1. First complete a paper copy of the evaluation form. Then rewrite the material from the existing abstract into a structured form as completely as possible. Keep a copy of this initial rewrite for later use in counting words under different headings. Each heading should begin on a new line, but please do not use white space between headings, the abstract should be a continuous sequence of text.
2. Where the entries for headings are incomplete, seek additional material from the paper. If you do so, please keep a note of:
  - (a) Where in the paper you found the necessary information.
  - (b) What information was still missing at the end of this process.

Rewrite the abstract using the additional material. This version should then be checked by the designated team member, and any suggested changes need to be agreed, edited and recorded.

3. If the original authors respond with suggested changes, the abstract may need to be further revised. Note that only material available in the original paper should be included and a record should be kept of what is done about each of the suggested changes.

As general guidelines, editors were asked to constrain the abstract to having no more than two sentences for each heading (with the possible exception of the **Results** heading); to try to keep to an overall limit of 300 words; and to reuse the original wording wherever possible.

## A.2 Organisation of Material

With multiple editors involved we needed to use a common file naming convention for all of the files involved, as well as a common means of documenting the changes we made. The basic structure that we employed for this was:

```
<first-author>-<year><year-no>-<index>
```

So, an example filename might be:

```
kitchenham-2002c-0.doc
```

Subsequent versions of the file then replaced the index value of '0' with the following values:

- 1 to designate the first structured version produced in *Step 1*. Any sentences omitted were stored in a file with a value of **1x** as index.
- 2a for the edited abstract using material from the paper if necessary.
- 2b for the version agreed after internal review.
- 3 as revised after feedback from the original authors.

We also kept a note of where in the paper any material was extracted to augment the information in the existing abstract.

## Appendix 2 Completeness Questions used in Evaluating Abstracts

The questions used to judge each abstract are listed below.

1. Is the rationale for the study reported?
2. Is the aim/purpose of the study reported?
3. Is a hypothesis (or hypotheses) provided?
4. Is there any indication of where this study took place? (E.g. in industry or academia, what the application domain was, etc.)
5. Is the number of participants reported?
6. Are the types of participants (e.g. students) reported?
7. Is any information about the experience of the participants reported?
8. Is the skill level of the participants described?
9. Is there any description of how the study was performed?
10. Does it report how the participants were allocated to different tasks or conditions?
11. Is the way that the data was collected reported?
12. Is there any description of the form of analysis performed?
13. Are the main results summarised in the abstract?
14. Are actual numbers from the results presented in the abstract?
15. Is any statistical information provided about the results?
16. Are any conclusions drawn?
17. Are any limitations of the study identified?
18. Is there any discussion of required future research?

### Appendix 3 Demographic and Qualitative Data Questions

The following questions provided the ‘third page’ seen by participants after they had completed their judging of the two abstracts.

1. Did you have any knowledge about structured abstracts before taking part in this study? (yes/no)  
If your answer was ‘yes’, then please indicate the nature of your knowledge:
  - (a) Heard about them, but not seen them before: (yes/no)
  - (b) Read papers about their use: (yes/no)
  - (c) Read papers with structured abstracts: (yes/no)
  - (d) Created structured abstracts for your own papers: (yes/no)
2. Please report up to three things that you *like* about structured abstracts (if there is nothing that you like, please leave blank).
3. Please report up to three things that you *dislike* about structured abstracts (if there is nothing that you dislike, please leave blank).
4. Overall, do you prefer structured or conventional abstracts?
  - (a) Prefer structured abstracts
  - (b) Prefer conventional abstracts
  - (c) No preference
5. Please indicate which description fits you best:
  - (a) Full-time researcher
  - (b) Practitioner
  - (c) Post-graduate Research Student
  - (d) Post-graduate Student
  - (e) Undergraduate
  - (f) Other (please specify)
6. Please indicate years of experience of software engineering research or practice.
7. Any other comments?

### References

- Automated Readability Index (2006) In: Wikipedia, the free encyclopedia. Retrieved 29 September 2006 URL:[http://en.wikipedia.org/w/index.php?title=Automated\\_Readability\\_Index](http://en.wikipedia.org/w/index.php?title=Automated_Readability_Index)
- Bayley L, Eldredge J (2003) The structured abstract: an essential tool for researchers. *Hypothesis* 17(1):11–13
- Booth A (2003) Bridging the research–practice gap? The role of evidence based librarianship. *New Review of Information and Library Research*, pp 3–23
- Brereton O, Kitchenham B, Budgen D, Turner M, Khalil M (2007) Lessons from applying the systematic literature review process within the software engineering domain. *J Syst Softw* 80(4): 571–583
- Budgen D, Kitchenham B, Charters S, Turner M, Brereton P, Linkman S (2007a) Preliminary results of a study of the completeness and clarity of structured abstracts. In: *EASE 2007: evaluation and assessment in software engineering*. BCS-eWiC, pp 64–72
- Budgen D, Kitchenham B, Charters S, Turner M, Brereton P, Linkman S (2007b) Protocol for an experimental study of the use of structured abstracts. Technical report, EBSE Project. Version 1.3.2.
- Editorial (2004) Addressing the limitations of structured abstracts. *Ann Intern Med* 140:480–481

- Flesch R (1948) A new readability yardstick. *J Appl Psychol* 32:221–233
- Flesch-Kincaid Readability Test (2006) In: Wikipedia, the free encyclopedia. URL:[http://en.wikipedia.org/w/index.php?title=Flesch-Kincaid\\_Readability\\_Test&oldid=77211134](http://en.wikipedia.org/w/index.php?title=Flesch-Kincaid_Readability_Test&oldid=77211134). Retrieved 29 September 2006
- Glasser B, Strauss A (1967) *The discovery of grounded theory*. Aldine, Chicago
- Hartley J (2000) Typographic settings for structured abstracts. *J Tech Writ* 30(4):355–365
- Hartley J (2003) Improving the clarity of journal abstracts in psychology: the case for structure. *Sci Commun* 24:366–379
- Hartley J (2004) Current findings from research on structured abstracts. *J Med Libr Assoc* 92: 368–371
- Hartley J, Sydes M (1996) Which layout do you prefer? An analysis of readers' preferences for different typographical layouts of structured abstracts. *J Inf Sci* 22(1):27–37
- Hartley J, Sydes M (1997) Are structured abstracts easier to read than traditional ones? *J Res Read* 20:122–136
- Hartley J, Benjamin M (1998) An evaluation of structured abstracts in journals published by the British Psychological Society. *Br J Educ Psychol* 68:443–456
- Jedlitschka A, Pfahl D (2005) Reporting guidelines for controlled experiments in software engineering. In: *Proc. ACM/IEEE international symposium on empirical software engineering (ISESE) 2005*. IEEE Computer Society Press, pp 95–195
- Jedlitschka A, Ciolkowski M, Pfahl D (2008) Reporting experiments in software engineering. In: Shull F, Singer J, Sjøberg D (eds) *Guide to advanced empirical software engineering*, Chapter 8. Springer, London
- Kitchenham B (2004) Procedures for undertaking systematic reviews. Technical Report TR/SE-0401. Department of Computer Science, Keele University and National ICT, Australia Ltd. Joint Technical Report
- Kitchenham B, Dybå T, Jørgensen M (2004) Evidence-based software engineering. In: *Proceedings of ICSE 2004*. IEEE Computer Society Press, pp 273–281
- Kitchenham B, Budgen D, Brereton P, Turner M (2007) 2nd international workshop on realising evidence-based software engineering (REBSE-2): overview and introduction. In: *Proceedings of REBSE-2 workshop, ICSE 2007*. IEEE Computer Society Press, pp 1–5
- Kitchenham B, Brereton P, Owen S, Butcher J, Jefferies C (2008) Length and readability of structured software engineering abstracts. *IET Softw* 2:37–45
- Kitchenham B, Pflieger SL, Pickard L, Jones P, Hoaglin D, Emam KE, Rosenberg J (2002) Preliminary guidelines for empirical research in software engineering. *IEEE Trans Softw Eng* 28: 721–734
- Kitchenham B, Al-Khilidar H, Babar MA, Berry M, Cox K, Keung J, Kurniawati F, Staples M, Zhang H, Zhu L (2006) Evaluating guidelines for empirical software engineering studies. In: *Proceedings ACM/IEEE international symposium on empirical software engineering (ISESE 2006)*. IEEE Computer Society Press
- Milliken G, Johnson D (1992) *Analysis of messy data—volume 1: designed experiments*. Chapman and Hall
- Petticrew M, Roberts H (2006) *Systematic reviews in the social sciences: a practical guide*. Blackwell
- Senn S (2002) *Cross-over trials in clinical research*, 2nd edn. Wiley
- Sharma S, Harrison JE (2006) Structured abstracts: do they improve the quality of information in abstracts? *Am J Orthod Dentofac Orthop* 130(4):523–530
- Sjøberg D, Hannay J, Hansen O, Kampenes V, Karahasanovic A, Liborg N, Rekdal A (2005) A survey of controlled experiments in software engineering. *IEEE Trans Softw Eng* 31(9):733–753
- Webster J, Watson R (2002) Analysing the past to prepare for the future: writing a literature review. *MIS Quarterly* 26, xiii–xxiii



**David Budgen** is a Professor of Software Engineering and Chairman of the Department of Computer Science at Durham University in the UK. His research interests include software design, design environments, healthcare computing and evidence-based software engineering. He was awarded a BSc(Hons) in Physics and a PhD in Theoretical Physics from Durham University, following which he worked as a research scientist for the Admiralty and then held academic positions at Stirling University and Keele University before moving to his present post at Durham University in 2005. He is a member of the IEEE Computer Society, the ACM and the Institution of Engineering & Technology (IET).



**Barbara A. Kitchenham** is Professor of Quantitative Software Engineering at Keele University in the UK. From 2004–2007, she was a Senior Principal Researcher at National ICT Australia. She has worked in software engineering for nearly 30 years both in industry and academia. Her main research interest is software measurement and its application to project management, quality control, risk management and evaluation of software technologies. Her most recent research has focused on the application of evidence-based practice to software engineering. She is a Chartered Mathematician and Fellow of the Institute of Mathematics and Its Applications, a Fellow of the Royal Statistical Society and a member of the IEEE Computer Society.





**Stuart M. Charters** is a Lecturer of Software and Information Technology in the Applied Computing Group, Lincoln University, NZ. Stuart received his BSc(Hons) in Computer Science and PhD in Computer Science from Durham University UK. His research interests include evidence-based software engineering, software visualisation and grid computing.



**Mark Turner** is a Lecturer in the School of Computing and Mathematics at Keele University, UK. His research interests include evidence-based software engineering, service-based software engineering and dynamic access control. Turner received a PhD in computer science from Keele University. He is a member of the IEEE Computer Society and the British Computer Society.



**Pearl Brereton** is Professor of Software Engineering in the School of Computing and Mathematics at Keele University. She was awarded a BSc degree (first class honours) in Applied Mathematics and Computer Science from Sheffield University and a PhD in Computer Science from Keele University. Her research focuses on evidence-based software engineering and service-oriented systems. She is a member of the IEEE Computer Society, the ACM, and the British Computer Society.



**Stephen G. Linkman** is a Senior Lecturer in the School of Computing and Mathematics at Keele University and holds an MSc from the University of Leicester. His main research interests lie in the fields of software metrics and their application to project management, quality control, risk management and the evaluation of software systems and process. He is a visiting Professor at the University of Sao Paulo in Brazil.