Teaching psychology students to write structured abstracts: An evaluation study

James Hartley, Judy Rock & Claire Fox, Keele University

Background: *Considerable evidence suggests that structured abstracts in scientific journal articles are more informative than traditional ones, but no one (to our knowledge) has written about asking psychology undergraduates to write structured abstracts for their laboratory reports.*

Aim: *Our aim was to assess whether or not the quality of such abstracts would be improved if such students were required to write structured abstracts rather than traditional ones.*

Sample & Method: *117 first-year psychology students at Keele wrote traditional abstracts for their laboratory reports during the academic session 1998–1999, and 122 wrote structured abstracts during the session 1999–2000. We compared 50 abstracts from each of these groups.*

Results. *We found that the structured abstracts were significantly longer (M = 217 vs 181 words, t = 3.23, p<0.005 one-tail) but that they did not contain significantly more key information than the traditional ones, as measured with an information checklist (M = 9.2 vs 8.6, t = 1.52). Nor was there any significant difference between them as measured with the Flesch Reading Ease formula (M = 40 vs 37, t = 1.36). However, independent raters, in two separate studies, rated the quality of the structured abstracts to be significantly higher than that of the traditional ones (M = 6.9 vs 6.1, t = 3.36, p<0.005 and 7.1 vs 6.0, t = 5.20, p<0.001 respectively).*

Conclusions: *These results give only partial support to the notion that teaching students to write structured abstracts will lead to improved abstracts.*

READERS OF THIS article may have already noticed that it has a 'structured' abstract. Instead of the traditional single 'block-paragraph' the abstract is organised under several sub-headings. Such structured abstracts have now replaced traditional abstracts in many medical research journals, and they are now beginning to appear in other disciplines such as psychiatry (e.g. *British Journal of Psychiatry*), non-medical science journals (e.g. *Bioinformatics, Fruits, Pharmaceutical Research*) and psychology (e.g. *British Journal of Clinical Psychology, British Journal of Educational Psychology, British Journal of Health Psychology, Legal and Criminological Psychology* and (sporadically) in the *British Journal of Medical Psychology*. Structured abstracts are also requested now by the British Psychological Society for judging conference submissions in place of the previously required four-page summary.

An overview of 31 studies of structured abstracts in medical and psychology journals suggests (Hartley, 2004) that they:

- contain more information (11 out of 12 studies);
- are easier to search and to read (five studies) – although two studies have queried this;
- are possibly easier to recall (one report of a series of studies);
- facilitate peer review for conference proceedings (three studies); and
- are generally welcomed by readers and by authors (six studies).

However, there have been some qualifications. Structured abstracts:

- take up more space (11 studies);
- sometimes have confusing typographic layouts (two studies); and
- may be prone to the same sorts of errors of omission and distortion that occur in traditional abstracts (eight studies).

Several authors have suggested why structured abstracts are so popular. In our view this is because they force authors to be more explicit about their content, and to do this in a systematic way. With structured abstracts it is difficult to leave out any key information and to vary the sequence of the information presented. In addition, the spatial arrangement of the text displays its underlying structure and this enhances readability (Hartley, 1994, 2000; Hartley & Sydes, 1997).

In the light of these research findings it seems that it might be useful to teach psychology students how to write structured abstracts for their laboratory reports. To our knowledge, although some colleagues have instituted the practice, no-one has as yet reported any actual findings on the topic. In this study we present the results that we obtained when we compared traditional abstracts written by one cohort of first-year psychology students with structured abstracts written by the following cohort. We predicted, on the basis of the research described above, that the structured abstracts would: (i) be longer; (ii) contain more information; (iii) be more readable; and (iv) be of a higher quality than the traditional ones.

Method

The first author of this paper proposed to the Keele psychology department towards the end of the 1998–1999 academic session that we should introduce structured abstracts into the laboratory reports of the 1999–2000 session, and this was agreed. This involved making changes to the departmental booklet on how to write a laboratory report provided to the students each year, and to one of the sessions in the laboratory classes.

First-year psychology students at Keele each write four laboratory reports during the academic session. (They are dual-honours students, so they may well do four more other pieces of written work in their other subject.) During the 1998–1999 session students were introduced to the conventions of report writing in their second laboratory class on research design held in the first semester. As a part of this class they carried out an exercise on the comprehension of traditional abstracts in order to introduce them to the elements that should be included. During the 1999–2000 session this component of the class was replaced with one on structured abstracts, and the students' attention was drawn to the relevant information in the departmental booklet. In this class the students worked in groups of four to re-write two traditional abstracts (from a choice of six) in a structured form, and then to evaluate other groups' efforts. In both 1998–1999 and 1999–2000 the students worked in groups on their last research topic in the second semester (in May) but they wrote up their reports individually.

One-hundred-and-seventeen laboratory reports were handed in for marking at the end of the 1998–1999 session. All but five of these were word-processed. One-hundred-and-twenty-two laboratory reports were handed in for marking at the end of the 1999–2000 session. All but six of these were word-processed. One report in this latter group did not contain an abstract.

The laboratory reports were marked by graduate teaching assistants in both sessions, with the marking moderated by the same member of staff in each year. Following moderation, and before the reports were returned to the students, we asked all the students by e-mail if they had any objections to their abstracts being used in our planned study. No student objected, and one or two specifically gave their permission. Accordingly, we photocopied all of the abstracts written for the final report in each of the sessions 1998–1999 and 1999–2000.

We then selected 50 of the traditional abstracts and 50 of the structured ones for detailed analysis. The basis of our choice was that two of the five graduate teaching assistants involved had marked the laboratory reports for both sessions and it seemed appropriate to use these abstracts. Twenty-five abstracts were taken from each marker

on each of the two sessions, making 50 traditional and 50 structured ones. As each marker had marked slightly more than 25 reports in each of their marking sessions, we included only abstracts written by three or more students on the same topic, thus excluding any abstract on a single topic (of which there were few).

We scanned each abstract into a computer and then, to ensure comparability between them for any subsequent computer-based analyses, we re-formatted each one, using the same typeface, type-size, inter-line length, spacing and unjustified text to avoid end-of-line-hyphenation. We also standardised certain other features of the texts so as not to affect differentially the computer-based readability software. Thus, for example, we always printed words and phrases such as 'self-esteem', and 'first-year students' with hyphens, and words such as 'short term memory' and 'context dependent' without them. We also standardised throughout the spacing of any statistical phrasing (e.g. $t = 1.96$, d.f.19, $p<0.05$). Finally, we corrected any gross grammatical errors or spellings (e.g. we changed, 'These findings show that the research hypothesis can be excepted' to 'These findings show that the research hypothesis can be accepted').

We then used Microsoft's *Office '97* software to carry out computer-based readability measures that assessed the overall lengths of the abstracts, their average sentence lengths, their percentages of passive sentences, and their Flesch Reading Ease scores. (We deleted the sub-headings from the structured abstracts in order not to increase the computer-based readability scores by having such one-word 'sentences'.)

In order to assess the content of the abstracts we slightly revised the checklist previously used for this purpose by Hartley and Benjamin (1998) in their study of abstracts in journals published by the British Psychological Society (see Appendix). This revised checklist was completed for all of the abstracts by the first author and, independently, by the third one. Both assessors wrote on each student's abstract the numbers corresponding to each of the checklist items where they thought that these items were present. They then met to agree between them an overall information score for each abstract that was arrived at by summing the number of agreed items marked, and resolving any discrepancies. The original checklist was also changed slightly to include, at the end, a space for making an overall rating (out of 10) for the quality of each abstract. The quality ratings of the two independent assessors were then averaged (without agreement being sought). In addition, two other assessors – experienced members of staff – also rated each of the 100 abstracts for quality without seeing or using the checklist. These ratings were also averaged (without agreement being sought) to arrive at a second quality score that was made independently of the checklist.

Results

Table 1 shows the results that we obtained. It can be seen that, with the computer-based measures, the structured abstracts were significantly longer than the traditional ones on two counts – the mean total number of words written (M = 217 vs 181, $t = 3.23$, $p<0.005$ one-tail), and the mean average number of sentences (M = 10.7 vs 9.3, $t = 2.43$, $p<0.01$ one-tail). There were no significant differences between the mean percentages of passive sentences (M = 35 and 38, $t = 0.65$) and between the mean Flesch readability scores (M = 40 and 37, $t = 1.36$). The data from the information checklist suggested that the structured abstracts contained more information than did the traditional abstracts (M = 9.2 vs 8.6), but this difference was not statistically significant ($t = 1.52$). The ratings of the quality of the abstracts made on the checklist were significantly higher for the structured abstracts than they were for the traditional ones (M = 6.9 vs 6.1, $t = 3.36$, $p<0.005$, one-tail). This was also true of the ratings made independently of the checklist (M = 7.1 vs 6.0, $t = 5.20$, $p<0.001$, one-tail).

Correlations between the scores obtained on the main measures are shown in Table 2.

Table 1: The mean scores (and standard deviations) obtained on the various measures for the 50 traditional and the 50 structured abstracts.

		Traditional abstracts	Structured abstracts	*t* value
No. of words	M s.d.	181 53	217 55	3.23***
No. of sentences	M s.d.	9.3 3.1	10.7 2.9	2.43**
% of passive sentences	M s.d.	38 18	35 19	0.65
Flesch Reading Ease score[1]	M s.d.	37 12	40 10	1.36
Checklist information score[2]	M s.d.	8.6 1.9	9.2 2.0	1.52
Checklist quality rating[3]	M s.d.	6.1 1.1	6.9 1.1	3.36***
Independent quality rating[3]	M s.d.	6.0 1.1	7.1 1.0	5.20****

*p<0.05 one-tail test; **p<0.01 one-tail test; ***p<0.005 one-tail test; ****p<0.001 one-tail test

[1] Flesch Reading Ease scores range from 0–100. The higher the score the more readable the text (see Hartley, 1994, for more details).

[2] The maximum score possible was 17 per abstract – see Appendix.

[3] The maximum rating possible was 10 per abstract.

Table 2: Intercorrelations between the main measures used in this study.

Measures	Traditional abstracts	Structured abstracts
Information content and quality ratings[1]	0.63***	0.65***
Information content and length	0.34*	0.44**
Quality ratings[1] and length	0.27	0.48***
Quality ratings[1] and readability	-0.33*	-0.30*

[1]Average of four ratings of quality

*p<0.05; **p<0.01; ***p<0.001 (two-tail tests).

Here it can be seen that the ratings of quality (here pooled across the four sets of data) correlated significantly with the information content scores. Of greater interest, perhaps, were the findings that the correlations between information content and quality, and information content and length were greater for the structured than they were for the traditional abstracts.

Discussion

These results, whilst clear, did not fully support our original hypotheses. The structured abstracts were significantly longer but they did not contain significantly more key information. Nor were the structured abstracts significantly more readable, as predicted, but they were rated as being of a significantly higher quality on both of the separate quality measures.

Changes in length and in information content

The findings of this study supported the hypothesis that the structured abstracts would be longer. The structured abstracts were, on average, 20 per cent longer than the traditional ones in terms of the number of words used but, as shown in Table 1, this actually translated into a difference of only one or two sentences. It is possible, of course, that this difference in length might simply reflect the fact that the instructions to the students in their booklets differed slightly with respect to length. The 1998–1999 booklet told the students that the abstract 'must be no longer than 200 words', whereas the 1999–2000 one said that the abstract 'is usually no longer than 200 words.'

With the extra text we imagined that there might be extra key information, and this is supported to some extent by the higher correlation between information content and length for the structured abstracts shown in Table 2. We had anticipated that, by comparing the checklists, we would be able to see where the structured abstracts differed from the traditional ones in terms of what extra information was included. However, this did not turn out to be realistic, since the differences between the checklist scores did not appear to differ much in terms of their means and distributions. One problem here lies with the fact that the checklist scores can under-represent the amount of extra information available in the structured abstracts. As noted before (Hartley & Benjamin, 1998), a brief or a detailed mention of an item on the checklist would ensure an appropriate tick, regardless of the amount of detail. Thus, for example, a sentence to the effect that, 'This experiment was a modification of one carried out by X and Y (1994)' would obtain a check for background information, but so too would this one, 'Stress is a common part of everyday life. Many researchers believe that stress in some people leads to depression because, instead of coping with the stress, they become overwhelmed. Different people cope with different stressors differently and past research has shown that students in different years of study have different kinds of stressors and reactions.'

Generally speaking there was little disagreement between the two assessors using the checklist, although certain items had to be clarified. Sometimes, for example, we found it difficult to distinguish between 'conclusions' and 'implications' (see the footnote to the Appendix.) And, although the average checklist score of 9 out of a possible 17 may seem low it is actually quite high. Hartley and Benjamin (1998), for example, reported an average score of 6.4 (out of a possible 22) for 30 traditional abstracts and 9.1 for 30 structured abstracts published in the *British Journal of Educational Psychology*. (We estimate that, if readers apply the checklist to the abstract for this particular paper, it should score 13 – as our abstract gives no information on the sex distribution of the participants, mentions no limitations of the study, and does not include any suggestions for further research or indicate that any such suggestions will be discussed in the paper.)

Thus the current checklist is a blunt instrument. It suffers from the fact that each item is given an equal weighting when clearly

some items are more important than others. Some more detailed suggestions for refining the checklist for future use are given in the footnote to the Appendix.

It is possible, of course, that many of the structured abstracts that the students wrote were too long – perhaps about a half of them. Part of the art of abstracting lies in being succinct, so the lengthy piece quoted above is actually inappropriate in an abstract. Thus one implication of our findings is that tutors need to pay more care to ensuring that their students write abstracts within the stipulated word length of 200 words or less.

Changes in readability

The findings of this study did not support our hypothesis that the structured abstracts would be more readable and, indeed, the correlations shown in Table 2 between the readability scores and the quality ratings were negative for both forms of the abstracts! We had anticipated that writing guided by subheadings would clarify the task, and make writing simpler. However, it appeared that, although the means were in the predicted direction, this might have arisen by chance. It is of interest therefore, to note, in this connection, that differences favouring the readability of structured over traditional abstracts have only been found in studies where the same authors have written both versions of the abstracts (Hartley, 2004).

Some readers, of course, might question the value of using a readability formula to assess the readability of abstracts. Indeed, there are several reasons for accepting this view, particularly as such formulae were not devised for measuring such complex text, and they only measure surface features of the text as opposed to deeper linguistic structures. Nonetheless, such measures do supply objective procedures that can be applied consistently to different texts – provided that the same computer package is applied (Mailloux *et al.*, 1995; Sydes & Hartley, 1997). More sophisticated computer-based measures of text structures have now become available since this enquiry was carried out (e.g. see Grasser *et al.*, 2004).

The Flesch readability score (for the 100 abstracts written by the students) averaged about 39. Such a score suggests that these abstracts were written at an appropriate level for undergraduate readers (see Hartley, 1994) although the raters perhaps thought that they were insufficiently academic. This figure compares well with those given for abstracts in other psychology journal articles. For example, Tenopir and Jacso (1993) reported a mean Flesch score of 19 for 300 abstracts taken from journals published by the American Psychological Association in the early 1990s; Hartley (2003) reported a mean score of 21 for 24 abstracts taken from the *Journal of Educational Psychology* in 1992; and Hartley and Benjamin (1998) reported a mean score of 21 for 30 traditional abstracts and 35 for 30 structured abstracts taken from the *British Journal of Educational Psychology* in 1997. (The abstract to this present paper has a Flesch score of 57.)

Furthermore, the mean percent of passive sentences in our students' abstracts was similar in both formats – of the order of 36 per cent. This figure compares well with the average of 32 per cent reported for the abstracts of articles in the *Journal of Educational Psychology* (Hartley, 2003), and with the 33 per cent for excerpts from the introductions to highly influential articles in psychology (Hartley, Sotto & Pennebaker, 2002).

Perceived changes in quality

The findings of this study supported the hypothesis that structured abstracts would be of a perceived higher quality than traditional ones. As noted above, we used two measures to assess this. The first required the two people completing the abstract checklist to give a rating out of 10 for each abstract as they completed each form. Then these two ratings were averaged, and the means of the averaged scores are those shown in Table 1. The second measure required two people to rate the quality of the abstracts by marking

them out of 10 without seeing the checklist. Again we averaged their two ratings and their mean average scores are also presented in Table 1. We used these two approaches because we anticipated that using the checklist might affect the judgements of quality made in the first comparison. However, it appears from the data shown in Table 1 that using the checklist had little effect upon these holistic ratings.

The correlation for the quality ratings between the two assessors using the checklist was 0.43 for the structured and 0.39 for the traditional abstracts (0.44 overall). Similarly, the correlations for the quality ratings between the two assessors without the checklist was 0.47 for the structured and 0.42 for the traditional abstracts (0.50 overall). We anticipated that these inter-correlations might be lower for the traditional abstracts than for the structured ones as this would support the notion that traditional abstracts are harder to assess than structured ones (because of their less structured format). However, these differences were not significant.

The ratings of quality are of interest because they suggest that differences between structured and traditional abstracts appear more marked when less quantitative methods of measurement are used. The significant difference between the mean ratings reported in Table 1 may, of course, have resulted from a combination of factors. It appears from the correlations shown in Table 2 that the information content was important but so, too, was the length of the abstracts. In addition, the results may have been affected by the fact that structured abstracts look – and are – easier to follow because of their standard format (Hartley & Sydes, 1997; Hartley, Sydes & Blurton, 1996).

To some extent the correlations reported in Table 2 parallel ones that have been found in other studies. James (1976), for example, found a correlation of 0.46 between the lengths of assignments and their grades at the Open University, and Norton and Hartley (1986) similarly found correlations of over 0.60 between the length of answers in an examination paper and the grades awarded. Furthermore, although the correlations between the raters assessing the quality of the abstracts were low, they were in fact higher than the correlations typically found between when different markers assess (handwritten) essays (e.g. see Hartog & Rhodes, 1936; Marshall & Powers, 1969; Branthwaite, Trueman & Berrisford, 1981; Newstead, 1996). Thus the correlations reported in this study lead us to the conclusion that the various factors that contribute to grading quality are actually confounded in 'real-life' studies such as this one. It would be interesting, but of less practical importance, to ask judges to make quality ratings of the two types of abstracts when they have been standardised for length and style of presentation.

Concluding remarks

In this study we have shown that there are certain advantages to having students write structured as opposed to traditional abstracts for their laboratory reports. The structured abstracts were longer and were rated by experienced assessors as being of a higher quality than were the traditional ones. These overall results, however, although encouraging, were not quite as pleasing as we had anticipated, partly because perhaps we were expecting too much. After all, our students had only written three abstracts throughout the year before writing the ones that we assessed. But current research on academic writing would suggest that it takes a long time and much practice to acquire the necessary skills of writing in a particular genre (Lea & Street, 2000; Lillis, 2001; Richardson, 2004). Nonetheless, there were some gains, and there were no losses. Practice in writing structured abstracts may, therefore, help students become more proficient academic writers in the future.

Correspondence

Professor James Hartley
School of Psychology,
Keele University, Staffordshire ST5 5BG.
E-mail: j.hartley@psy.keele.ac.uk

References

American Psychological Association (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington: APA.

Branthwaite, A., Trueman, M. & Berrisford, T. (1981). Unreliability of marking: Further evidence and a possible explanation. *Education Review, 33*(1), 41–46.

Graesser, A.C., MaNamara, D.S., Louwerse, M.M. & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behaviour Research Methods, Instruments & Computers, 36*(2), 193–202.

Hartley, J. (1994). *Designing instructional text* (3rd ed.). London: Kogan Page.

Hartley, J. (1999). Applying ergonomics to applied ergonomics: Using structured abstracts. *Applied Ergonomics, 30*, 535–541.

Hartley, J. (2000). Typographic settings for structured abstracts. *Journal of Technical Writing and Communication, 30*(4), 355–365.

Hartley, J. (2003). Improving the clarity of journal abstracts in psychology. *Science Communication, 24*(3), 366–379.

Hartley, J. (2004). Current findings from research on structured abstracts. *Journal of the Medical Libraries Association, 92*(3), 368–371.

Hartley, J. & Benjamin, M. (1998). An evaluation of structured abstracts in journals published by the British Psychological Society. *British Journal of Educational Psychology, 68*, 443–456.

Hartley, J., Sotto, E. & Pennebaker J. (2002). Style and substance in psychology: Are influential articles more readable than less influential ones? *Social Studies of Science, 32*(2), 321–334.

Hartley, J. & Sydes, M. (1996). Which layout do you prefer? An analysis of readers' preferences for different typographic layouts of structured abstracts. *Journal of Information Science, 22*(1), 27–37.

Hartley, J. & Sydes, M. (1997). Are structured abstracts easier to read than traditional ones? *Journal of Research in Reading, 20*(2), 122–136.

Hartley, J., Sydes, M. & Blurton, A. (1996). Obtaining information accurately and quickly: Are structured abstracts more efficient? *Journal of Information Science, 22*(5), 349–356.

Hartog, P. & Rhodes, E. C. (1936). *The marks of examiners.* London: Macmillan.

James, A. (1976). Does the amount written on assignments bias the grades awarded? *Teaching at a Distance, 7*, 49–54.

Lea, M. & Street, B. (2000). Student writing and staff feedback in higher education: An academic literacies approach. In M.R. Lea & B. Steiner (Eds.), *Student writing in higher education.* Buckingham: Open University Press.

Lillis, T. M. (2001). *Student writing.* London: Routledge.

Mailloux, S.L., Johnson, M.E., Fisher, D.G. & Pettibone, T.J. (1995). How reliable is computerised assessment of readability? *Computers and Nursing, 13*(5), 221–225.

Marshall, J.G. & Powers, J.M. (1969). Writing neatness, composition errors and essay grades. *Journal of Educational Measurement, 6*(2), 97–101.

Newstead, S.E. (1996). The psychology of student assessment. *The Psychologist, 9*(12), 543–547.

Norton, L. & Hartley, J. (1986). What factors contribute to good examination marks? The role of notetaking in subsequent academic performance. *Higher Education, 15*, 355–371.

Richardson, P.W. (2004). Reading and writing from textbooks in higher education: A case study from Economics. *Studies in Higher Education, 29*(4), 505–521.

Sydes, M. & Hartley, J. (1997). A thorn in the Flesch: Observations on the unreliability of computer-based readability formulae. *British Journal of Educational Technology, 28(*2), 143–145.

Tenopir, C. & Jacso, P. (1993). Quality of abstracts. *Online, 17*(3), 44–55.

Appendix

The abstract checklist used in this study. (Respondents check only the features that are present in the abstract by marking the item number on the text, and the number of checks is summed.)

Abstract No. _____

1._____ Is anything said about previous research or research findings on the topic?

2._____ Is there an indication of what the aims/purpose of this study were?

3._____ Is there information on who the participants were?

4._____ Is there information on the numbers of participants?

5._____ Is there information on the sex distribution of the participants?

6._____ Is there information on the ages of the participants?

7._____ Is there information on how the participants were placed in different groups (if appropriate)?

8._____ Is there information on the measures used in the study?

9._____ Are the main results presented in prose in the abstract?

10.____ Are the results said to be (or not to be) statistically significant?

11.____ Are actual numbers (e.g. means/correlation coefficients/t values) given in the abstract?

12.____ Are *p* values reported for the results?

13.____ Are any conclusions drawn?

14.____ Are any limitations of the study mentioned?

15.____ Are any implications drawn from the findings?

16.____ Are suggestions for further research mentioned in the abstract?

17. ____ Does the abstract say that suggestions for further research are discussed in the report?

Overall rating out of 10 for this abstract: _____

Footnote: A limitation of this checklist is that all items are treated equally, where some are obviously more important than others, and some are not applicable to every abstract. However, the same method was used for scoring both sets of abstracts.

We had some difficulties using this checklist that might be overcome in a future study by making slight modifications. Thus for:

Item 6. We accepted this for students if the word 'first-year' or some year-group designation was given, despite the fact that ages vary within year groups.

Item 7. We did not accept 'a quota sample' as an indication of method of recruiting participants, and of course, item 7 is not applicable for studies where all of the participants do the same task(s).

Item 10. We accepted this if the results were followed by a *p* value, even if nothing was said in the text about the findings being significant.

Item 13. This could probably be re-written 'Are any conclusions/implications drawn?' and Item 15 deleted. Then items 16 and 17 could be collapsed into one item along the lines of, 'Does the abstract say anything about further research?'

In future it might be possible to weight some items as being more important than others – such as those that reflect the requirements of the American Psychological Association – see the APA's *Publication Manual* (APA, 1994, 8–11).