

Investigating Referees' Requirements in an Electronic Medium

P. Wright and A. Lickorish

MRC Applied Psychology Unit, Cambridge, England

The performance of eight academic refereeing two scholarly journal articles is examined when the articles are presented either on paper or on a CRT screen. Three aspects of performance are analysed: (a) details of how the task was undertaken by each referee, (b) individual assessments of the two presentation media, (c) the effect of presentation medium on the evaluation of each article's merit. The results showed that there was considerable variation among referees in the way they undertook the task and in their evaluation of each article. Referees tended to be slower with the CRT text and made several criticisms of the facilities available to them in the electronic medium. Suggestions were put forward concerning the support needed for refereeing academic papers presented in an electronic medium. In this study evaluative assessments of the merits of each article were not affected by the presentation medium.

Society is in the midst of transforming itself to use the power of computers throughout its entire fabric — wherever information is used — and that transformation depends critically upon the quality of human-computer interaction. — Card, Moran, and Newell, 1983, p.3.

1. Introduction

The impact of the new information technologies is changing the ways in which professional groups communicate with each other. Scholarly communications are no exception. Inglis (1983) suggests that the academic author has a great deal to gain from the new information technologies because "the tasks on which the academic author spends most of his time, editing and rewriting, are the very tasks which the computerised text editor is most able to assist."

A greater use of word processing facilities does not necessarily lead to a move away from print on paper as the primary source of communication among academics. Yet those who have an eye for looking into crystal balls suggest that, in time, the traditional printed journal may be replaced by a video disc, accessed from a microcomputer. The technology has not yet made this practicable, but the forecasts are that the economics both of production and storage could make this a very attractive option before too long (Blunden,

1981). In the mean time the growth of electronic mail and teleconferencing systems offers the academic community some potential advantages over postal and telephone services (Meadows, 1983). Perhaps the three most obvious advantages are: much faster long distance communications, potential savings in effort and therefore costs, and enhanced opportunities for extended interactions among readers and writers. These advantages make the electronic medium attractive as a means of communication not only between publishers and their readers but also between editors and those whose opinions they seek on the suitability of a manuscript for publication: the referees.

Whatever the final form of the publication, it seems likely that the new information technologies will play an increasing part in the preparation of material prior to publication. Sophisticated graphics packages allow alternative page makeups to be explored easily by editors (e.g., Walker, 1983; Davies Cooper and Cooper, 1984). There are optical character recognition systems which convert typescript into electronic form without further manual keying being necessary. Already, as Inglis has noted, many authors choose to prepare their texts on word processing systems. With the increased availability of electronic networks, such authors might welcome the opportunity to submit to the editor an electronic version of their text rather than a paper version. Naturally some topics may be more suitable for this than others. A journal article rich in photographic illustrations, particularly colour plates, would not at the moment fare very well in electronic transmission. But there are many other kinds of journal article where such factors do not arise. Even where they do, a parallel submission of text and figures may be feasible, with the text being sent electronically and the accompanying camera-ready figures transmitted on paper.

It is by no means a matter of trivial convenience to transmit text electronically. This method of communication can result in a sizable reduction in the time taken by the editorial review process. The normal sequence of interactions when an article is submitted to an academic journal is listed below. Each of the arrows indicates a communication delay, usually postal. Where the post is intercontinental this can easily be a delay of 10 days.

1. author ---> editor
2. editor ---> first referee
3. first referee ---> editor
4. editor ---> second referee
5. second referee ---> editor
6. editor ---> author

Even if the editor forwards the article to both referees at the same time, there will still be four waiting periods (perhaps a total of 40 days). This is essentially dead time during which nothing at all is done to further the course of the article. This dead time is even further increased if, as often happens,

authors are invited to revise and resubmit (Bradley, 1982). Electronic transmission removes these dead periods. In principle it is possible that all six steps, from author to editor and back to author, could be accomplished in a single day if the referees were able to deal with the article as soon as they received it.

Apart from speeding things up, the savings in cost and effort made by retaining the text in electronic form can also be considerable. Many journals require three copies of the paper to be submitted, some require six copies. Producing these multiple versions via a photo-copier takes time and, unless expensive equipment is available, laborious collation of the pages. The photocopying charge can itself be sizable for a lengthy paper. International airmail charges must then be added and it is clear that economic factors can make a good case for at least exploring some of the alternatives. The text file could be transmitted by telephone in a matter of minutes. The exact time would depend upon transmission rates, but at 1200 baud (120 characters per second) a 5000 word article would take only 4.2 minutes to send. Even on a transatlantic phone call this is cost effective, and there are systems available which will automatically send files from one computer to another over the telephone network during the night at cheap rates. There are no problems of collating pages and the recipient can make as many subsequent copies as are needed, or even transmit it further electronically.

The advantages discussed so far have been those which accrue to the author of an article or the editor of a journal. In the middle of this system are the referees. In principle they too may benefit from a change to electronic communication. If they wish to work with a paper copy then they are free to print out the file they have received, but in an electronic medium they have the option of combining the activities of reading and making comments. These comments can then be revised by the referees if they wish and returned to the editor without secretarial intervention, often a further source of lost time. Such a suggestion presupposes that the refereeing process can adequately take place in an electronic medium. This assumption may not be well founded. During the early stages of the British Library's experimental project on an electronic journal (BLEND, see Shackel, 1982) several people who were asked to referee an electronically transmitted article chose to do the task using a paper version of the text. The reasons for this are unclear. It might be that there is something about the electronic medium which makes it inappropriate for a task such as refereeing, or it may be that the activity of refereeing requires special support in an electronic medium. In order to examine these possibilities let us consider the requirements that referees have, and how these might be affected by a change from paper to video terminal.

2. Requirements of referees

2.1 Legibility. A basic requirement of referees, as for all readers, is that the text be legible. In this respect the readers of electronic text may often be at a disadvantage compared with those reading print on paper. There is a wide variability in the typographic quality within both media, so generalisations are not very meaningful, but at the present stage of technological development it is often the case that the electric typewriter produces more legible displays than many CRT screens.

2.2 Movement. A second general requirement of referees is the ability to move back and forth in the text, comparing the information in different sections. Here the paper medium has the advantage that the sheets can be spread out and several pages viewed concurrently. The limited size of the CRT screen prevents this. It is also a fact that for printed texts readers have acquired the skills of page turning, which may include leaving a finger in the place to which one will return, and can call upon these skills in relatively automatic fashion as the need for them arises. Comparable manoeuvres may be possible in an electronic medium, but their execution will be less familiar. As a consequence, the attention they require may detract from the primary task of following and evaluating the author's argument.

2.3 Annotation. A third requirement of referees relates to the ability to comment on the article as it is being read. Such comments may sometimes be written directly on the text itself or on other occasions they may be noted on a separate sheet. They are essentially private remarks which referees are addressing to themselves, in contrast to the considered opinions which will later be forwarded to the editor. From anecdotal information it would seem that these annotations often tend to reflect affective responses (such as agreement, disagreement, or puzzlement) rather than being well-articulated comments about the substance of the article. It is possible that these concurrent comments are an optional strategy which only some writers use, and perhaps not for all papers. But it is important to appreciate that when they occur they form part of the reading task rather than being related to the subsequent task of writing a report on the article.

2.4 Communication. A fourth requirement of referees is that they formulate an opinion about the text and communicate this to the editor. Often two parallel communications are requested, one being suitable for further transfer to the author and indicating where improvements might be made to the text, the other being predominantly a judgement about the article's suitability for publication. Here again referees have the choice of composing these comments by any method and, if they wish, the keyboarding and transmission to the editor can be done by secretaries. But when referees choose to generate the electronic report themselves, there arise further issues about how the use of

word processing facilities may change the character and the quality of what is written. At present we know very little about the psychology of report writing.

The above discussion has sketched out the referee's task using an armchair-based task analysis. When it comes to empirical evidence concerning how referees go about their task, again there seems to be relatively little hard data to draw upon. Several studies of refereeing have cast doubts on the objectivity and reliability of the process. For example, not only is disagreement common among referees but changing the author's name or academic affiliation appears to influence the judged quality of the article (Peters and Ceci, 1982), so too does the pattern of data being reported by the author (Gordon, 1980). If the process is as labile as this it may well be influenced by a change in presentation medium. At present one can only guess whether these known deficiencies in the refereeing system will disappear or will be exacerbated by a change from the display medium of paper to that of the CRT screen. A better understanding of the constituents of the refereeing activity might provide some basis for assessing how these will be affected by changing the medium of communication. The findings from one such small-scale study are summarized in the next section.

3. Investigation of how people referee journal articles

3.1 Design and procedure

It would obviously be inappropriate to ask members of the general public to carry out a professional task such as refereeing an academic paper. Therefore a group of 8 senior academics, 2 women and 6 men, have been kind-hearted enough to act as unpaid volunteers. All are tenured members of the scientific staff of the Medical Research Council, working at the Applied Psychology Unit in Cambridge, England. All have considerable experience in refereeing papers submitted for publication in academic journals.

The expertise of these volunteers lay in different areas of cognitive psychology, so it was necessary to find journal articles of general interest on which these referees could be expected to pass an informed judgement. Two articles were chosen from those available on the BLEND system. One was in the *Computer Human Factors Journal* and dealt with a comparison of the feedback obtained from software editing tools and that obtained from human editors (Hartley and Frase, 1983 and 1984). The other was in the *Bulletin* and dealt with the prospects for the development of electronic journals (Pullinger, 1983). Referees were asked to evaluate the suitability of these articles for "an interdisciplinary journal in the general area of design and human factors". All the referees were familiar with one such journal, namely *Applied Ergonomics*.

Another reason for choosing these articles was their length. Each was roughly 2800 words and took about 15 minutes to read. This meant that both

articles could be dealt with by referees in approximately one hour. Only one referee took so long on the first article that the second was done on a subsequent occasion. For convenience the two articles will be referred to as A and B throughout the following discussion. Because of differences in the use of space by each author, although roughly matched in numbers of words, article A required 23 screen pages whereas article B required only 16 screen pages.

The experimental design required each volunteer to referee both articles. One article was presented as white text on a black background on a 12-inch Hitachi monitor driven by an Apple II computer. The other article was the printout from the italic font of an Epson FX-80 dot-matrix printer. For each article the layout of the screen and paper versions was identical, except for the font changes mentioned. Both used single line spacing. The text itself was confined to a maximum line length of 65 characters and each page had only 24 lines. The printout was centered on the page with wide headings and footings and corresponded exactly to the visual display on the screen. The texts were altered as little as possible from their "published" versions on BLEND, but additional line spacing was introduced where appropriate to have page breaks correspond to thematic breaks within the article. Every line of text had a line number on the left, both in the screen and paper versions. On the right of the screen display there was a reserved margin of 4 characters which could be used by referees to record annotations as they read the article (see below for details).

Four referees began with the article on paper and four had the screen article first. Of each group of four, half started with article A and half with article B. All referees were given written instructions which explained the purpose of the experiment and the criteria to be applied to the papers: "Your task is to make comments on each paper that can assist an editor in deciding whether the paper is suitable for publication. At the editor's discretion your comments will be forwarded to the author(s). The points which you should bear in mind when assessing the paper are listed below; but any negative comments are much more helpful when they are specific and related to particular parts of the text rather than being global expressions of dissatisfaction."

In order to encourage referees to consider several facets of the text, and perhaps increase the homogeneity of the standards being applied, the following written checklist remained available for consultation by referees throughout the experiment:

POINTS TO ASSESS WHEN REFEREEING

1. Does the paper advance our knowledge of issues relating to design (either specifically or in general)?
2. Are the procedures and techniques which are reported in the paper appropriate to the issues being discussed?

3. Are the issues adequately related to the experimental research and design literature?
4. Is the paper intelligible to a non-specialist audience?
5. Please rate the paper on a scale from 1 to 5 (1 = good; 5 = bad) for each of the following characteristics:
 - (a) content, (b) presentation, (c) appropriateness.
6. Should the paper be accepted for publication:
 - (a) as it stands, (b) after minor revision, (c) after major revision
 - (d) after further work, (e) not at all

The instructions to referees also discussed ways of commenting. These instructions said, "You are welcome to annotate the text as you read, you also have a notepad for any notes that you may wish to make, but your comments to the editor should be written on the headed paper provided." The paper was headed "The Journal of Communication Studies".

When the article appeared on the screen, referees were told that they could use the three symbols ?, !, and S to make "comments" in the right-hand margin. These keys were made distinctive on the keyboard by fitting green caps to the keys, and they were grouped together by re-assigning the keys Q,S,X. These symbols would appear at the cursor location in the margin, and this location could be controlled by two other green keys labelled U (for up) and D (for down) which were assigned to the qwerty keys U and N. Two green keys labelled B and F were assigned to the qwerty keys F and L, and enabled readers to move backwards to the previous page and forward to the next page.

After the referees had evaluated both articles, they completed a short four item questionnaire about the two display methods. The first of these questions was a series of multiple choice items with the answers all being selected from the following five-point scale:

- 1 = text on screen much better
- 2 = text on screen slightly better
- 3 = no difference
- 4 = text on screen slightly worse
- 5 = text on screen much worse

For half the volunteers the word "screen" was replaced by the word "paper". The second question asked about the need for extra facilities when dealing with electronic text. The third question asked about the acceptability of reduction in page size to allow space for writing comments on the screen. The final question solicited any other comments.

3.2 Results.

There are three potentially independent aspects of the results and each of these will be considered in turn:

Table I. Summary of the times in minutes taken by referees to deal with articles on paper or on a CRT screen.

	<i>Total time</i>	<i>First read</i>	<i>Write report</i>
Paper			
Article A	36.62	14.01	18.97
Article B	22.40	14.23	8.17
<i>Mean</i>	29.51	14.12	13.57
Screen			
Article A	41.72	18.06	21.48
Article B	29.45	19.98	9.48
<i>Mean</i>	35.59	19.02	15.48
% Screen Slower Than Paper			
Article A	13.9	28.9	13.2
Article B	31.5	40.4	16.0
<i>Mean</i>	20.6	34.7	14.1
Screen	35.59	19.02	15.48

(a) The way referees conducted the refereeing task when the article was on screen and on paper. Relevant performance measures include how long it took, how referees divided their time between reading and writing, how much they moved back and forth through the text.

(b) The assessments made by referees of the two presentation media. This includes their preferences, criticisms, and suggestions for improvements which would facilitate refereeing in an electronic medium.

(c) The impact of medium on the evaluative assessment of the article being refereed. This focusses on the content of the final report written for the editor.

3.2a Referees' performance. Table I summarises the times taken to complete the refereeing task when the article was on paper and when it was on the screen. Six of the eight referees took longer refereeing the screen article. As a group the referees were taking 20.6% longer to deal with the screen article, but this difference was not statistically significant (Wilcoxon $T = 6$).

From the total time, the period spent on the first reading of the article and that spent writing the report for the editor are shown separately in Table I; subtraction of these two periods from the total time will indicate the time spent in a second reading of the article and in activities such as turning to other parts of the text in order to make comparisons. Table I shows that it was the time spent on the first reading which accounted for most of the difference between

screen and paper. Referees spent on average an extra 5 minutes reading when the article was on the screen whereas they spent only an extra 2 minutes in writing the report to the editor. The time spent on the first read through of the article was in the order of 34.7% more when the articles were on the screen than when they were on paper, which is comparable to the decrement of 37.3% reported by Wright and Lickorish (1983) in a proof-reading task when detected errors were recorded on the screen. However, in the present study this difference was shown by only six of the eight referees and was not statistically significant (Wilcoxon $T = 5$, for significance at $p < 0.05$, $T < 5$).

The extra 2 minutes spent writing can be accounted for by the slightly longer comments which referees made when the article was on the screen (225 words) compared with when it was on paper (196 words). Detailed analysis of the content of the referees' reports will be given in section 3.2c, which deals with the way the change in media influenced the evaluative assessment of the texts.

Four of the referees annotated the text as they read. One did so only when the article was on paper, one only when it was on the screen, and two did so in both conditions. All four referees who did not make marginal annotations made notes on the pad as they read, so did one of those who made marginal annotations. For those who used the pad, the mean number of notes made was 9 for the screen article and 6 for the paper article.

Four of the referees carried out a second read of the article. Two did so for both the screen and paper articles, two did so only when the article was on paper. In all but one instance the second readings were very much faster than the first reading, usually less than 20% of the first reading time. Clearly these numbers are too small for safe generalisations, but the emerging picture seems to be one suggesting that there is no common set of procedures adopted by experienced referees for carrying out the task of refereeing a journal article.

Variability is also evident in the reading patterns adopted. For example, some referees made quick flips back to a previous page. The mean number of such flips per referee was 5.3 for the screen article and 3.3 for the article on paper. This mean difference was not statistically reliable, and only four of the eight referees ever made more than two such flips; all four did this when the article was on the screen and three of these made more than two flips when it was on paper. So if a "quick flip" facility had been provided it would have been used by only half this group of referees. Extended searches for earlier information were even less common than quick flips to a previous page. Although only two referees never made an extended search during the course of refereeing, there were only two people who made more than two such searches; one did this for both the screen and paper articles, the other referee did it only for the article on paper.

In interpreting these data as characterising what referees do when they referee, it must be borne in mind that the nature of the task inevitably had a

Table II. Referees' responses to questionnaire rating scale. (Cell values show how many referees gave this rating, max = 8)

	<i>Screen Worse</i>		<i>No Diff</i>	<i>Paper Worse</i>	
	1	2		3	4
Between sections	6	1			1
Legibility	4	3	1		
Between pages	4	1			3
Annotate*	2		3		1
Make comments		2	6		
Follow theme		2	5		1

*Two referees wrote comments which could not be interpreted on this scale.

certain artificiality. The referees were conscious that they were being monitored, they were working under time pressure and evaluating papers which did not lie in their particular area of expertise. Nevertheless, it had been anticipated that these experienced referees would have established some stable strategies for doing a task which they do frequently as part of their professional life. In so far as this assumption is correct, the data suggest people develop different refereeing strategies. Referees differ both in the way they read an article and in the way they integrate writing activities with such reading.

3.2b Assessment of media

Referees were asked to contrast the two media for each of six factors related to actions that referees might want to carry out. The five-point rating scale used to answer these questions was mentioned earlier. The questions themselves were presented in the following order:

- (a) How legible was the text?
- (b) How easy was page turning?
- (c) How easy was moving between different sections of text?
- (d) How easily could you follow the author's argument?
- (e) How easy was it to make marginal annotations?
- (f) How easy was it to compose comments to the editor?

In Table II, which shows the distribution of answers, the questions have been re-ordered so that the upper rows of the table show those factors which the referees said were at a disadvantage for the screen article. The referees considered that having the text on the screen made no difference to activities such as making annotations, or commenting to the editor, or following the author's argument. However seven of the eight referees felt that legibility was

reduced and that moving about within the article, particularly moving about between different sections of the text, was much worse when the article was presented on the screen. These aspects of the electronic display were mentioned by several referees in their written comments.

In response to the question about additional facilities which would have been welcomed when refereeing the screen text, 20 suggestions were made. This is an indication that the system used in the present experiment was far from perfect. But it should also be noted that people have different expectations concerning texts presented under computer control. Several of the facilities requested are not feasible with conventionally printed texts. The comments made by referees are summarized below, where they have been categorised in relation to four areas of interface design: legibility of the display, ease of access to specific information, movement within the text, writing facilities.

Legibility Several referees commented on legibility problems, three mentioning text characteristics (such as the need for descenders, although these were present in both the screen and printout displays used here). One request was for more space between the lines and another was a general comment about eye strain when reading material on VDUs. There is an additional reason for thinking that this screen display may have been less legible than the printed text. Wright and Lickorish (1983) reported similar differences in a proof-reading task using the same screen display but their comparison was with a different dot matrix printer, and they later suggested that much of the difference they had found may have been attributable to annotation factors (Wright and Lickorish, 1984). Nevertheless, the legibility implications have no generality outside the specific typographic characteristics of the displays used in this study. As the technology improves there is no inherent reason why legibility differences should persist.

Access There were five requests for improved access to specific parts of the text, and four of these specified that the access should be via a single key. Such single key access was requested to the contents, the abstract, the reference list, and the specific points that the editor wanted referees to consider. The remaining request for better access concerned the simultaneous display of different sections of the text (i.e., multi-windowing). Such simultaneous comparisons are much more easily achieved with paper than within the confines of the small screen sizes which are most commonly used at the moment.

Movement Eight comments from referees related to moving back and forth within the text. Three referees simply noted that movement would be irksome in a paper longer than that used in this study. One person requested a scrolling facility, another suggested that it should be possible to use the cursor to make

selections from the contents page and jump straight to the required part of the text. Two people wanted the ability to do keyword searches, and one asked for the marginal annotation keys to include a "new page" option with a repeat function which enabled changing multiple pages. The tradeoffs here are not easy to compute. Nor is it simply a question of the economic costs of the alternative systems. The more facilities that are provided, the harder the system is to learn. On the other hand, any one referee may choose only to bother with a manageable subset of the available facilities. Perhaps an ideal refereeing package would need to be customizable, to satisfy the preferred strategies of individual referees.

Writing Only two comments were made by these referees about facilities needed to support their writing activities in a more satisfactory way. One request was for the opportunity to mark within the text itself, and the other was for an auto-repeat function on the marginal annotations. It is possible that if the referees had had greater experience with handling text on a VDU screen (some but not all were regular users of word processing packages running on microcomputers) then full screen editing facilities might have been requested. For editors receiving such modified text from referees there are some obvious difficulties relating to a clear separation between the original text and the referees' suggested changes. With increasing use of colour (or even of reverse video) maintaining such separation need not be an insurmountable problem.

3.2c Evaluation of articles

Given that the CRT display used in this study was felt by those using it to be inadequate in a number of respects, it is of interest to consider whether this will have coloured referees' assessments of the merits of the articles. There is evidence in other contexts, such as examinations, that the characteristics of the visual presentation (e.g., handwriting — Briggs, 1980) may influence the examiner's evaluation of the content.

All referees were asked to assess on a five point scale three separate aspects of each article: its content, presentation and appropriateness for the defined non-specialist target audience. On each of these measures there was a non-significant tendency for the article on paper to be given on average poorer ratings (see Table III). This bias occurred for article B on all three evaluations and for article A on the evaluations of presentation and appropriateness. Such bias may have arisen because the referees were trying to compensate for any carry-over effects which they felt might have arisen due to their irritation with the electronic medium. Whether such compensation would be similarly made outside an experimental setting it is not possible to say.

Referees were asked to make an overall recommendation about the suitability of each article for publication. The distribution of recommendations for each article are shown in Table IV where it can be seen that there was

Table III. Referees' evaluations of three aspects of each article. (Mean rating on 5 point scale, 1 = good 5 = bad)

	<i>Content</i>	<i>Presentation</i>	<i>Appropriateness</i>
Paper			
Article A	3.50	4.00	3.25
Article B	2.75	3.00	2.50
<i>Mean</i>	3.13	3.50	2.88
Screen			
Article A	3.00	2.75	1.75
Article B	2.75	3.50	2.50
<i>Mean</i>	2.88	3.13	2.13
Screen	2.9	3.1	2.1

Table IV. Distribution of each of the eight referee's final recommendations on publication for articles A and B.

<i>Recommendation</i>	<i>Screen</i>	<i>Paper</i>	<i>Overall</i>
1. YES: as it stands	A	A	AA
2. YES: minor revisions	BB	A	A BB
3. ?: major revisions		AA	B AA B
4. NO: needs further work	A B	BB	A BBB
5. NO: not at all	AA B	B	AA BB

considerable variation in opinion. Such differences among reviewers are not uncommon (Peters and Ceci, 1982).

The summary column shown on the right of Table IV indicates that recommendations about article A were fairly evenly split. Three referees thought that it should be published either as it stood or after minor revisions, and three people thought that it should not be published at all or at least not until further work had been carried out. The diversity in respect of article B is only slightly less, and this is the result of agreement that the article should not be published as it stood. Given such diversity it is clearly optimistic to imagine that the effects of presentation medium might be detectable. Yet it is possible to ask whether those who were tough-minded were equally tough on screen and paper, and similarly whether those who were tenderhearted maintained this outlook across media. Figure 1 shows the recommendations on publication for each of the eight referees, as a function of presentation medium, text

content and serial order. The picture that emerges seems to be one of unsystematic variation rather than one of any general trends.

The reports for the editor which were written by each referee were analysed in several ways. The total number of words in each report was counted. It has already been mentioned that the presentation medium had only a small influence on length (paper = 196 words; screen = 225 words). However the length of each referee's report was related to their personal evaluation of the article. The evaluation measure used was obtained by pooling the ratings across the assessments for content, presentation and appropriateness. Each of the two articles seen by a given referee was allocated to either the category "better" or the category "worse" for that referee. This categorisation depended only on the two evaluations made by that referee. For article assigned to the "better" category (mean evaluation rating 2.37) referees wrote on average 256 words, whereas for the article assigned to the "worse" category (mean evaluation rating 3.60) referees wrote on average 166 words (Wilcoxon $T = 1$, $p < 0.02$). It is not possible to say from the present study how general a characteristic of referees' performance this is. There may be a tendency to make fewer comments at both the good and the bad ends of the scale. The

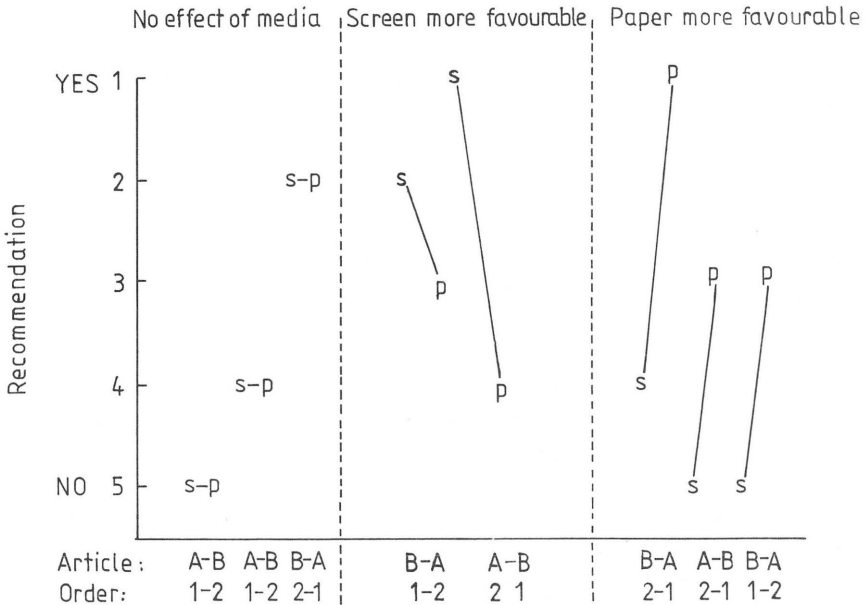


Figure 1. The evaluation made by each referee of the suitability of the two articles for publication, as a function of whether the article was presented on a CRT screen (S) or on paper (P).

Table V. Analysis of content of referees' reports to editor. (Mean number of comments per referee)

<i>Comment</i>	<i>Article A</i>	<i>Article B</i>	<i>Overall Mean</i>
Facts from text:			
paper	2.3	5.5	3.9
screen	7.0	1.0	4.0
Evaluative comments:			
paper	6.3	12.5	9.4
screen	9.3	7.8	8.6
Suggestions:			
paper	3.5	3.5	3.5
screen	4.5	3.8	4.2
Total comments per referee:			
paper	12.1	21.5	16.8
screen	20.8	12.6	16.7

articles used in the present study, in the opinion of these referees, appear to have sampled the bad end rather than the good end of the distribution. In support of this interpretation of the relation between the amount written and evaluation of the article, one referee wrote down on the questionnaire that if the article had been better written they would have made fuller comments to the editor, but that as it was they felt it was not worth doing anything more to the article.

The content of each referees' report was analysed for the frequency with which information from the article was cited, evaluative comments were made, and suggestions were put forward either about presentation or about further research. Table V summarizes the results of this analysis, but again it must be remembered that there were wide ranging differences among the referees. Some people wrote short essays to the editor, others restricted themselves to addressing those points to which the editor had asked them to give attention. An indication of the extent of individual differences is given in Table V, bearing in mind that the referees who read article A on the screen were those who read article B on paper. For example, in the rows headed "Facts from the text" the referees reading article A on the screen made more than three times as many comments on each article as did the other subgroup of referees.

Table V shows that the most frequent category of comments made in the reports to the editor were evaluative (51%). These comments were wide ranging and conveyed both positive and negative feelings. Some referees used phrases such as "well written", "enjoyable", "surprising" of the same article which

other referees found “disjointed”, “predictable”, and “not worth doing”. Because of the diversity of comments and the small sample size there were no observable differences due to the change from paper to screen. There is no evidence in these data for the kinds of negative halo effects which have been reported elsewhere (Briggs, 1980).

4. Implications for support of on-line refereeing

Although generalisations from this study must be made cautiously, in view of the small sample size both of referees and journal articles, there is enough evidence here to indicate some of the reasons why participants in the BLEND experiment were reluctant to referee journals on CRT screens. At a purely practical level it would almost certainly have taken them longer to referee an electronic article. In order to be able to redress this disadvantage in the design of future systems it is necessary to have some understanding of the reasons for the slower performance. These reasons concern cognitive costs imposed by the computer-based communication system. For convenience these costs can be considered to fall mainly into three broad problem areas: perceptual problems, memory problems, writing problems. These three areas will be considered in more detail below but technical developments, in both hardware and software, will enable some of these problems to be reduced or eliminated, if system designers are aware of the requirements of users such as referees.

Perception Perceptual problems relate not only to the legibility of the characters on the screen itself but also to the referees’ need to be able to see different parts of the text, either simultaneously or in rapid succession. One technical solution to this problem is a larger screen with multi-windowing facilities. But even this does not adequately meet the referee’s need to know where certain information is located and how to access it. When the reader formulates an objective (such as making Figure 3 visible in a window alongside a certain passage of text) there are both conceptual and physical aspects of the operation to be performed which can disrupt the efficiency with which the article is read. The implication is that providing multi-windowing facilities will not necessarily solve these problems for referees. Indeed disrupting the linearity of the text might even make it harder for readers to keep track of where they are and where they are going. They may even fail to notice where they have not been. What is needed in addition to a multi-windowing display is a procedure for accessing different sections of an article which is as cognitively undemanding as turning the pages of an MS in typescript.

Memory From the referees’ responses to the questionnaire it appeared that people found it as easy to follow the thematic construction of the author’s argument when the article was on paper. Yet several referees mentioned, either in the questionnaire section for “general comments” or in conversation

afterwards, that when the text was on the screen they appeared to lose some of the incidental location cues which they normally pick up when reading a wad of paper sheets. That is to say, people felt less certain whereabouts certain information was after they had read it. Several studies have shown that the readers of articles printed on paper remember many irrelevant aspects relating to the location of information on the page and to its serial order within the articles (e.g., Rothkopf, 1971). Probably an ideal solution would be to make the visible language of the electronic journal sufficiently distinctive so that readers had available "background" cues, which they might incidentally pick up while reading and subsequently use for retrieval - just as they seem to use the thickness of a pile of typewritten pages. However, it is not immediately obvious what visual characteristics to select for this purpose. Wright and Lickorish are currently exploring the ways in which colour might serve such an incidental cueing function in some kinds of text.

Writing Although comments were made about the annotation facilities, referees did not mention writing the report for the editor as being a problem in this study. Among the reasons for this may be that the report was written by pen on paper for both the screen and paper articles. This is the traditional method of communication with which these referees were familiar. Moreover, half the referees did not annotate the text but made notes on a pad while reading. As a consequence these non-annotators could access their notes as easily when writing the report about the screen article as they did with the paper article. We can only speculate what requirements might have been mentioned if these referees had been sitting at their own terminals reading the article on a VDU screen and composing their report for the editor with full screen editing facilities available. It is likely that this would not only have increased the demand for split-screen displays (for viewing whichever part(s) of the article the referee wished), but there would also have been a demand for multi-tasking facilities so that referees could annotate what they read and also write separately to the editor, these two writing tasks being capable of being performed as overlapping, essentially concurrent, rather than sequential activities. The present data suggest that any such multi-tasking facilities would need to be very flexible to meet the varied working patterns adopted by referees. Enabling the tasks of reading and writing to be engaged in concurrently is an aspect of text display which does not have to be considered when articles are presented on paper.

5. Future prospects for referees

The techniques of communication are changing rapidly as the new information technologies become more widely available. It would therefore be rash to try predicting how specific communities, such as academics, might be affected. There is a current impetus to develop the "fifth generation" of computer systems, which will provide knowledge based and therefore more intelligent tools for many academic functions (Feigenbaum and McCorduck, 1983). One fairly trivial area of application for such an "expert system" is in some of the routine aspects of refereeing/editing, such as checking the correspondence between references cited in the text and those listed at the end of an article. With regard to developments outside the domain of intelligent assistance, three further points can be made with some certainty. One concerns the technology itself; the second concerns the overlap and intermingling between visible and auditory language; the third concerns some of the invisible correlates of visible language.

Technology The investigation reported here was undertaken because of the authors' involvement with the British Library project on an electronic journal (Shackel, 1982). Yet for a variety of reasons it seems unlikely that electronic communications among the academic community will have developed sufficiently for this kind of interchange to be feasible for the next couple of years. One reason is simply that the present economic climate is not encouraging the use by many university departments of networks and workstations that have the necessary power for handling multiple extensive texts. This is necessary if one wishes to access journal length articles and seems likely to be a state of affairs which may change as networks become more widespread.

Among our colleagues here in Cambridge are several who frequently referee papers while travelling by train to meetings. At the moment the highly portable "lap computers" do not have the memory capacity that would enable single texts of 20 to 30 pages to be handled easily. However the predictions are that by the end of 1984 there will be available 32 bit machines with ample processing capacity for this sort of task (Winer and Winer, 1984).

Even if the money were available, the technology is not yet ready for the demands to be made upon it. The limitations relating to articles that require high quality graphics have already been mentioned. A temporary solution is the separate and parallel submission to the editor of figures and text. But even the successful transfer of text will depend on the computer industry's readiness to overcome the current limitations of incompatible systems which are reluctant to communicate with each other. International standards for electronic data transmissions are being worked out, but the standards once agreed have to be implemented by product developers. It all takes time.

Visual auditory Not only will the new technologies create demands for changes in the characteristics of visible language in many domains, they will also blur some of the distinctions between the uses of visual and audio communication. At present the circumstances in which each modality is used are reasonably well defined. Auditory communication is used predominantly for real time person-to-person interaction, visual communication is used when the interaction is asynchronous. In future, voice may be an option for both input and output. Perhaps the referees of an electronic journal article might have their spoken comments captured and transmitted electronically to the editor, who might then have the option of considering these in either spoken or written form. There is already evidence that dictated letters are faster, and therefore more economic, than handwritten-to-typist letters, without there being any evident drop in the quality of the letters (Gould, 1982). Those concerned with the design of visible language have yet to work out the implications for generating messages whose modality may be changed by the recipient. For example, what happens to headings?

Invisible factors One of the points which was emphasized when discussing the results of the present study concerned the difficulties people reported when wanting to move about within the text. This is one of the invisible correlates of visible language. Its importance to successful communication is easily overlooked, perhaps because it has virtually no counterpart in auditory language, although people do hunt for specific portions on a tape recorder and on a dictaphone. When text is printed on paper both readers and writers make use of the opportunity to move back and forth between different sections. The crucial role of this aspect of visible language becomes more obvious from comparisons of people's interaction with text presented in different media. It is not confined to ease of movement, it also includes ease of integrating the reading task with other activities. Sometimes these other activities will be decision making (e.g., the manager taking decisions on the basis of spread sheet information). In the present instance the other activities were annotating and report writing. From the findings of this study it seems likely that irrespective of the ease of reading itself, the acceptability of a new medium for presenting visible language will depend on how smoothly some of these other activities can be integrated with the use of the displayed information.

References

- Blunden, B. (1981). Computer aids in publishing. In E. Than (ed.) *Computer Graphics Manual* Key papers from computervision '81. London: Nord Media Ltd.
- Bradley, J.V. (1982), Editorial overkill. *Bulletin of the Psychonomic Society*, 19, 271-274.
- Briggs, D. (1980). A study of the influence of handwriting upon grades using examination scripts. *Educational Review*, 32, 185-193.
- Card, S.K., Moran, T.P., and Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, N.J.: Lawrence Erlbaum Assoc.
- Davies Cooper, R.F., and Cooper, C.L. (1984). Effect of new technologies on the work and methods of the typographic designer. *Design Studies*, 5, 21-29.
- Feigenbaum, E.A., and McCorduck, P. (1983). *The Fifth Generation: artificial intelligence and Japan's challenge to the computer world*. London: Pan Books.
- Gordon, M.D. (1980). The role of referees in scientific communication. In J. Hartley (ed.) *The Psychology of Written Communication*. London: Kogan Page.
- Gould, J.D. (1982). Writing and speaking letters and messages. *International Journal of Man-Machine Studies*, 16, 147-141.
- Hartley, J., and Frase, L.T. (1983). Human and Computer aids to writing. Archived in electronic form as part of the British Library Electronic Network Development experiment, in *Computer Human Factors 2*, entries 1-51.
- Hartley, J., and Frase, L.T. (1984). Human and computer aids to writing. *IEEE Transactions on Professional Communication*, 27, (in press).
- Inglis, D.A. (1983). New developments in typesetting and their implications for graphic and instructional designers. *Information Design Journal*, 3, 183-197.
- Meadows, J. (1983). Scholarly communication in transition. *Journal of information Science*, 7, 81-97.
- Peters, D.P., and Ceci, S.J. (1982). Peer review practices of psychological journals: The fate of published articles submitted again. *The Behavioural and Brain Sciences*, 5, 187-225.
- Pullinger, D. (1983). Reading electronic journals on-line. Archived in electronic form as part of the British Library Electronic Network Development experiment, in *Bulletin 7*, Entries 1-35.
- Rothkopf, E.Z. (1971). Incidental memory for location of information in text. *Journal of Verbal Learning and Verbal Behavior*, 10, 608-613.
- Shackel, B. (1982). The BLEND system: program for the study of some "electronic journals". *Ergonomics*, 25, 269-284.
- Walker, C. (1983). Between the lines. *Design*, 409, 48-49.
- Winer, D., and Winer, P. (1984). Portables - 1984 and beyond. *Byte*, 9, 243-262.
- Wright, P., and Lickorish, A. (1983). Proof-reading texts on screen and paper. *Behaviour and Information Technology*, 2, 227-235.
- Wright, P., and Lickorish, A. (1984). Ease of annotation in proof-reading tasks. *Behaviour and Information Technology* (in press).