4. LEGIBILITY and the RATING of SECOND LANGUAGE WRITING:

AN INVESTIGATION of the RATING of HANDWRITTEN and WORD-PROCESSED IELTS TASK TWO ESSAYS

Annie Brown
The University of Melbourne

IELTS™
English for International Opportunity

Publishing details

International English Language Testing System (IELTS)

Research Reports 2003

Volume 4

Editor: Robyn Tulloh

IELTS Australia Pty Limited
ABN 84 008 664 766
Incorporation in the Australian Capital Territory
Web: www.ielts.org

National Library of Australia
Cataloguing-in-Publication Data
2003 ed
IELTS Research Reports 2003 Volume 4
ISBN 0 86403 045 2

Introduction

Handwriting and neatness of presentation has long been seen as a contaminating factor in the assessment of writing ability. In particular it has been invoked as a possible reason why girls tend to perform better in relation to boys on free-response writing tests than they do in forced-choice formats in tests of first and second language proficiency.

Over the years there have been a number of studies in the area of first language writing assessment which have investigated the impact of handwriting on overall judgements of writing quality. Some of these have involved correlations of teacher-assigned ratings of writing quality with independent judgements of handwriting (eg. Stewart and Grobe, 1979; Chou, Kirkland and Smith, 1982). Others have involved experimental designs where the same essays are presented to raters in different presentation formats involving good handwriting, poor handwriting and, in some case, typed scripts (Chase, 1968; Marshall and Powers, 1969; Briggs, 1970, Sloan and McGinnis, 1978. Bull and Stevens, 1979, McGuire 1995). The findings indicate in general that the quality of handwriting does have an impact on the scores awarded to essays, and that increased legibility results in higher ratings; in all the studies except that by McGuire (1995), the essays with better handwriting or the typed scripts received higher scores.

Given the great interest over the years in handwriting and its impact on assessments of writing proficiency within the field *of first* language literacy, it is surprising that there are hardly any studies of the effect of handwriting in the assessment of *second* language writing. The only study that could be traced involving essays written by non-native speakers, Robinson (1985), produced similar findings to the majority of the first language writing studies; essays written by students whose L1 did not use the Roman alphabet tended to receive lower scores than essays written by 'expert' writers.

The lack of research into the impact of handwriting on assessments of L2 writing proficiency is all the more surprising in a field where reliability and validity issues are generally well understood, and where much attention is paid in the research literature to identifying and examining the impact of construct-irrelevant variance on test scores. One could argue that it is particularly important in formal L2 writing test contexts to examine and evaluate the impact of extraneous variables such as handwriting and presentation, because it is often on the basis of such tests that decisions regarding candidates' future life or study opportunities are made. Moreover, it is particularly in writing contexts such as these, where the writers typically have to write under considerable time pressure, that it may be most difficult for them to control the quality of handwriting and general neatness of layout. It is rare, for example, in formal tests that writers have time to transcribe a draft of the essay into a more legible and well-presented script. Also, as Charney (1984) points out, in a test context the constraints imposed on the *rater* may result in handwriting playing a larger part in the assessment than it should. Charney argues that the assessment constraints, that is limited time and multiple assessment foci, mean that raters have to read essays rapidly and this may force them to "depend on those characteristics [such as handwriting] in the essays which are easy to pick out but which are irrelevant to 'true writing ability'".

It is, perhaps, natural to assume that the same situation would hold for assessments of L2 writing as for L1 writing, that is, that poor handwriting would have a negative impact upon scores. Such an expectation seems logical - as Chou *et al* (1982) point out, a paper that looks good and is easy to read is likely to create a better impression on a rater than one which is messy or difficult to read. Chou *et al* also point out that it is not only handwriting *per se* that

may influence ratings, but that crossings out and re-sequencing of pieces of text may also make a bad impression as they can be interpreted as being indicative of a student who is unprepared for writing and unsure of how to sequence his or her ideas. They seem to be implying, then, that it may not simply be that poor writing is difficult to process (and therefore assess) but also that raters may make negative inferences about the character or personality of the writer on the basis of their script untidiness.[1]

There is no obvious reason to suppose that the same features of writing would have a different impact in a L2 writing context. It may even be that students with poor handwriting are even more disadvantaged in L2 contexts because of the centrality of 'fluency' as an aspect of communicative quality. It is difficult to read fluently when illegible handwriting and poor presentation hinder access to the text. Huot (1993) argues that under test marking conditions where rapid reading is required, poor handwriting is likely to impede fluent reading. On the basis of verbal protocol studies of the rating of L2 writing, in fact, it appears that raters do react in much the same way as they do when rating L1 writing; in both contexts it has been found that raters comment frequently and negatively on legibility (see, for example, Huot, 1988, 1993; Cumming, 1990, 1998; Vaughan, 1991; Weigle, 1994; Wolfe and Feltovich, 1994; Milanovic, Saville, & Shen, 1996).

Whilst one could argue that in these days of increased computerisation of tests and test centres, the issue of handwriting in second language tests is likely to become redundant. However, the fact that students may be offered one mode of testing or the other depending on the facilities available at a given test centre, or may even be offered a choice of test mode, the issue of equity arises. The current study arises in the context of a move to administer IELTS in both computerised and pen-and-paper formats, and examines whether the supposed effects do, in fact, occur in second language contexts, that typed essays or neatly presented essays receive higher ratings, or whether, given the different rating focus (less on ideas and content and more on the mechanics of writing) or the different test purpose, a different dynamic holds.

Protocol analysis is perhaps the most common approach taken to an examination of factors influencing the assessment of L2 writing. In fact, most of the claims-made about the impact of handwriting in second language writing are based upon verbal protocol studies produced by raters rather than on actual analyses of score data. These protocols go under many names (think-aloud, talk-aloud, concurrent) and basically involve asking raters to rate a series of writing texts, and, while doing so, to describe the thought processes they go through. Since its first use in applied linguistics, to examine the process of writing, protocol analysis has come to be a popular and widely-used methodology. There is a common belief that verbal protocols produce valid data in terms of assessment focus. Huot, for example, claims that "the examination of the verbal reports of the raters in the study provides the best possible insight into the process of reading and rating student writing" (1993: 215). He also maintains that this methodology "bypasses the need for manipulating student essays or correlating scores with textual features" (1990: 207). In fact, so widespread have protocol studies of the rating process become that questions about the veracity of the verbal reports and the validity of the inferences drawn from the analyses tend to be swept aside in the enthusiasm for the rich and detailed information that is generated. In fact many of the studies discussed above classify and quantify the focus of comments in order to draw conclusion about what aspects of performance raters focus on. Vaughan (1991), for example, concludes that handwriting is

[1] A study of the rating of speaking (Brown, 2000) found just this, that raters constantly make inferences about the personality or maturity of the candidate on the basis of what they say or how they interact with the interviewer.

salient to raters on the basis that the number of references to handwriting was second only to content. There appears even to be an assumption that there is a direct relationship between *how much* attention raters pay to particular features (in terms of numbers or frequency of comments) and the *weighting* accorded that feature in the overall assessment.

Whilst a study of what raters claim to be the influences on their ratings (Milanovic, Saville, and Shen 1996) supports the inferences that are drawn from protocol studies regarding the salience of handwriting, in that legibility was regarded by two-thirds of raters as having an effect in varying degrees on their rating, there is, in fact, no *direct evidence* that the number or frequency of comments produced in protocols on any one aspect of the performance directly reflects the contribution of that feature to the score. Indeed, it has been argued that verbalisations (as produced by raters when rating writing scripts) are neither a full nor an accurate representation of cognitive activity, especially where tasks are complex and cognitively demanding (Ericsson and Simon, 1984). It may be that raters are cognitively able to focus on different aspects of performance simultaneously, so that any linear verbal description is necessarily incomplete. It has been claimed that certain processes are automatic (particularly for experienced raters) and therefore not accessible to description, and it has also been argued that raters are likely to comment more on those aspects of performance which are easy to single out and describe, that is those which refer to the more obvious or surface features of the text. Of course, as well as referring to the features which influenced their ratings, raters may refer to aspects of performance which are commentable but peripheral; in describing what they are doing (the rating process), they may include features which they take note of while reading but which are not relevant to the assessment (the rating outcome). Finally, they may be influenced by the fact that their comments are being taped and analysed to report things that they feel the researcher would want to hear, or which they perceive will show them up in a good light; conversely they may leave out reactions and comments which they feel might be interpreted as inappropriate or irrelevant.

The second research question is concerned with the validity of verbal protocols as evidence of raters' assessment focus. Using the same scripts, and with the focus still on handwriting, the appropriateness is examined of the assumption commonly made about verbal protocols, namely that there is a relationship between the number of comments made about a particular aspect of performance and the effect of that aspect of performance on subsequent ratings. On the basis of these assumptions, it was hypothesised that essays with poor legibility (and hence a marked score difference between the typed and handwritten versions) would attract more negative comments on handwriting and presentation than those with good legibility.

Question 1: The Impact of Presentation on Ratings

The salience of legibility as a factor in raters' judgements is examined within a controlled experimental study using IELTS Task Two essay, within a in which a comparison was made of scores awarded to scripts which differed only in relation to the variable 'handwriting'.

On the basis of previous studies in L1 contexts it was hypothesised that scores awarded to the handwritten and typed versions of the essays would be significantly different, with higher scores being awarded to the typed versions. In addition it was hypothesised that the score differences would be greater for those scripts where the handwritten version had particularly poor legibility.

Methodology

Forty IELTS scripts were selected at random from administrations held at one test centre within a one-year period. The scripts were selected from five different administrations and involved five different sets of essay prompts.

Each of the Task Two essays was retyped. Original features such as punctuation, spelling errors and paragraph layout were retained, but aspects of text editing which would be avoided in a word-processed essay, such as crossings out, insertions and re-orderings of pieces of text, were tidied up. See Appendix 4.1 for examples of handwritten and typed versions of essays.

IELTS ratings

Next, in order to produce stable and comparable ratings for the two script types, that is handwritten and typed (henceforth H and T), each essay was rated six times. In order to ensure that ratings awarded to an essay in one script type did not affect scores awarded to the same essay in the other format, raters did not mark both versions of the same essay. Rather, each of twelve accredited IELTS raters involved in the study rated half of the typed scripts and half of the handwritten scripts, each being from a different candidate. Appendix 4.2 sets out the rating design. Because the need to ensure that raters did not see the same essay twice meant that the raters could not all rate the same set of scripts, the classical calculation of inter-rater reliability (correlation) could not be used here. However, the overlapping design shown in Appendix 4.2 made it possible to use Multi-faceted Rasch Analysis (Linacre, 1989; Linacre and Wright, 1992) to both estimate a measure of inter-rater agreement (expressed in terms of 'fit') and to examine relative rater severity.

Although in normal operational studies it is left to the discretion of raters as to whether they rate the essays globally or analytically, for the purposes of this study, in order to investigate whether poor legibility had most impact on one particular assessment category the raters were instructed to assess all the essays analytically. Thus ratings were awarded to each script for each of the three Task Two analytic categories: *Arguments, Ideas and Evidence*, *Communicative Quality*, and *Vocabulary and Sentence Structure*. A final overall bandscore was calculated in the normal way, by an averaging and rounding of the three analytic scores. Raters also took the length of each essay into account in the usual way.

Legibility judgements

In addition to the IELTS ratings, judgements were made of the legibility of each handwritten script. A six point scale was developed specifically for the purposes of this study. Drawing on discussions of legibility in verbal report studies such as those discussed above, legibility was defined as a broad concept which included letter and word formation, general layout (spacing, paragraphing and lineation), and editing and self-correction. The four judges (all teachers of writing in first or second language contexts) were given written instructions to accompany the scale (Appendix 4.4).

Results

Inter-rater and inter-judge agreement

Before discussing the findings, in order to confirm that the score data is reliable and therefore useable, we turn briefly to an analysis of inter-rater (IELTS ratings) and inter-judge (handwriting judgements) agreement.

Judge	Count	Measure (Severity)	Model S.E.	Infit MnSq	Std	Outfit MnSq	Std
1	40	-1.46	0.24	1.0	0	1.0	0
2	40	1.13	0.27	0.7	-1	0.6	-1
3	41	0.87	0.26	1.1	0	1.1	0
4	39	-1.2	0.25	0.8	0	0.8	0
5	40	-1.9	0.24	1.5	1	1.4	1
6	40	-0.29	0.25	0.9	0	0.9	0
7	40	2.7	0.28	1.5	1	1.5	1
8	40	-1.27	0.25	0.7	-1	0.7	-1
9	40	0.4	0.26	0.9	0	0.9	0
10	40	1.25	0.27	0.8	-1	0.8	0
11	40	-0.01	0.26	0.7	-1	0.7	-1
12	40	-0.23	0.26	1.1	0	1.1	0
Mean	40	0	0.26	1	-0.2	1	-0.2
S.D.	0.4	1.29	0.01	0.3	1.1	0.3	1.1

Table 1: IELTS ratings - Inter-rater agreement

Inter-rater agreement was estimated using Many-facet Rasch Measurement (MFRM). MFRM describes agreement or similarity between raters in terms of measures of 'fit' and 'severity' (Table 1). Raters 5 and 7 were found to be the least in agreement with the others, both lying outside what is normally considered acceptable fit, that is 0.7 to 1.3 (Linacre, 1999). However, given the relatively small number of ratings undertaken by each rater (ie. forty), and the fact that the fit values lie only *just* outside the acceptable values, it is not appropriate to place too much emphasis on this misfit with regard to the acceptability of the score data. The same two raters, Raters 5 and 7, also turn out to be the most lenient and the most severe with a difference of almost two bandlevels. Whilst this difference seems somewhat large (certainly for operational testing), given that scores for each script were averaged across six raters and that each rater rated twenty of each type of script, the range of severity levels amongst raters should not affect the comparison of the two script types

Inter-judge agreement for the judgements of legibility (scored on a scale of 1-5) is shown in Table 2. Whilst it would be considered low for rater agreement in operational testing, lying between .62 and .75, given that this is not an operational rating and the judges were neither trained nor familiar with making such judgements, this level of agreement can be considered acceptable.

	Judge 2	Judge 3	Judge 4
Judge 1	.688	.696	.623
Judge 2		.752	.740
Judge 3			.685

Table 2: Legibility ratings – Inter-judge agreement

Score data

Turning now to the score data, Appendix 4.3 shows the ratings (averaged across the six raters) awarded for both the analytic and overall score categories for each version (H and T) of each essay. Table 3 below presents the same data summarised (meaned) for each score category. It shows that both the analytic and overall scores were on average marginally higher for the handwritten scripts than for the typed scripts. The handwritten scripts achieved a mean rating of 5.30 as opposed to 5.04 for typed scripts for Arguments, Ideas and Evidence (AIE), 5.60 as opposed to 5.34 for Communicative Quality (CQ), 5.51 as opposed to 5.18 for Vocabulary and Sentence Structure (VSS), and an overall bandscore (OBS), averaged across the three categories, of 5.48 as opposed to 5.17. The spread of scores was similar for both type of script. Although one might expect the score difference to be least marked for Arguments, Ideas and Evidence, as this category is the least concerned with presentation issues, and most marked for Communicative Quality as handwriting which is difficult to read will inevitably make the essay less immediately 'communicative', the score difference appears to impact relatively evenly across all assessment categories. In contrast to what was expected scores were marginally higher for the handwritten scripts than for the typed scripts in all categories.

Rating Category	Handwritten Mean	Handwritten SD	Typed Mean	Typed SD	Difference (H-T)	Z	Sig.
AIE	5.30	*0.81*	5.04	*0.82*	.26	-2.570	.01
CQ	5.60	*0.85*	5.34	*0.82*	.26	-3.530	.00
VSS	5.51	*0.80*	5.18	*0.79*	.33	-4.230	.00
OBS	5.48	*0.77*	5.17	*0.70*	.31	-3.723	.00

Table 3: Score analysis of handwritten and typed scripts

In order to investigate the significance of the score differences for the two script types for each rating category, Wilcoxon matched-pairs signed-ranks test was carried out (also Table 3). As can be seen, although the difference in mean scores is relatively small (0.26 for AIE and CQ, 0.33 for VSS, and 0.27 for OBS), it is nonetheless significant for all rating categories.

The second analysis looked more narrowly at the impact that different degrees of legibility have on ratings. On the basis of findings within the L1 writing assessment literature, it was considered likely that the score differences across the two script types (H and T) would be insignificant for highly legible scripts but significant for ones which were difficult to decipher. A comparison was made of the score differences for the ten essays judged to have the best legibility and the ten judged to have the worst.

Table 4 shows the average score difference for the two script types for each set of ten essays. As expected, the score difference between the H and T versions for the candidates with the best handwriting was found to be relatively small (ranging from 0.05 to 0.17 of a band), whereas for those with the worst handwriting it was somewhat larger (ranging from 0.5 to 0.62, ie. at least half a band).

Wilcoxon matched-pairs signed-ranks test was carried out in order to determine the significance of the score differences between the two script types for each group. For the scripts with the best handwriting, none of the differences were significant. For those with the worst handwriting AIE was not significant but then other three categories were, CQ at the .05 level, and VSS and OBS at the .01 level.

Rating Category	Least legible (n=10)	Z	Sig	Most legible (n=10)	Z	Sig
AIE	0.50	-1.84	*.07*	0.13	-1.13	*.26*
CQ	0.50	-2.20	*.03*	0.15	-1.19	*.23*
VSS	0.62	-2.67	*.01*	0.17	-.94	*.35*
OBS	0.59	-2.45	*.01*	0.05	-.20	*.83*

Table 4: Score Differences according to handwriting quality

In summary, then, the analysis has shown as hypothesised that there was a small but significant difference in the scores awarded to typed and handwritten versions of the same essay. Also as expected the score difference between handwritten and typed essays was greater for essays with poor legibility than for those with good legibility, being on average less than 0.1 of a band for the well-written essays and slightly over half a band for the poorly written ones. However, contrary to expectations, it was the handwritten scripts which scored higher, and the handwritten scripts with poor legibility which had the greatest score difference between versions. In effect this means that rather than being *disadvantaged* by bad handwriting and poor presentation, test candidates are *advantaged.*

It is interesting to reflect more closely on why this might have happened. As noted earlier, a major difference in the rating of L1 and L2 writing is that in L2 assessments there is a heavy central focus on mechanics, whereas this is far less important in L1 assessments (Cumming 1998). It may be, then, that poor legibility has the effect of masking or otherwise distracting from mechanical errors. Given that in L2 writing assessment, raters usually have a limited time in which to make their judgement, it may be that the extra effort required to decipher illegible script distracts from a greater focus on grammar and accuracy, so that candidates are somehow then given the 'benefit of the doubt'. The corollary of this, of course, is that errors stand out more (are more salient) when the essay is typed or the handwriting is clear.

Question 2: The Validity of Verbal Protocols as an Indicator of Assessment Focus

In order to investigate the second research question (that of the relationship between comments on handwriting and impact on score), four 'protocolers', all accredited IELTS raters, were asked to rate the handwritten versions of the essays whilst providing think-aloud protocols.

Methodology

Twenty-one scripts were selected; seven with the greatest score difference between the two versions (which had the worst handwriting), seven with the smallest score difference between versions (which had the best handwriting), and seven which were ranked as average for handwriting. Each protocoler completed all twenty-one essays in a single individual sitting.

On arrival, each protocoler was given the IELTS Bandscale and rating instructions for Writing Task Two and time to re-familiarise themselves with the materials as they would before the start of an operational rating session. Those who wished to complete some practice rating before the protocol session commenced were given four IELTS scripts based on the one essay prompt.

Next the researcher introduced the participants to the protocol technique. Each protocoler was given a set of instructions to read (Appendix 4.5). For each script they were instructed to switch on the tape, read out the script ID and then commence their reading and rating, at the same time reporting verbally what they were doing and what they were taking notice of. When reading the text or the bandscale, they were instructed to read aloud the relevant bits. They were instructed to write down their assessment once they had completed the rating, and then to switch off the tape. They were told they should have a break whenever they felt the need for one. Each protocoler undertook three practice protocols, after each of which the researcher provided feedback on the quality of the protocol data (sufficiency, clarity, etc.).

Results

All references by the four protocol raters to handwriting and other presentation factors affecting legibility (ie. those aspects of writing which can make it difficult to read the handwritten script but which are not relevant to word-processed scripts) were identified for each of the twenty-one essays. These comments reflected the criteria in the legibility judgements - other than the handwriting itself (neatness, size, spacing, etc.) other aspects of the general layout which affected legibility included crossings out, insertions and re-orderings of paragraphs. Each comment was also coded as positive, negative or neutral.[2]

Examples of these comments are presented below. The code for each excerpt refers to:
The rater (A, B, C, or D), the handwriting category (good, poor or average), the script ID, and the comment polarity (positive, negative or neutral).

[2] The neutral category included 'backhanded' compliments such as "Oh, here's a sentence I think I can read".

A-G-E7-POS	It's easy to read, it's nice handwriting.
A-P-C8-NEG	You can just see how you lose all sense when you can't read the writing. .
B-P-C7-NEG	[*Many*] what word is that [*Many ... people remain in their closed environments Many... people remain in their closed environments because they do not wish to lose their high social norms*] I can't read the second word in the second paragraph to work out what it's meant to be about.
D-P-C8-NEG	It's really very hard to decipher the handwriting, I'm finding it very difficult.
D-G-D5-NEUT	I'm guessing the person's from Indonesia from the handwriting, could be wrong, seems to be that sort of script.
B-P-D3-NEUT	This one writes like I think, goes back and says oh I should've said that or I should've said that, maybe they've said it properly cause they fixed up what they were writing. Maybe they even re-read what they wrote which is unusual.
A~P~E5-NEG	This gets confused because this person has indicated that the script is, the paragraphs get rearranged but I'm not sure how.
B-M-A2-NEG	and A2 is really large scrawly writing, scrawled over two pages. gets progressively untidier probably gonna be a bit difficult to read. Lots of scribbling and rearranging and changing word order and bits and pieces.

As has been found in other studies, the overwhelming majority of comments were negative. In fact positive comments tended to be produced only within certain routinised overviews of the script, an approach used regularly by one rater, Rater C, but to some extent by all of them. Rater C began every assessment with a 'first impressions' overview of the paragraphing, handwriting and length, for example:

> *Oh this one's nice, what beautiful writing. Hm, looks about the right length, I don't think I need to check this one [counting] yeah no this one looks fine. Nice clear paragraphing. Okay so it creates an overall good impression. Let me read the first paragraph...."*

Neutral comments about handwriting made reference to ethnicity (the rater guessing or making assumptions about the candidate) and neutral comments about untidy layout made reference to the fact that the crossing out and insertions were an indication that the writer had gone back to edit his or her work.

Due to the relatively small number of comments in general, it was considered inappropriate to undertake a statistical analysis of significance. Instead a simple count was made of the number of comments involving legibility for each of the three sets of scripts (Table 5).

Handwriting Grading		Scripts			
	Total *nt= neutral*	**A**	**B**	**C**	**D**
Good	8+ 1nt	2+	3+	3+	1nt
Average	3+ 4-	2-	1-	3+ 1-	0
Poor	29- 3nt	11-	9- 2nt	5-	4- 1nt

Table 5: Legibility comments

As handwriting and layout were the basis of the judgements of legibility, as hypothesised one would expect that the number and direction of comments in this category would correlate with the degree of legibility. There does, in fact, appear to be a trend for scripts judged to have poor handwriting and layout to receive more negative comments than those judged to have good handwriting and layout, which means that the number of comments does appear to reflect the impact on score, as those with the worst legibility had the greatest difference in scores for the two versions. The difference between the sets of essays is quite marked, with a total of twenty-nine negative handwriting comments and no positive ones received by the seven essays with the worst handwriting, no negative comments but eight positive comments received by those with the best handwriting, and a balance of three positive and four negative comments received by scripts with average handwriting.

Interestingly, despite a link between the number of comments and the strength of influence on scores, the polarity of the comments made by the raters reflects the opposite of what they did; scores were higher where there were more negative comments, not lower. As was commented earlier, it may be that raters do not mark scripts down for legibility but that poor legibility masks mechanical problems (grammaticality or spelling, for example), so that the essays are instead marked up. One of the raters commented that she deliberately allocates more time to the script if the handwriting is poor in order to ensure that she is fair on the candidate, but it cannot be assumed that all raters do this.

With regard to the validity of the verbal protocols, then, on the basis of the findings here it can be argued that not only do raters make evaluations which are not reflected in the scores (ie. the negative comments on handwriting), but also there appear to be interactions going on between different assessment features (ie handwriting and (possibly) grammaticality) which raters are almost certainly unconscious of and which are not reflected in their protocols.

Conclusion

This study investigated two research questions. Firstly, in the context of a move to deliver IELTS in two alternative modes, pen-and-paper and computer, it explored the impact of legibility on ratings awarded to IELTS Task Two essays. The study found that legibility plays a small but significant role in scores, and that the size of the effect is relative to the quality of handwriting and presentation. However, the direction of the effect in this study was unexpected; whereas it had been hypothesised on the basis of numerous studies of L1 writing assessment that poor legibility would lead to lower scores, the opposite was, in fact, the case. Given that the assessment of L2 writing differs crucially from L1 assessment in that there is a much stronger emphasis on 'linguistic' features (syntax, grammar, vocabulary),

Marshall, J.C. and Powers, J.M. (1969) Writing neatness, composition errors and essay grades. *Journal of Educational Measurement* 6: 97-101.

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Studies in Language Testing 3* (pp. 92-114). Cambridge: CUP and UCLES.

Robinson, T.H. (1985) Evaluating foreign students' compositions: the effects of rater background and of handwriting, spelling and grammar. The University of Texas at Austin: Unpublished PhD thesis.

Sloan, C.A. and McGinnis, 1 (1978) The effect of handwriting on teachers' grading of high school essays. (ERIC Document reproduction Service No. 220 836).

Stewart, M.R. and Grobe, C.H. (1979) Syntactic maturity, mechanics, vocabulary and teacher's quality ratings. *Research in the teaching of English* 13: 207-215.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex Publishing Corporation.

Weigle, S. (1994) Effects of training on raters of ESL compositions. *Language Testing* 10:197-223.

Wolfe, E., & Feltovich, B. (1994). Learning to rate essays: A study of scorer cognition. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA. (ERIC Document Reproduction Service No. ED 368 377).

Appendix 1 Scripts D3 and E8 (Typed version)

D3

As science and technology is developing, scientists make great advances in many subjects such as genetics. Some advances are helpful for human being. However, if some techonologys are used on humanbeing, ethics should be considerated in advance.

As genetics engineering is developping rapidly cloning animals becomes reality. Corp and animals breeding benefits from this technology But some scientists intend to clone humanbeings. That can cause ethical problems. Moreover, cloning humanbeings has other negative effect such as reducing resistance of human and similarity of people who lead to confused. Consequently, Many countries ban cloning humanbeing by law.

On the other hand, fertility treatment and organ transplants from animals to humans are good for sterile people and sick people whose organs are out of order. Thus, that is a blessing for them these projects should be developped.

In conclusion, while scientists pursue scientific goals, the enthics shold be considerated carefully if they experiment on human.

E8

Nowadays all over the world a huge movement of people in rural areas to urban areas occurs. Although people could move to the place where they hope to improve their life this immigration could cause some problems.

It is true that the life in urban areas, to some extent, is better than in rural areas. In the urban areas there are many jobs with reasanable salary available. Also, the transport and living facilities are more developed. Furthermore, the life in the urban areas seems to be more excited. So people in the rural areas hope that they could improve their life considerably if they move to the cities. For exampe, in India, China there is a large immigration to big cities. According to the statistics, every year 30 million Chinese moved from rural areas to cities.

However, this immigration could lead to many problems. First, the cities could become overpopulated. So the number of available jobs could be reduced, and the number of unemployed people rose considerably. As a consequence, crime in cities could increase. Second, if many people from rural areas come to cities they could face the problem of accommodation when houses and building in cities are also limited. Lastly, when cities is overpopulated the problem of pollution and transport also becomes difficult to solve.

In conclusion, I believe that people could move to the place where they could improve their life. However, this could lead to many problems. So goverments should have appropriate immigration policy and economical development plans so that people in rural areas could improve their life in their villages.

Sample Handwritten texts

Poor legibility

As science and technology is developing, scientists make great advances in many subjects such as genetics. Some advances are helpful for human being. However, if some technologys are used on humanbeing, ethics should be considerated in advance.

As genetics engineering is developping rapidly cloning animals becomes reality. Corp and animals breeding benefits from this technology But some scientific intend to clone humanbeings, that can cause ethical problems. Moreover cloning humanbeings has other negative effect such as reducing resistance of human and similarity of people who lead to confused. Consequently, Many countries ban cloning humanbeing by law.

On the other hand, fertility treatment and organ transplants from animals to humans are good for sterile people and sick people whose organs are out of order. Thus, these projects should be developped. That is a blessing for them

High legibility

It is very clear at the moment that all countries had better help each other in many ways. Funding from overseas is the easiest and quickiest way to help developing countries. However, if the government misspend this money, which is always a very large amount of money, and this money is not contributed to the poor of these countries, whether international aid is still a proper solution to help ?

A clear advantage of funding from overseas is that the government can spend it immidiately on a problem. For example, when there was a flooding disaster in the south of Thailand in 1990, Thai government received money from Japan and some European countries to help all people that lost everything from this disaster. The government provided accommodation, food and health care for free and people appreciated this international moral undoubtedly. Moreover, the government contributed to these people by giving money from overseas for their jobs

On the other hand, some countries such as Nicaragua, do not spend this money correctly. United Nation and USA subsidized a huge amount of money to the Nicaragua's government in 1987 to build all infrastructure and to support all studies. But the government spent money more 50 percent providing guns, bombs and tanks in order to

Appendix 2 Rating design

Script No.	Handwritten version							Typed version					
	Raters							Raters					
A1	1	3	5	7	9	11		2	4	6	8	10	12
A2	1	3	5	7	9	11		2	4	6	8	10	12
A3	1	3	5	7	9	11		2	4	6	8	10	12
A4	1	3	5	7	9	11		2	4	6	8	10	12
A5	1	3	5	7	9	11		2	4	6	8	10	12
A6	1	3	6	7	9	12		2	4	5	8	10	11
B1	1	3	6	7	9	12		2	4	5	8	10	11
B2	1	3	6	7	9	12		2	4	5	8	10	11
B3	1	3	6	7	9	12		2	4	5	8	10	11
B4	1	3	6	7	9	12		2	4	5	8	10	11
B5	1	4	6	7	10	12		2	3	5	8	9	11
B6	1	4	6	7	10	12		2	3	5	8	9	11
B7	1	4	6	7	10	12		2	3	5	8	9	11
B8	1	4	6	7	10	12		2	3	5	8	9	11
C1	1	4	6	7	10	12		2	3	5	8	9	11
C2	1	4	5	7	10	11		2	3	6	8	9	12
C3	1	4	5	7	10	11		2	3	6	8	9	12
C4	1	4	5	7	10	11		2	3	6	8	9	12
C5	1	4	5	7	10	11		2	3	6	8	9	12
C6	1	4	5	7	10	11		2	3	6	8	9	12
C7	2	3	6	8	9	11		1	4	5	7	10	12
C8	2	3	6	8	9	11		1	4	5	7	10	12
D1	2	3	6	8	9	11		1	4	5	7	10	12
D2	2	3	6	8	9	11		1	4	5	7	10	12
D3	2	3	6	8	9	11		1	4	5	7	10	12
D5	2	3	5	8	9	12		1	4	6	7	10	11
D6	2	3	5	8	9	12		1	4	6	7	10	11
D7	2	3	5	8	9	12		1	4	6	7	10	11
D8	2	3	5	8	9	12		1	4	6	7	10	11
D9	2	3	5	8	9	12		1	4	6	7	10	11
D10	2	4	5	8	10	12		1	3	6	7	9	11
E1	2	4	5	8	10	12		1	3	6	7	9	11
E2	2	4	5	8	10	12		1	3	6	7	9	11
E4	2	4	5	8	10	12		1	3	6	7	9	11
E5	2	4	5	8	10	12		1	3	6	7	9	11
E6	2	4	6	8	10	11		1	3	5	7	9	12
E7	2	4	6	8	10	11		1	3	5	7	9	12
E8	2	4	6	8	10	11		1	3	5	7	9	12
E9	2	4	6	8	10	11		1	3	5	7	9	12
E10	2	4	6	8	10	11		1	3	5	7	9	12

Appendix 3 Mean scores (each script, all rating categories)

	Handwritten version					Typed version			
SCRIPT	AIE	CQ	VSS	Final		AIE	CQ	VSS	Final
A1	4.83	4.67	4.83	4.67		5.00	4.83	4.83	4.67
A2	4.50	4.33	4.00	4.33		5.00	4.50	4.17	4.50
A3	6.83	6.83	6.67	6.83		6.83	6.50	6.50	6.50
A4	4.50	5.00	4.67	4.67		4.33	4.67	4.67	4.50
A5	6.50	6.50	6.00	6.33		6.33	5.83	6.00	6.17
A6	4.17	4.67	4.50	4.50		4.50	4.83	4.33	4.50
B1	5.83	6.33	6.33	6.33		6.00	6.00	6.17	6.17
B2	5.83	6.17	6.17	6.17		6.17	6.33	6.00	6.17
B3	5.00	4.83	5.00	4.83		5.50	5.17	5.00	5.33
B4	5.33	5.00	5.00	5.00		5.00	5.17	4.67	5.00
B5	5.67	5.50	5.17	5.33		5.67	5.00	4.83	5.33
B6	7.00	6.50	6.33	6.50		6.67	6.67	6.33	6.50
B7	5.83	5.83	5.67	5.83		5.00	5.50	4.83	5.17
B8	4.83	4.83	5.00	5.00		4.83	5.17	4.83	4.83
C1	4.33	5.00	5.00	4.67		4.17	5.50	5.00	4.83
C2	5.50	5.33	5.50	5.50		4.67	4.83	4.67	4.67
C3	5.50	5.83	5.83	5.67		5.17	5.17	5.00	5.00
C4	5.00	5.17	5.00	5.00		5.00	4.67	4.67	4.67
C5	5.00	4.83	4.83	4.67		3.83	4.67	4.33	4.33
C6	4.33	5.33	5.50	5.17		3.17	4.33	4.67	4.17
C7	5.50	6.83	6.83	6.33		4.50	5.67	5.17	5.17
C8	3.50	4.17	4.50	4.17		4.17	4.33	3.67	4.00
D1	4.00	5.00	4.33	4.33		4.17	4.50	4.17	4.33
D2	5.33	5.83	5.33	5.50		5.33	5.33	5.50	5.33
D3	5.67	6.33	6.33	6.17		4.50	5.67	5.33	5.00
D5	5.33	5.50	5.17	5.33		5.33	5.67	5.33	5.50
D6	5.00	5.17	5.17	5.17		5.33	5.67	5.33	5.33
D7	4.83	7.17	6.83	6.33		4.50	6.00	6.17	5.33
D8	4.83	5.50	5.33	5.33		5.00	5.00	5.17	5.17
D9	5.67	7.83	7.50	7.00		6.17	7.33	7.33	6.33
D10	5.67	5.33	5.67	5.67		4.67	5.00	4.83	4.83
E1	6.17	6.17	5.83	6.00		4.83	5.17	5.33	5.17
E2	4.67	4.83	5.00	5.00		4.00	4.50	4.33	4.33
E4	6.83	6.83	6.67	6.83		5.50	6.17	6.17	6.00
E5	5.00	5.33	4.83	5.17		4.17	4.67	4.33	4.50
E6	5.17	5.33	5.50	5.33		4.67	4.83	5.17	4.83
E7	5.67	5.67	5.83	5.67		5.50	5.67	5.50	5.67
E8	6.83	6.83	6.67	6.83		6.67	6.67	6.67	6.67
E9	5.33	5.33	5.17	5.17		5.17	5.50	5.50	5.50
E10	4.50	4.67	5.00	4.83		4.50	4.83	4.67	4.83

AIE = Arguments, Ideas and Evidence
CQ= Communicative Quality
VSS= Vocabulary and Sentence Structure

Appendix 4 Instructions for legibility judgements

Legibility judgements

What I am interested in the **legibility** of the essay. I'm NOT talking about the quality of the syntax, coherence, etc (that is, the ENGLISH), but the handwriting and general layout. Things that are relevant will be letter formation and word legibility, and tidiness of script (for example if there are a lot of crossings out and insertions and it is difficult to follow where the text goes), ie.:

- Handwriting style / script (neatness, size, spread, slope, etc)
- Layout (tidiness, crossings out, insertions, etc)

Problems in reading the text could be occasional (particularly 'messy' bits) or could be constant. Please make a judgement of your overall impression of how easy it was to 'read' the essay, taking all of this into account.

1	Barely legible. Great effort needed to make sense of the handwriting and layout. Almost illegible.
2	Quite difficult to make sense of the handwriting and layout. Effort required.
3	Noticable difficulties in making sense of the handwriting and layout.
4	Minor difficulty experienced in making sense of the handwriting.
5	Generally quite easy to make out, minimal adjustment necessary.
6	Reads very smoothly – no problems at all with this handwriting.

The main thing is in interpreting these categories is to be consistent in the level of difficulty that is associated with each level on the scale.

A1 _____	B1 _____	C1 _____	D1 _____	E1 _____
A2 _____	B2 _____	C2 _____	D2 _____	E2 _____
A3 _____	B3 _____	C3 _____	D3 _____	E4 _____
A4 _____	B4 _____	C4 _____	D5 _____	E5 _____
A5 _____	B5 _____	C5 _____	D6 _____	E6 _____
A6 _____	B6 _____	C6 _____	D7 _____	E7 _____
	B7 _____	C7 _____	D8 _____	E8 _____
	B8 _____	C8 _____	D9 _____	E9 _____
			D10 _____	E10 _____

Appendix 5 Protocol instructions

I am going to ask you to rate a set of 20 Task 2 writing scripts (the essay). I would like you to rate them as far as possible in the usual way. However, there will be one important difference: we are conducting a study of the processes used by raters when they rate writing scripts, and I would now like you to talk and think aloud as you rate these scripts, while this tape recorder records what you say.

First, you should identify each script by the ID number at the top of the page. Then, as you rate each script, you should vocalise your thoughts, and explain why you give the scores you give. You may rate holistically or analytically, as is appropriate for the particular script.

It is important that you keep talking all the time, registering your thoughts all the time. If you spend time reading the script or the rating scale, then you should do that aloud also, so that I can understand what you are doing at that time. In order to make sure there are no lengthy silent pauses in your rating, I propose to sit here, and prompt you to keep talking if necessary. I will sit here while you rate and talk. I will say nothing more than give you periodic feedback such as 'mhm', although I will prompt you to keep talking if you fall silent for more than 10 seconds.

Trial

Perhaps you would like to try it out, to see what it's like. Let's look at this script, first of all. In a moment, I'll ask you to rate it as you would a normal script. First, introduce the script with its ID number - at the top of the page, here - then talk and think aloud while you rate it. I will sit here. Any questions? OK, I'll put the tape recorder on, and let's go.