CHAPTER FOUR

IMPRESSION JUDGEMENTS ON READABILITY

The aim of the present chapter is to test H1, i.e., that the readability of the writing products by the participants improved after instruction had ceased. More specifically, my objective is to compare the readability of the three pre-treatment and the three post-treatment essays in order to find out whether my prediction that the latter will be more readable can be sustained. I will begin by describing how the participants' performance in such essays was converted into readability scores, after which I will use those scores in order to test H1.

4.1 Converting writing performance into readability scores

To convert writing performance in the pre and post-treatment essays into readability scores, two preliminary steps had to be taken: first it was necessary to define how, and then by whom, the essays would be graded. These questions obviously presuppose the more fundamental question of what is meant by the term readability, which

was operationally defined in section 3.2.1 of chapter three. The definition draws on Clyne (1984) and Schema theory.

For Clyne, as stated in chapter two, the main factor of readability in English expository prose is clarity or whatever ensures the reader will gain access to text. Clarity or processing ease seems to be the most logical measure of the readability of the essays upon which this study is based inasmuch as the essays in question are expository texts, which means that their main function is to inform[1]. For an expository text to achieve its goal, its author must convey his message to readers clearly. The factors which ensure written discourse is clear are not direct functions of text, but of an agreement between writers and readers which is conveyed through text. This is in accordance with Schema theory, which maintains that what differentiates discourse from text is that the former is reader-dependent. That is to say, discourse depends on how a reader in a given context interprets text. In the words of Carrel (1982: 482),

> "In the schema-theoretical view of text processing, what is important is not only the text, its structure and content, but what the reader or listener does with the text."

Written discourse can therefore only said to be readable when the text that serves as a bridge between the writer and his interlocutors is clear, i.e., it causes no

processing difficulties to the latter. From this point onwards, readability will - therefore be assessed by measuring the extent to which written discourse conveys information to the reader in a clear way.

Having defined readability in this way, it was established that in the present part of the analysis clarity or processing ease would be measured via the impression method. Of the three different ways of marking essays described by Heaton (1975), the impression method was thought to be more appropriate than both the analytical and the error-count (or accuracy-based) methods.

The error-count method is by definition the one which has the least to do with processing ease or clarity, for an error-free text may not necessarily be easier to process than one which is dotted with errors. In fact, an error-free piece of written discourse may be so longwinded and unclear to the reader that it can be a lot more difficult to decode than a well-organized text tainted with a large number of spelling and grammar mistakes.

The analytical method, in turn, involves synthesing the evaluation of separate components of text, such as spelling, grammar, punctuation, fluency etc. It therefore consists of a series of impression marks which may be

useful when it comes to identifying specific problems in text, but which are probably very difficult to put together in a way which summarizes overall processing ease.

Unlike the error-count and analytical methods, the impression method offers a holistic perspective of discourse, which enables one to access and measure readability directly. That is to say, the impression method takes into account both the more central and the more ancillary factors which might affect overall readability, and automatically asigns them their proper weight, without the reader having to decompose readability consciously, into parts which would be extremely difficult, if not impossible, to synthesize into one meaningful overall score[2]. The impression method is also the most convenient method for marking of a large number of essays, as in the case of the 24 pre-treatment and 24 post-treatment essays relevant to this part of the analysis.

Using the impression method in order to assess readability obviously requires the use of a scale. According to the definition of readability adopted, I take it that written discourse ranked top on this scale is very clear and causes no difficulties to a given group of readers; written discourse ranked bottom on this same scale is not accessible to the same group of readers. The values in between these two extremes are theoretically limitless, but in practice they should be confined to a number which poses

no problems for the users of this scale (the readers) to distinguish between them. The following ordinal scale, which was validated by two native speakers of English who agreed that its intervals were semantically distinct from one another, was utilized to convert impression-judgements by a given group of readers into readability scores[3]:

1 = The essay is completely confusing and does not adequately convey its message.

2 = The essay is confusing and conveys its message with considerable difficulty.

3 = The essay is not always clear and conveys its message with some strain.

4 = The essay is clear and causes the reader few difficulties.

5 = The essay is very clear and gives no difficulties to the reader.

Insofar as the above scale is above all reader-dependent, it is obvious that it only makes sense if it is used by readers who are likely to share roughly the same amount of background knowledge on the content of the texts being evaluated. Because the pre and post-treatment essays in question were meant to be written according to the

conventions underlying the discourse of English expository prose, I decided to have them assessed by native speakers of English who shared a high degree of familiarity with this kind of discourse. At the same time, however, because impression judgements on readability can be quite significantly distorted by a knowledgeable reader's opinion on content, it was thought best to have them graded by a group of native-speaker readers who would not be overly influenced by factors which had more to do with opinions on the subject-matter of the essays than on readability. I therefore decided that all readers had to be equally unfamiliar with the subject-matter of the essays in question. Moreover, as James (1984) so aptly observed, the subject specialist tends to be overly tolerant with respect to communication breakdowns which his specialized knowledge enables him to overcome, and I specifically wanted to avoid making any allowances for such breakdowns. Thus what the readers chosen had in common was that they were native speakers of English highly familiar with the discourse of English expository prose but unfamiliar with the topics covered in the essays by the participants: they were sixteen Edinburgh University postgraduate students and members of staff working in areas different from those the participants were specialists in[4].

The 48 pre and post-treatment essays were distributed among the above readers so that in the end two different readers had to score the full set of pre and post-treatment essays

by the same participant. The reason for having distributed the essays in this way was that I did not expect any of the above readers to have the time to assess 48 essays (3 pre-treatment essays + 3 post-treatment essays x 8 participants) on topics unfamiliar to him all on the same day, let alone expect his or her judgements not to be influenced by fatigue[m]. The drawback of doing so, it could be argued, is that no matter how homogeneous the sixteen readers were expected to be, their interpretation of the values on the readability scale established would probably vary as a function of beyond control differences in personal interest in the topics of the different essays. However, the objective of assigning readability scores to the essays was to assess the progress of the participants along the succession of essays rather than to cross-compare their individual performances. Thus although it was crucial that all essays by the same participant be judged by a single reader, it did not matter so much that the essays by different participants should be assessed by different readers.

Once the scale and the readers who would use the scale to evaluate readability had been established, the essays by each participant were masked and shuffled into a random order so that their readers would be ignorant of the original order in which they had been written. The readers were then given the following instructions in writing:

a. Read the six essays enclosed in any order you wish, but all in one go.

b. Do not allow the technical words you are not familiar with stop you. You are to concentrate on your impression of the overall readability and clarity of the essays rather than on trying to understand their content in detail.

c. Give an impression mark to each essay according to the readability values set in the 1-5 scale provided. Half-marks allowed.

d. Write down your score to each essay next to its corresponding symbol on the scoring sheet enclosed.

The above instructions were repeated orally and the readers were allowed to make questions if they had any doubts concerning the procedure. No time limit was imposed for the task.

Having thus assigned the pre and post-treatment essays impression marks on readability, before handling them it was necessary to check whether the two respective readers of the sets of essays by the same participant had agreed often enough for me to feel confident about their ratings. Given ordinal scale used, the Spearman rank-order correlational analysis was the one chosen for this purpose.

Six out of the eight correlation coefficients were +0.5 or over, a figure that was accepted as indicating that there was sufficient agreement between six out of the eight pairs of readers. However, the remaining two coefficients obtained , +0.1 and -0.5, indicated that the former pair of readers had not reached any significant agreement, and that the latter pair had actually disagreed. This was rather problematic because the number of essays was relatively small, which meant that any statistical computation applied to the readability scores would be especially sensitive to such disagreements. In consequence, before proceeding any further, the two sets of essays in question had to be reassessed until some significant agreement by any two readers was reached. Each of these sets was therefore duly scored by a third reader, both of whom were again native speakers of English highly familiar with the discourse of English expository prose but unfamiliar with the topics of the essays in question. When the correlation coefficients were then recalculated, it was found that both third

readers had agreed more with one of the original readers than the original readers among themselves. The ratings given by the most discrepant original readers were therefore discarded at the expense of the new ratings provided by the third readers. The eight final pairs of readability scores and their respective correlation coefficients are summarized in table 4.1 below. The fact that it was not unduly problematic to obtain such positive coefficients in itself gives some indication that the method used to arrive at the readability scores was reliable.

Table 4.1: Readability scores assigned to the eight sets of pre and post-treatment essays plus correlation coefficient per pairs of scores (*scores on the left by first reader; scores on the right by second reader)

	PARTICIPANTS			
	Cida	Dony	Elisa	Gustavo
PAIRS OF SCORES PER ESSAY*	5 : 2	2 : 1	2.5 : 2.5	3.5 : 4
	4 : 2	3 : 2	2 : 3.5	3.5 : 4
	2 : 2	1 : 2	2.5 : 1	3 : 3
	3.5 : 1.5	3 : 3	3.5 : 4	3.5 : 1.5
	5 : 2	4 : 3	4.5 : 5	4 : 3
	3 : 1	5 : 2	4 : 4.5	4 : 5
COEFF.	+0.5	+0.5	+0.8	+0.5

Table 4.1 (cont.):

	PARTICIPANTS			
	Henrique	Silvia	Thelma	Wilson
PAIRS OF SCORES PER ESSAY*	4 : 4	3 : 2	3 : 2.5	3 : 3
	1 : 3	4 : 2	3 : 3	3 : 4
	5 : 4	4 : 3	4 : 4	3 : 2
	5 : 4	3 : 2.5	4.5 : 4.5	4 : 3
	4.5 : 3	5 : 3	5 : 5	4 : 4.5
	5 : 4	4 : 2	3.5 : 4	4 : 4.5
COEFF.	+0.7	+0.6	+1.0	+0.7

4.2 Were the post-treatment essays more readable than the pre-treatment essays?

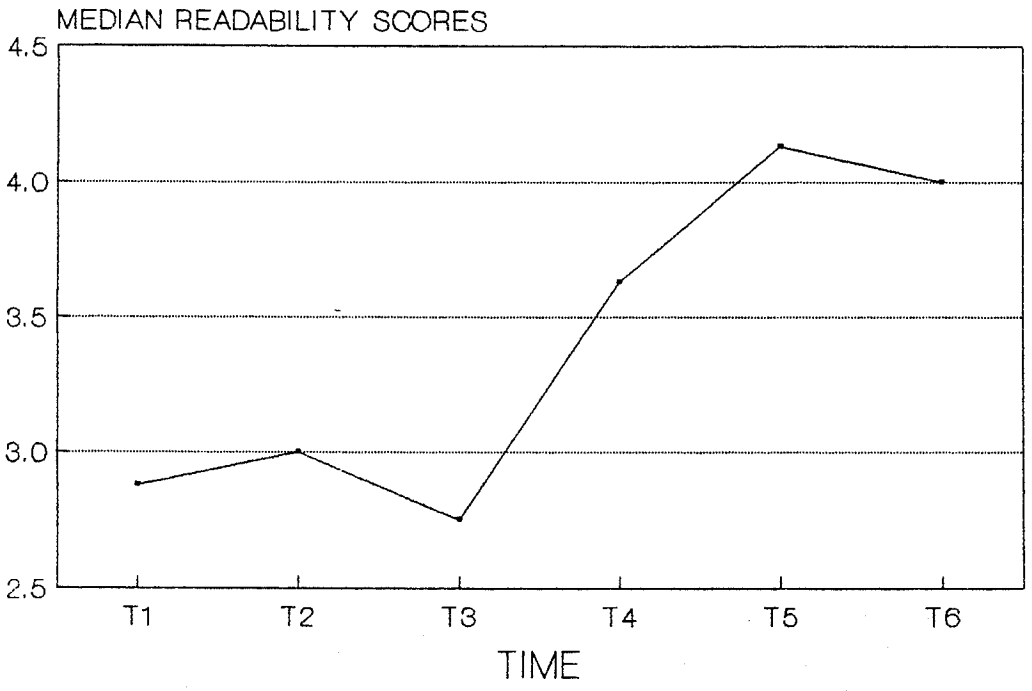
I shall now describe how the final two readability scores given to each of the 48 essays were processed, and how the readability of the pre and post-treatment essays were subsequently compared. Given the ordinal scale used, an option was made for non-parametric statistical methods.

The first step was to extract the median readability score for each individual essay so that the scores by all readers would be taken into account. Having obtained the median score for each essay, the next step was to unmask the essays and sort them out according to the order in which they had been written. That is to say, the eight median scores given to each of the three pre-treatment essays (T1, T2 and T3) and each of the three post-treatment essays (T4, T5 and T6) were distributed as required in a time-series design. Next, the median readability score for each T was computed. Table 4.2 below summarizes the median scores per essay and the overall medians per T, which were then mapped onto the graph in figure 4.1.

Table 4.2: Distribution of median readability scores per essay and overall median readability score per T.

PARTICIPANT	T1	T2	T3	T4	T5	T6
Cida	4.00	3.00	2.00	2.75	4.25	3.25
Dony	1.50	2.50	1.50	3.00	3.50	3.50
Elisa	2.50	2.75	1.75	3.75	4.75	4.25
Gustavo	3.75	3.75	3.00	3.75	3.50	4.50
Henrique	4.00	2.00	4.50	4.50	3.75	4.50
Silvia	2.50	3.00	3.50	2.75	4.00	3.00
Thelma	2.75	3.00	4.00	4.50	5.00	3.75
Wilson	3.00	3.50	2.50	3.50	4.25	4.25
MEDIAN	2.88	3.00	2.75	3.63	4.13	4.00

Figure 4.1: Median readability scores from T1 to T6



It can be seen from the gradients in figure 4.1 that the biggest improvement in readability occurred between T3 and T4 (+0.88). It can also be seen that the three post-treatment group medians (T4, T5 and T6) were higher than the three pre-treatment group medians (T1, T2 and T3), which is already an indication that the post-treatment writing products by the participants were more readable, and that the improvement which took place was maintained after the treatment had ceased.

To find out whether or not time or reading and writing practice alone (as opposed to instruction) could have affected these results, it seems appropriate to examine the curves pertaining to pre and post-treatment performance separately. It can be seen from figure 4.1 that before the treatment was introduced readability increased very little from T1 to T2 (+0.12) and then, from T2 to T3, dropped below T1 (-0.25). After the treatment had ceased, readability increased quite substantially from T4 to T5 (+0.5) and then dropped slightly from T5 to T6 (-0.13), to a point which was nevertheless above T4. The fact that readability both increased and dropped twice, once before and once after the treatment, suggests that time or reading and writing practice alone did not in themselves result in improved readability. In other words, neither the pre-treatment curve between T1 and T3 nor the post-treatment curve between T4 and T6 indicate that practising reading and writing, which is what the participants did during

those two phases of the experiment, or time alone, contributed towards a consistent increase or decrease in readability.

Since neither time nor reading and writing practice alone seemed to have affected the results in a specific direction, to find out more about how the post-treatment writing products by the participants compared with the pre-treatment equivalents, I found it legitimate to compare overall pre-treatment readability and overall post-treatment readability as two unitary blocks. Table 4.3 below summarizes the overall pre and post-treatment readability medians per participant.

Table 4.3: Comparison of overall pre and post-treatment readability medians per participant

PARTICIPANT	PRE median	POST median	CHANGE
Cida	3.00	3.25	+0.25
Dony	1.50	3.50	+2.00
Elisa	2.50	4.25	+1.75
Gustavo	3.75	3.75	0.00
Henrique	4.00	4.50	+0.50
Silvia	3.00	3.00	0.00
Thelma	3.00	4.50	+1.50
Wilson	3.00	4.25	+1.25
CENTRAL TENDENCY:	3.00	4.00	+1.38

The above results indicate that although there does not seem to have been any post-treatment improvement in readability in the essays by Gustavo and Silvia[6], the post-treatment overall readability medians for the essays by all other participants were higher than the pre-treatment equivalents. In addition to this, from the bottom row of

table 4.3 it can be seen that the central tendency for the group as whole (which was computed by extracting the median of the individual medians) leaves no doubt about evidence of a general improvement in readability. If this is interpreted in association with the fact that there were no significant fluctuations between the pre or post-treatment readability scores upon which those medians are based (before and after the treatment readability both increased and decreased), one might infer that the instruction provided during the experimental treatment is more likely to have been the cause of improvement than time or reading and writing practice alone. Evidence that the participants were able to produce more readable writing products after instruction had ceased is further strengthened by the fact that:

a. the group readability medians for T4, T5 and T6 were higher than the equivalent medians for T1, T2 and T3 (table 4.2);

b. the biggest improvement observed occurred from T3 to T4 (figure 4.1).

Although the present results are highly encouraging, it would be precipitate to attribute the improvement perceived to the specific instruction provided during the experimental treatment without examining its effects in further detail. After all, it could be argued that any type

of writing instruction could in the end promote some kind of improvement in readability. In other words, it would be wrong to equate the improvement perceived with the pedagogy tested during the treatment without having a measure of whether or not the peculiarities of the experimental treatment played an important role in such a development.

To draw any significant conclusion about the relationship between the instruction provided and the above evidence of improved readability, a more extensive analysis of the data is required. For the matter, I opted for analysing and interpreting only a selected part of the data - the post-treatment revisions of the pre-treatment final drafts - in much greater depth. The next three chapters will deal with that data, the last of which will finally examine the effects of instruction.

Notes to chapter four

1. Elsewhere in the literature this particular function has been referred to as transactional (Brown and Yule 1983), descriptive (Lyons 1977), ideational (Halliday 1970), referential (Jackobson 1960), and representative (Buhler 1934).

2. In a later part of this study (chapter six), readability will however be analysed in parts. It shall nevertheless be seen that no attempt will be made to add up the parts, although the overall picture they make will be discussed in the light of holistic impression judgements on readability.

3. Half-marks were allowed as a means of capturing differences finer than the wording of the values in the scale.

4. As stated in chapter three, the participants wrote essays in immunology, pharmacology, medicine, geology and communication studies. The native speaker of English readers responsible for evaluating those essays were specialists in applied linguistics, linguistics, cognitive sciences, artificial intelligence and anthropology. Care was taken to have the set of essays in communication studies assessed by the specialists in artificial intelligence, who were considered to be the readers who had had less contact with humanities. It will be seen in chapter six, however, that it was belatedly discovered that one of the applied linguists responsible for evaluating the essays by one of the pharmacologists (Silvia) was an experienced teacher of medical English.

5. According to Underhill (1982), one of the major sources of unreliability in the marking of written texts is that a single reader may assign different scores to the same essay from one day to the next. For this reason, it was thought important to have the essays marked all in one go.

6. See note 4 above and chapter six for a possible reason why Silvia's post-treatment writing products were not thought to be more readable.