A Critical Review of Readability and Comprehensibility Tests

Barbara Woods
Gianna Moscardo
and
Tanya Greenwood

Interpretive text appears on signs, exhibit labels, and maps, and in brochures, displays and guidebooks. Interpretive signs or labels are a particularly common form of communication with visitors. While interpreters hope their signs will inform, excite, educate and interest, many readers will be able to recall a sign that failed to attract or hold their attention.

For an interpretive sign to be effective at getting across its message to visitors it must:

1. be seen,
2. be legible,
3. be sufficiently interesting to visitors to actually be read,
4. be understood, and
5. be remembered.

Additionally, if the sign has the goal of changing visitor behaviour, it must also motivate visitors to change their attitudes and/or values and to translate the information given by the sign into changed behaviour patterns.

Each of the five stages listed above can involve multiple factors or design elements. For example, the first stage involves a consideration of the placement or location of the sign, while the second stage involves issues of type size and font, the type of material used to construct the sign, as well as colours and contrasts. Factors which can

Abstract
Many factors contribute to the effectiveness of interpretive signs and labels including placement, size, length, organisation of information, layout, and level and complexity of language used. Most authors agree that the key to success lies in keeping the visitor and their needs central to the design process. One way of examining the complexity of the language used by interpreters is to conduct a readability or comprehensibility test on interpretive text. This paper reviews the most commonly suggested tests, their origins, functions, limitations and strengths. The paper then reports on two exploratory studies comparing readability and comprehensibility measures calculated for a sample of interpretive signs in North Queensland. The results indicated that comprehensibility tests and directly asking for reader comments were much more valuable in revising text than the readability measures examined.
influence visitors' motivation to read include length, topics, use of headings, content of headings, use of illustrations and the style of language used. The reader is directed to Woods (1997) for a more detailed discussion of the full range of factors that can influence the effectiveness of an interpretive sign.

There are three main sets of factors that contribute to the success of an interpretive sign in the fourth stage. To be able to understand a piece of interpretive text, visitors must be able to understand the language used, to follow the structure of the argument, story or information being presented and be able to incorporate the information into their existing systems of knowledge. Of particular interest to the present paper is the issue of visitors being able to understand the language used by the interpreters.

Many texts and guides for the design and use of interpretive text suggest the use of readability tests as a way of ensuring that the level and complexity of language used is appropriate for the intended audience. A common message in this literature is that simplicity and brevity are desirable features of interpretive text, and that readability tests offer interpreters a guide to the level of simplicity they have achieved (Bitgood, 1996; Knudson, Cable & Beck, 1995; Kool, 1985; MacIntosh, 1976; Serrell, 1996; Sorsby & Horne, 1980, Stansfield, 1981). The following quote from Kaye summarises this message.

The two most common drawbacks of labels are that they are too long and are composed at too high a comprehension level. Labels written at a seventh to ninth grade level are comfortable for most adults. Studies reveal that adults like texts written at the fourth grade level even more. By counting the number of sentences and syllables in 100-word samples it is possible to measure the educational level of a text - and adjust it accordingly. The graph for estimating readability, developed by Edward Fry (1977) of Rutgers University Reading center, is a highly reliable tool for gauging comprehension level (1993, p. 22).

While not all the authors listed have been so supportive of the readability tests, most suggest that using a test is a useful procedure. It is also the case that while most of the these authors are careful to note that testing reading level alone will not tell the interpreter all they need to know, few analyse the readability tests themselves. Further there appears to be no agreement as to the best test to use. Kaye (1993), Bitgood (1996) and Sorsby and Horne (1980) support the Fry test, while Knudson, Cable and Beck (1995), MacIntosh (1976), and Serrell (1996) suggest Flesch’s Ease of Reading test. Knudson, Cable and Beck (1995) also suggest the FOG test, which in turn is also recommended by Kool (1985). Bitgood (1996) and Kool (1985) recommend the use of the CLOZE test and Davenish (1990) describes FORCAST.

The present paper was written with the aim of addressing the ambiguity in the interpretation literature over the use of readability tests. The paper will therefore describe some of the suggested tests, their aims and how they are used, and it will then examine the strengths and weaknesses of these tests. These strengths and weaknesses will be illustrated with examples from
two exploratory studies which adapted and applied various readability and comprehensibility measures to the text found on a series of interpretive signs in the tropical rainforests of North Queensland, Australia.

**Introducing readability tests**

The basic process in most readability tests involves counting the number of some combination of syllables, words, sentences and paragraphs to estimate the difficulty of the language level used. The purposes of using readability tests in interpretation are to ensure the language style is not too difficult for the average visitor, and to assist in avoiding unnecessary scientific jargon. Robert Gunning for example, developed the “FOG” index in the 1940’s to serve as a warning against “drifting into needless complexity in writing” (Gunning, 1968:12). Readability measures have generally been developed as an ‘objective’ measure against writing complexity, and to estimate the reading or education level required for comprehension of the text.

An examination of the reading and reading research literature reveals the existence of many readability formulae. Klare (1974), for example, summarises and describes 27 sets of formulae, some of which include up to 10 individual sub-formulae. Mayer, Marsiske and Willis (1993) suggest that there are more than 30 basic formulae from which many derivatives have been proposed. It also worth noting that new tests are developed on a regular basis. Harrison and Bakker (1998), for example, recently proposed two new measures based on lexical density and packet length, which refers to a group of words within syntactic punctuation marks. Many of these newer formulae are, however, complex and require detailed understanding of linguistic and grammatical terms and so it is not surprising to find that some of the earliest and simplest formulae are still the most commonly used. Generally it appears that the Flesch test for Reading Ease, as well as Fry, SMOG and FOG tests are widely used (Harrison & Bakker, 1998, Mayer et al., 1993; Newton, 1992; Reid, 1984). Appendix A contains the formulae and procedures for calculating some of these most commonly cited and used formulae. Studies comparing the simple and the more complex tests in terms of how well they predict reading performance, generally conclude that the small advantage in accuracy achieved by the more complex tests is not sufficient to justify the extra time and difficulty in calculating them (Klare, 1984; Mayer et al., 1993; Newton, 1992).

**Criticisms of readability tests**

The most common criticism of readability tests is that they are too simple and fail to consider, or examine, any of the many other variables which may influence reading or comprehension (Bitgood, 1996; Harrison & Bakker, 1998; Irwin & Davis, 1980; Klare, 1974 & 1976; Knudson et al., 1995; Serrell, 1996; Sorsby & Horne, 1996; Wagenaar, Schreuder & Wijhuizen, 1987; Weaver & Kintsch, 1991). Readability tests do not measure how interesting the material will be to the intended audience and a readability score can be computed even for random sequences of words or sentences that have no meaning (Meade and Smith, 1991; Newton, 1990). It is certainly true that a positive readability score does not guarantee that a piece of text can in fact be successfully read.
Another criticism is that readability tests do not provide any useful information on how to actually improve the comprehensibility of a piece of text (Klare, 1976; Mayer et al., 1993).

In considering these two criticisms it is important to remember that the tests were not intended to be measures of comprehensibility. They were originally developed with the aim of ranking school textbooks in terms of difficulty in order to assist teachers in the selection of appropriate texts for children of different ages. A related goal for many of the tests was to predict the reading age required to successfully use the text (Harrison & Bakker, 1998). In essence such tests are often challenged for not achieving things they were never meant to achieve. However, their use can be seductive and many of the developers themselves were aware of the limitations and add cautions to their use. Gunning (1968, p. 12), for example, warns that “nonsense written simply is still nonsense”.

Further, it should be noted that studies have found positive correlations between readability scores and other measures of reading ease and/or comprehension. Klare (1984) reviews several studies which found that readability test scores were related to:

1. the probability of readers actually reading a piece of text completely,
2. the amount of information remembered by readers,
3. the length of time taken to read a passage, and
4. readers’ ratings of difficulty.

Harrison and Bakker (1998) also found positive relationships between their new proposed readability measures, Flesch and FOG scores and readers’ ratings of readability. This would suggest that readability scores do tap into some aspects of text difficulty that are recognised by, and relevant to, readers.

A second and more challenging problem is that of what the scores actually mean. It is not unusual for different predictive formulae to yield disparate reading levels for the same passage of text (Borun & Miller, 1980). Also it is not clear what a reading grade score actually means (Kirkwood, Wolfe, Maynes, Millar, Sword & Word, 1980). Harrison and Bakker (1998) ask the question, “If a passage has a Reading Grade Level (RGL) of 7, does this mean that a person with an RGL of 7 would be able to understand 100% of the text, or 75%, or 50%” (p. 124). It is even more difficult to interpret these scores in situations other than the evaluation of school texts. Writers in these contexts require information on how many adults read at what grade levels. Unfortunately it is difficult to find adult literacy tests which provide results in terms of grade levels. Recent international literacy investigations provide scores on three measures which range from 0 to 500 and cross-tabulations of these levels with educational experience do not reveal consistent patterns (OECD, 1996). In an Australian study, for example, more than one third of the respondents with a university degree were placed in the two lowest levels of literacy (ABS, 1996).

**Comprehensibility tests**

One suggested solution to some of the problems associated with readability tests is to use a comprehensibility test (Bitgood, 1996; Kirkwood, et al., 1980; Kool, 1985; Newton, 1990). These tests involve deleting part of the text and investigating the effect of those deletions on comprehension. Examples include the CLOZE procedure (Taylor, 1953), where every fifth word is deleted, and the OPIN procedure (Denner & Pehrsson, 1994) where the second half of the sentence is deleted. Respondents are required to fill in the missing spaces to make sense of the text. Such performance-based tests are said to be preferable to traditional readability tests because they measure the actual comprehension of visitors and thus take into consideration the interaction between the text and the processing capacities of people (Denner & Pehrsson, 1994; Klare, Sinaiko & Stolurow, 1972; Sawyer, 1991). However, they are more expensive and time consuming to do because they require a sample of readers and so are not as appealing to many practitioners as the simple readability formulae. Similar problems also exist with methods involving readers rating various aspects of pieces of text (Harrison & Bakker, 1998).

---

1. Select a passage of text
2. Delete every fifth word and replace it with a blank. Ensure the blank spaces are of the same size.
3. The respondents must then guess the exact word that belongs in the blank.
4. Comprehension levels:
   **INDEPENDENT:**
   A person will not require assistance to attain a high level of understanding (CLOZE greater than 60%)
   **INSTRUCTIONAL:**
   A person will understand if provided with some instructional assistance (CLOZE between 40 - 60%)
   **FRUSTRATION:**
   A person is unable to read the material effectively (CLOZE less than 40%)

---

Table 1: The CLOZE Procedure (from Taylor, 1953)
Summary and aims of the present research

In summary there has been considerable discussion of the value of readability tests within the reading literature, but little critical assessment conducted within the interpretation literature. There is evidence that the tests can be useful for identifying potential problems with the reading difficulty of proposed text, but there is little consensus about which test is most useful. Again comparisons between tests have usually been conducted on school text book passages with little evidence of their comparative usefulness for examining interpretive text. One reason for this lack of research in interpretive settings is that many of the tests are designed for use with samples of 100 words taken from larger pieces of text. Few interpretive signs or labels achieve that length and so the formulae require modification.

The present research program had the overall aim of adapting and applying a series of the more frequently suggested readability tests to a sample of interpretive signs and comparing these to reader or audience based measures of difficulty or comprehension. The goal was to conduct some preliminary evaluations of the relative usefulness of the readability test for predicting audience perceptions of difficulty and levels of comprehension. Two exploratory studies were conducted to begin investigating the potential usefulness of readability tests in the development of interpretive text.

Study One: Comparing CLOZE, Fry, Flesch, SMOG and FORCAST Measures for Interpretive Signs

A sample of 20 signs were selected from a larger group of 153 interpretive signs from tropical rainforest areas in North Queensland, Australia. The reader is directed to Woods (1997), and Moscardo, Verbeek and Woods (1998) for a more detailed description of the larger sample of signs. The signs were chosen so that each had a minimum of 50 words and they were selected to cover as wide a range of locations and content areas as possible. The sample was also chosen to include signs produced by different public and private agencies. These selection criteria reflect some of those suggested by Klare (1984).

Fry readability scores were calculated for each sign. These tests were chosen because they were among those most commonly recommended or because the literature suggested that they were frequently used. Adjustments were made to the formulae for the SMOG, FORCAST and Fry tests to take into account the shorter text length of the passages and details of these are contained in Appendix A.

The 20 signs were also assessed in terms of levels of compre-
hension, using the CLOZE test. A total of 150 respondents were involved in calculating the CLOZE scores. Each sign was assessed using the CLOZE procedure by 15 respondents, and each respondent completed two signs. The respondents were all first year university students, thus the sample for this test had a uniform 12 years of education. The use of a sample with a uniform reading age is recommended by Kirkwood et al. (1980). Instructions for the administration of the CLOZE comprehensibility test are outlined in Figure 1.

Results

The first component of the analysis was to investigate the CLOZE scores for each of the signs. Table 1 provides the mean and median scores, as well as the standard deviation, range of scores and the percentage of readers falling into each of the three levels identified in Figure 1. In this table the signs have been ranked in order from the highest to the lowest mean CLOZE score.

Using the mean score as the main indicator four of the signs fall into the Independent category, 5 fall into the Frustration category and the majority fall into the intermediate Instructional category. It is worth noting that many of the signs in this Instructional category had illustrations and figures which could assist comprehension. Overall the ranking changes little if other measures, such as the median or the percentages of the sample in each category, are used. Therefore it was decided to use the mean CLOZE score as the major measure for further analysis.

The next step in the analysis was to investigate the readability scores and compare these with each other and the mean CLOZE scores. Table 2 lists the four readability scores and the mean CLOZE scores for the 20 signs. Again the signs were ranked in order from highest to lowest mean CLOZE score.

The most striking feature of this table is the lack of consistency between the five measures. Sign 5, for example, is rated as difficult according to the Flesch score, requires 16 years of education according to the FRY test, has a reading grade of 7 according to the SMOG test and a reading grade level of 10 according to the FORCAST test. Both the Fry and Flesch tests and Stansfield’s (1981) interpretation of FORCAST scores, predict that readers would have difficulty with the sign, but two thirds of the sample demonstrated few difficulties when completing the CLOZE procedure.
Cathedral Facts

The Cathedral Fig is a Banyan or Green Fig (Ficus Virens). Green Figs commonly have a strangling habit. A strangler fig such as the Cathedral Fig starts life high in the forest canopy where a seed is dropped by a bird or bat into the fork of a tree. The seed germinates and the fig may live on high as an epiphyte for many years. When conditions are favourable the fig sends fine cable-like roots down the host’s trunk to the ground. Roots act as a feeding tube and the plant grows rapidly. Roots fuse to encircle the host tree. By restricting the sap flow to the canopy leaves the strangler fig may finally kill its host. Strangler figs are a common feature of upland tropical rainforests.

Figure 3: Sign 12

On the Boardwalk Today

Watch out for Woompoo Pigeons (Ptilinopus magnificus)

Woompoo pigeons may be active around the Centre today as mates and nesting sites are being sorted out. Listen for their human-like call “wolack-a-woo” whilst on the walk. Although magnificently coloured they may be hard to spot. Woompoo have a burgundy chest, light grey head and iridescent green wings splashed with yellow . . . . . not your average shopping mall pigeon! They live entirely on the canopy and never come to the ground, although they may be observed noisily feeding in the lower branches of a favourite fruiting tree.

Figure 4: Sign 16

An examination of the actual sign (see Figure 2) provides some insights into the likely sources of the discrepancies in scores. The readability tests all assume that polysyllabic words are difficult or unfamiliar words, and that longer sentences will contain multiple concepts or concepts with modification or qualification. While it may be the case that these two assumptions are often appropriate, in the case of sign 5 many of the polysyllabic words are common ones such as rainforest, popular, excellent, and impressive. Also the sign has longer sentences but these are all descriptions of a place rather than abstract or complex discussions.

Sign 12 is an example of the opposite problem (see Figure 3). In this case most of the readability measures suggest that the sign should not be too difficult. Forty percent of the sample, however, could not complete more than forty percent of the blank spaces and overall the CLOZE test would suggest that most readers would have some difficulty with the sign. While signs 12 and 16 (see Figure 4) received similar CLOZE scores, one was rated as Fairly Difficult according to the Flesch test while the other one was rated as Fairly Easy. In both these cases the FRY and SMOG tests gave the signs the same scores, but the FORCAST scores differed.

Again an examination of the actual signs provides some insight into the results. The CLOZE test instructions were strictly adhered to in the present study. That is, respondents had to fill in the blank spaces with the exact word to be marked correct. In the case of sign 12 many of the deleted words (underlined in the Figures) had many synonyms or alternatives which reflected the correct meaning. The tenth deleted word, for example, was germinates. Ten out of the fifteen respondents replaced this with grows which was marked as incorrect even though the meaning or intended message of the sentence was understood. Thus the CLOZE scores underestimated the respondents’ levels of understanding.

Sign 19 (see Figure 5) provides another example of what appears to be a major inconsistency between the predictions of the readability tests and the findings of the CLOZE sample. In this instance all the readability tests agree that this sign should be very difficult to read, yet 80 percent of the sample correctly replaced between 40 and 60 percent of the blank spaces suggesting only moderate difficulty. In this case an examination of the sign reveals the use of several scientific labels.

Subtropical rainforest

This type of forest is found in the cooler uplands of Queensland to the lowlands of coastal New South Wales. In general they have less of everything typically found in tropical rainforests. Species which would indicate the presence of this forest type include Coachwood Ceratoptetulum sp. And Sassafras Doryphora sp. Ferns are common, with buttresses and lianas being absent.

Warm temperate rainforest

These rainforests occur at high altitude in cooler climates, from the Atherton Tablelands to eastern Victoria’s Gippsland. Vegetation in warm temperate rainforests is less diverse than in tropical rainforests with 3-15 species making up the canopy. Species common to temperate areas include Coachwood Ceratoptetulum sp. And Sassafras Doryphora sp. Ferns are common, with buttresses and lianas being absent.
for the plants described. These labels inflate the number of syllables and the number of words with more than three syllables and the length of the sentences. All these consequences are reflected in the higher reading difficulty predictions. In the context of the sentences, however, the scientific labels are not critical for understanding the main message of the sign. Also the placement of the blank spaces in this sign resulted in these difficult words remaining mostly intact. This latter feature suggests some caution is necessary when applying the CLOZE procedure to small passages of text.

Within the tests themselves the correlations were also generally low with the Flesch test most likely to be related to the others.

Summary

Overall all the tests suggested that the majority of signs would present many readers with some difficulties. Kaye (1993) suggested that text at a seventh to ninth grade was acceptable, but a fourth grade was preferable. He quotes no studies to confirm this conclusion and we assume that the grades refer to Fry scores as that is the test recommended. Using this indicator all the signs were much too difficult. Using Flesch’s categories 11 out of the 20 signs tested were rated as difficult or very difficult with only three classified as fairly easy. Unfortunately the inconsistencies between the test scores make any conclusions about difficulty hard to support. Many of those signs classified as difficult according to Flesch were not a problem according to CLOZE scores and vice versa.

The inconsistencies also present difficulties in using the scores to suggest improvements. If the scores had been highly correlated, groups of easier signs could have been contrasted with groups of more difficult signs. Comparisons of these groups could then have been made seeking to find patterns in the easier pieces of text that could provide guidance for improving the more difficult passages. Unfortunately no sign could be consistently given an easy or difficult label and so comparisons could not be made.

The CLOZE test, does, however provide some guidance for the interpreter or writer in that the response sheets can be examined for words which provide the same, or a similar, meaning, and which are consistently preferred by respondents. In other words we can examine the words the respondents used to replace the deleted words to find easier or more familiar alternatives. Table 4 presents some of the findings from the present CLOZE tests.

<table>
<thead>
<tr>
<th>Original sentences with deleted words underlined</th>
<th>Alternatives suggested by respondents (number of respondents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>If a dinosaur were seen today it would be alive (11) here (3)</td>
<td></td>
</tr>
<tr>
<td>From giant trees to small species on the forest floor (from a sign about palms) plants (5) palms (3)</td>
<td></td>
</tr>
<tr>
<td>In general they have less of everything typically found in tropical rainforests. usually (3) that is (7)</td>
<td></td>
</tr>
<tr>
<td>These conifers are important representatives in deciphering fossil records. tools (2) assets (2) sources (2)</td>
<td></td>
</tr>
<tr>
<td>Some plants have special features that allow them to survive on the forest floor. help (10)</td>
<td></td>
</tr>
<tr>
<td>Others cannot cope with the gloom and must be more aggressive in their quest for light. survive (9) live (3) become (9) will (3) have to (2)</td>
<td></td>
</tr>
</tbody>
</table>

* Total number of respondents for each sign was 15.

<table>
<thead>
<tr>
<th>Table 3: Correlations between the tests.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flesch</td>
</tr>
<tr>
<td>Fry</td>
</tr>
<tr>
<td>SMOG</td>
</tr>
<tr>
<td>FORCAST</td>
</tr>
<tr>
<td>CLOZE</td>
</tr>
</tbody>
</table>

* Significant at p<0.05.
As you journey down the Cardwell Range the major habitats of the Wet tropics can be seen. Most of the Wet Tropics is covered by tropical rainforest, but bordering the rainforest are other important habitats such as open forest, melaleuca swamp and mangroves. The sudden transition from open forest to rainforest can be easily seen. Open forests, where eucalyptus and wattles grow, are drier, sunnier places, while rainforests are covered with leafy, humid places. A layer of leaf litter can be found at the base of rainforests. Leaf litter comprises of fallen leaves, twigs, branches, bark and wood. Leaf litter breaks down to form humus or decomposed vegetable matter. Growing from the trunks of many rainforest trees are plants known as epiphytes. These plants can also grow from rocks. Epiphytes do not harm the host tree, or are not parasitic, as they do not penetrate the tree’s living tissue. The humid atmosphere inside the rainforest is caused largely by the great number of transpiring plants giving off water vapour and the absence of drying agents such as wind and bright sunlight. The bark of rainforest trees is often coloured and mottled by lichen. Lichen is made up of a green plant (algae) in partnership with a fungus. Rainforests are characterised by a dense canopy where at least 70% of sunlight is blocked. Rainforest trees have special rounded points called driptips. Driptips help the leaves shed moisture after rain. 

The text was designed to be interspersed through a large exhibition panel with illustrations.

**Study Two: Comparing Fry, Flesch, SMOG and FORCAST Measures with Directly Asking Readers**

The second study was developed both as a specific evaluation requested by interpretive staff and as a further opportunity to examine the value of readability measures. Given that the previous study suggested some difficulties were likely with applying readability tests to small passages of text the present study examined a larger passage of text developed for a single display. The previous study also suggested that the CLOZE test may be unreliable with shorter passages and so this study examined an alternative approach. In this study respondents were asked to read the text and to underline any words or phrases that they did not know to understand. The evaluated text had been developed for an interpretive display on the national parks of North Queensland.

A sample of 369 respondents were approached at various tourist attractions in the region where the interpretive display was to be built. The aim was to gather information from a group likely to be visitors to the display. Just under half of the sample (49%) were local residents travelling through the region for recreation or business, with 29 percent being Australian visitors from outside the region and the rest being international visitors. The mean age of the respondents was 36 years and the sample was evenly split between males (46%) and females (54%). Just under one third of the sample (32%) had completed secondary education, 15 percent had a vocational qualification and 53% were studying or had gained a university degree. This high proportion of young university graduates reflects the dominance of young independent long stay travellers (referred to locally as backpackers) in the profile of visitors to the region.

**Results**

Analyses of the proposed text (see Figure 6) resulted in a SMOG score of 11, a Fry score of 12, a FORCAST score of 12 and a Flesch ease of reading score of 77. According to Flesch the text is in the Fairly Easy category, but the other tests suggest a difficult piece of text. Overall the majority of the sample of potential readers had few problems with the text. Just over half (53%) stated that they could find nothing in the text that they could not understand or that presented them with any difficulties. A further 19% underlined one word or phrase, 10% underlined two words or phrases, 15% underlined three to five words or phrases and 3% underlined more than five words or phrases. The maximum number of problems identified was nine and this figure applied to only three out of the 369 respondents. Table 5 contains the words and phrases that visitors said were difficult or unfamiliar.

The relationship between levels of education, and the number of words or phrases seen as difficult or unfamiliar, was also examined. Table 6 provides a summary of the results. No statistically significant relationship was discovered. In general there was a very slight trend for increasing education to be associated with fewer problems. Even with this trend, however, 47% of those who stated they had only completed primary education did not acknowledge any problems with the text. An examination of the words and phrases underlined suggested that some of the difficult words were specifically relevant to Australia and so analyses were conducted to investigate any differences in understanding between Australian residents and international visitors. On this occasion there was a statistically significant difference (Chi-square = 28.6, p<0.001) reflecting a greater proportion of international visitors who had difficulties with the text (see Table 7).
Summary

As in the first study the readability tests gave different results for the same piece of text. Also, as in the first study, the readability tests were not good indicators of potential reader responses. In this instance the sample of readers did not suggest any major difficulties with the text. A list of unfamiliar and problematic words were generated, but even the most commonly mentioned words failed to be a problem for even half of the sample. Given that the text was read without any of the planned accompanying illustrations it seems likely that very few readers would experience difficulty understanding the text in the actual display.

Overall conclusions

The readability tests examined in the present study gave very inconsistent results and none of the tests did a very good job of predicting readers' responses. The CLOZE test was found to be useful both in suggesting levels of difficulty and in providing guidance for improving the comprehensibility of the text. Directly asking potential readers to identify difficult or unfamiliar words and/or phrases was also a valuable approach. In this case no numerical index was provided, but the procedure clearly identifies words and phrases that require modification. Also this technique has the advantage of considering all potentially problematic items. The authors believe that the results from the CLOZE procedure and from directly asking for feedback from a sample of potential readers, support the recommendation that interpreters engage in some formative evaluation of proposed text. The results do not, however, support the use of the readability tests analysed in this study (FRY, SMOG, FORCAST or Flesch's Ease of Reading test). It may be possible that other readability measures, especially those developed in more recent times may be valuable and this highlights the need for more systematic analyses of readability tests for interpretive text.

Table 5: Responses to Proposed Text.

<table>
<thead>
<tr>
<th>Words/phrases Not Understood/ Unfamiliar</th>
<th>Number of respondents (% of the sample)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epiphytes</td>
<td>95 (26%)</td>
</tr>
<tr>
<td>Melaleuca/melaleuca swamp</td>
<td>88 (24%)</td>
</tr>
<tr>
<td>Buttress/buttressed roots</td>
<td>59 (16%)</td>
</tr>
<tr>
<td>Driptips</td>
<td>46 (13%)</td>
</tr>
<tr>
<td>Transpiring</td>
<td>26 (7%)</td>
</tr>
<tr>
<td>Humus</td>
<td>22 (6%)</td>
</tr>
<tr>
<td>Lichen</td>
<td>20 (5%)</td>
</tr>
<tr>
<td>Wattles</td>
<td>18 (4%)</td>
</tr>
<tr>
<td>Mottled</td>
<td>12 (3%)</td>
</tr>
<tr>
<td>Cardwell Range</td>
<td>9 (2%)</td>
</tr>
<tr>
<td>Transition</td>
<td>7 (1.8%)</td>
</tr>
<tr>
<td>Wet Tropics</td>
<td>6 (1.6%)</td>
</tr>
<tr>
<td>Parasitic</td>
<td>6 (1.6%)</td>
</tr>
<tr>
<td>Bordering</td>
<td>4 (1%)</td>
</tr>
<tr>
<td>Habitats</td>
<td>3 (%8)</td>
</tr>
<tr>
<td>Mangroves</td>
<td>3 (8%)</td>
</tr>
<tr>
<td>Leaf litter</td>
<td>3 (8%)</td>
</tr>
<tr>
<td>Canopy</td>
<td>3 (8%)</td>
</tr>
<tr>
<td>Moon canopy</td>
<td>3 (8%)</td>
</tr>
<tr>
<td>Base of rainforests</td>
<td>2 (6%)</td>
</tr>
<tr>
<td>Rounded leaves</td>
<td>2 (6%)</td>
</tr>
</tbody>
</table>

Table 6: Crosstabulation of Levels of Education by Number of Words/Phrases identified as Difficult or Unfamiliar.

<table>
<thead>
<tr>
<th>Number of Words/Phrases Identified</th>
<th>% of Sample with Primary Education</th>
<th>% of Sample with Secondary Education</th>
<th>% of Sample with Vocational Qualifications</th>
<th>% of Sample with University Experience or Qualifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>47%</td>
<td>46%</td>
<td>53%</td>
<td>54%</td>
</tr>
<tr>
<td>1</td>
<td>23%</td>
<td>23%</td>
<td>10%</td>
<td>19%</td>
</tr>
<tr>
<td>2</td>
<td>6%</td>
<td>9%</td>
<td>6%</td>
<td>12%</td>
</tr>
<tr>
<td>3 to 5</td>
<td>18%</td>
<td>18%</td>
<td>28%</td>
<td>13%</td>
</tr>
<tr>
<td>&gt;5</td>
<td>6%</td>
<td>4%</td>
<td>3%</td>
<td>2%</td>
</tr>
</tbody>
</table>

Table 7: Crosstabulation of Australian residents and International Visitors by Number of Words/Phrases identified as Difficult or Unfamiliar.

<table>
<thead>
<tr>
<th>Number of Words/Phrases Identified</th>
<th>% of Sample Who Were Australian Residents</th>
<th>% of Sample Who Were International Visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>58%</td>
<td>31%</td>
</tr>
<tr>
<td>1</td>
<td>19%</td>
<td>19%</td>
</tr>
<tr>
<td>2</td>
<td>10%</td>
<td>11%</td>
</tr>
<tr>
<td>3 to 5</td>
<td>11%</td>
<td>34%</td>
</tr>
<tr>
<td>&gt;5</td>
<td>2%</td>
<td>5%</td>
</tr>
</tbody>
</table>
References


1: SMOG test (from Kool, 1985)

1. Count 10 consecutive sentences near the beginning of the text, 10 in the middle and 10 at the end.
2. In the 30 selected sentences, count every word of 3 or more syllables.
3. Estimate the square root of the number of polysyllable words counted.
4. Add 3 to the approximate square root.

As a formula: \( \text{SMOG} = 3 + \left( \frac{30}{\text{number of sentences in sign}} \right) \times \frac{X}{\text{number of words with 3 or more syllables in the sign}} \)

Variation for interpretive signs:

Interpretive signs are not likely to have 30 sentences. Thus the formula must be adjusted to allow for fewer than 30 sentences.

\( \text{SMOG} = 3 + \left( \frac{30}{\text{number of sentences in sign}} \right) \times \frac{X}{\text{number of words with 3 or more syllables in the sign}} \)

2: FORCAST test (from Davenish, 1990)

\[ \text{Reading grade level} = 20 - \frac{[\text{the number of 1 syllable words in a 150 word passage}]}{10} \]

Variation for use with interpretive signs

\[ 20 - \left[ \frac{\frac{150}{\text{number of words in sign}}}{10} \right] \times \frac{X}{\text{number of 1 syllable words in sign}} \]

3: Flesch Ease of Reading (from Flesch, 1948)

1. Count the total number of sentences and words in the text.
2. Count the number of syllables in the text.
3. Calculate the average sentence length in words (the number of words divided by the number of sentences).
4. Calculate the number of syllables per 100 words (number of syllables divided by number of words multiplied by 100).
5. Calculate the Reading Ease score with the following formula.

\[ \text{Ease of Reading} = 206.835 \times (\frac{\text{the number of syllables per 100 words}}{10}) - (1.015 \times \text{average sentence length}) \]

4: FRY test (from Fry, 1977)

1. Count the total number of sentences.
2. Calculate the average number of sentences per 100 words (number of sentences divided by number of words multiplied by 100).
3. Count the total number of syllables in the passage.
4. Calculate the average number of syllables per 100 words. (number of syllables divided by number of words multiplied by 100).
5. Use the graph to calculate reading level.