

TECHNIQUES FOR EVALUATING DRAFT LEGISLATION¹

*Duncan Berry*²

Outline: This paper advocates selective usability testing of draft legislation and canvasses various methods by which testing might be carried out.

Why usability tests of draft legislation are a good idea

People do not read legislation for pleasure. They read it only when they want to find out what the law is on a particular matter or when they want to solve a problem that has legal implications. When trying to understand a particular piece of legislation, readers need the information they are seeking to be presented clearly, precisely and in the first place they look. They judge the usability of legislation by how quickly it helps them find the information they are looking for or to solve the legal problem that is confronting them. If they cannot do this, they complain.

So how can legislative counsel optimise legislative documents so that their various audiences can better understand and make use of them? As I see it, a legislative document must meet three requirements.

- It must be sufficient, that is, contain the necessary information,
- It must be precise, that is, contain the correct information, and
- It must be usable, that is, be organised and written so that all those who have to use it can find what they need and can understand what they find in the time that they are willing to spend on it.

Usability is as important to a document as sufficiency and precision, because if the document's audiences cannot find or understand the information, they will not even get to find out whether it is precise or sufficient. Consequently, there must be ways to test the sufficiency, precision and usability of legislative documents. For usability, we need tests that show whether people who must use those documents can find what they need and can understand what they find.

In recent years, a number of methods have been developed for evaluating documents. What follows is an outline of these methods, with an indication of their respective advantages and disadvantages. The relevant research literature puts typical methods for evaluating text quality into three classes. These classes are—

- text-focused approaches,
- expert-judgment-focused approaches,

¹ This article was first published in *The Loophole* in March 1997 and is based on a paper presented at the CALC conference held in Vancouver, September 1996. Copyright: The Commonwealth Association of Legislative Counsel and the author.

² Deputy Government Crown Counsel, Law Drafting Division, Hong Kong Attorney General's Chambers.

- and reader-focused approaches.

Methods for evaluating text quality

(1) Text-focused evaluation

Text-focused methods function by asking a person (or a computer) to examine a text, attend to a set of text features and assess text quality by applying principles or guidelines that have been developed from ideas or research about how readers at a certain level and of a certain background will probably respond. There is little or no reader input. These methods include

- readability formulae,
- computer based stylistic analysis programs,
- guidelines and maxims, and
- checklists.

Readability formulae

A readability formula is a mathematical equation that is applied to prose texts in an effort to predict how difficult the text will be for a given group of readers. When a readability formula is applied, you get a score, which is some number between 0 and 100 or a reading grade formula³. Formulae are commonly used to see if a text meets a predetermined numerical goal or to compare two versions of a text.

Readability formulae are easy to learn, easy to use⁴ and inexpensive. They require no input from readers or testers and they provide an impartial and objective measure. So why are readability formulae not a good idea? After all, back in 1985, their use was enthusiastically supported by the Hon. J. H. Kennan, the former Attorney-General of Victoria, who said that in future all Victorian legislation would be subject to the “Flesch Test”. The problem is that readability formulae are used in circumstances for which they have not been tested. Furthermore, they are used not simply as a measure of textual comprehensibility but as the *only* measure.

Studies have shown that readability formulae do not validly or reliably predict how intelligible documents are to their readers. One study⁵, which involved rewriting some jury instructions so that they were more intelligible, found that improved comprehensibility did not always produce better readability scores. In fact, changes that improved comprehensibility often resulted in worse readability

³ Usually this is a school reading grade. There are hundreds of kinds of readability formulae. The most common one is the Flesch Reading Ease Scale, which is based on sentence length and the number of syllables per 100 words. The higher the number, the easier the text should be to read.

⁴ Computerisation makes formulae even easier to use.

⁵ Charrow, R.P., and Charrow, V.M. Making legal language understandable: “A psycholinguistic study of jury instructions”, *Columbia Law Review* (1979): 1306-1374.

scores! Another study⁶ showed that what mattered to readers was not the readability score but that the tested passage had fewer ideas in each sentence and that the connections between the ideas were clearer.

Now that computerised readability formulae are available, legislative counsel may be tempted to use them as guides when writing or editing a draft Bill or regulation. But even the advocates of readability formulae agree that this is an inappropriate use of readability formulae. Legislative counsel who merely shorten sentences and change words to get a better readability score miss the point. A readability formula only *correlates* certain features with reading difficulty: the features do not *cause* the difficulty. Readability formulae can seduce counsel into drafting short, simple legislative sentences, but, as many lawyers experienced in reading legislation may have suspected, sentences can be difficult to read simply because they are too short⁷. Very short sentences in fact inhibit the flow of ideas. It is not length that causes the difficulty in sentences: it is features such as passive verbs, noun strings and nominalisations⁸. A noun string, for example, is often shorter than the more understandable phrase that untangles the string. Similarly, sentences containing nominalisations are often shorter than the same sentences that use more understandable verb phrases⁹. The basic problem with readability formulae is that they are mechanistic. They do not interpret context, meaning, grammar or content.

Another problem with readability formulae is that proponents assume that readers are text-processing machines. However, cognitive psychologists and psycholinguists have shown that readers read a text not from the bottom up but from the top down. Readers construct meanings on the basis of schemes that they bring to the text. They create expectations about the direction the text is taking and look for words and sentences that satisfy or negate those expectations. They also look for contextual material, such as signposts and explanatory introductions, to help them to construct those meanings.

Yet another major problem with readability formulae is that they measure only those features that can be counted. But many factors for which there are no objective measures influence how compressible a text is, factors that may be even more important than the length of sentences and the words used in the text. Three critical factors that readability formulae do not measure are content, organisation and lay out. In one study involving more than 50 life assurance policies, all of which satisfied a Flesch test, it

⁶ Kintsch, W., and Vipond, D., "Reading comprehension and readability in educational practice and psychological theory", in Proceedings of the Conference on Memory, ed L Nilsson Hillsdale, NJ: Erlbaum, 1977.

⁷ See study by P.D. Pearson, "The effects of grammatical complexity on children's comprehension, recall and conception of certain semantic relations", Reading Research Quarterly, 10(1974), 155-192.

⁸ i.e. nouns made out of verbs.

⁹ See study by Flower, L, Haves, J.R., and Swarts, H. "Revising functional documents", in New Essays in Technical and Scientific Communication: Research Theory, Practice, ed. P.V. Anderson, J. Brockman, C.R. Miller, Farmingdale, NY: Baywood, 1983, 90-108.

was found that most still hide important information under obscure headings¹⁰. The sentences may be shorter but the reader does not know where to look for it. And because readability formulae can be used only to measure straight text, they provide no assistance on graphics or typography.

At best, readability formulae should be used as a screening device for an old document. If a text that was not created with a readability formula in mind gets a poor readability score, it almost certainly needs to be re-organised. It is probably not even a good idea to use a readability score as one of several criteria for developing a readable Bill. There are at least two reasons for this. Firstly, any particular score from applying a formula is arbitrary. Secondly, there is a danger that legislative counsel who are held to a readability requirement would end up writing a formula. When a readability formula is one of the yardsticks for a document, all other measurement tools tend to be ignored.

Computer-based stylistic analysis programs

Computer-based programs typically work by assessing readability using one or more standard formulae and by counting passive constructions, misspellings, and numbers of simple, compound or complex sentences and then by providing the evaluator with a statistical summary of the text problems by assigning particular features an average score through comparison of the use of the text feature (the number of passive verbs for example) against the proportion used in a “good text” template¹¹. Most of these programs cannot address the kinds of grammatical problems that poor writers often create. The fundamental drawback of most of the programs is that “they rely too much on lookup tables instead of a parser to determine the role words play in a sentence”¹².

Guidelines and maxims

Guidelines and maxims are perhaps the most popular text-focused method in use. They are usually aimed at providing advice on the linguistic, stylistic or graphic features of text. From the legislative counsel’s point of view, such guidelines as “omit needless words” or “use shorter sentences” are of little help. Either they are too vague and too generic as in the first example or require counsel to assume that all writing tasks are alike and involve the same simplistic prescriptions. Furthermore, counsel may have difficulty in deciding when and how to apply guidelines. Guidelines that are applied too rigidly can have the effect of stifling solutions to rhetorical problems. In sum, guidelines are not likely to help legislative counsel to adapt their drafts to the unique features of a given rhetorical situation.

¹⁰ Redish, J. *Beyond readability: How to write and design understandable life policies* Washington DC: American Council of Life Insurance, 1984.

¹¹ e.g. Unix’s Writer’s Workbench and the GM Star program. Two style checkers are worthy of note Grammatik III for IBM type PCs and Macproof for Macintosh PCs.

¹² See Richardson, S., Creed, W., and Chandler, R., “Critique as a teaching tool for writing classes”, in *The Dynamic Text Guide*, 9th International Conference on Computers and the Humanities (ICCH) and 16th International Association for Literary and Linguistic Computing (ALLC) Conference, Toronto: University of Toronto, Centre for Computing in the Humanities, June 5-10, 1989, pp. 57-58.

Checklists

Another text-focused method, checklists, typically works in one of two ways. One is where the evaluator (i.e. drafter or editor) is prompted to consider certain specified issues. Many checklists depend on recommending visual or verbal text features to use or those that should be avoided or used sparingly. Other checklists are in essence additive weighting procedures that ask the evaluator to rank the text's features according to a quality scale and then to assign a score to the text.

A disadvantage with checklists arises because of the difficulty of deciding which text features are most important and of assigning weights or numerical values to text features. There is usually disagreement about the value to be attributed to any particular text feature. Moreover, checklists usually fail to ask evaluators to judge the use of text features in relation to the given rhetorical context. For example, some checklists caution against using the passive voice, even though there are many rhetorical situations when its use is the most effective and appropriate linguistic choice.

(2) Expert-judgement focused evaluation

Methods involving the application of expert judgement constitute another widely used set of evaluation procedures¹³. These methods include:

- peer review,
- editorial reviews, and
- external reviews.

The first and third of these methods have been used for a number of years in at least one Australian State Parliamentary Counsel's Office.

Peer Review

With peer review, people who share a common background evaluate texts for matters of style, consistency, and the like. Peer reviews can be very informative in pointing out text problems, allowing the drafter to draw on the multiple perspectives of other counsel. Peer reviewers tend to be good at recognising stylistic issues at both the macro and the micro levels. Peers can also be helpful in making suggestions to solve problems involving organising the text.

One disadvantage of this method is that legislative counsel may receive divergent opinions about the problems that the text will create for readers. Another is that peer reviews can suffer from evaluators who work too frequently with texts of similar kinds and subject-matter. When evaluators always work with the same kind of texts, they can become insensitive to audiences' likely responses to texts of the same kind¹⁴. A further concern is that the method can be a way of socially constructing and

¹³ By experts, I mean individuals who have a lot of knowledge about the text, its audience or, in this instance, legislative drafting.

¹⁴ Bond, S.J., Hayes, J.R., and Flower, L.S., "Translating the law into common language: A protocol study",

institutionalising certain styles (as probably occurred with legalese for example).

Review by experts in the field

Reviews by subject-matter experts (SME) usually involve content evaluations of text, with a view to finding deficiencies in coverage, accuracy, authenticity of completeness. Such reviews are intended to provide detailed information about the ways in which the content of the text is inaccurate or misleading.

Although this method can provide valuable feedback about difficulties with a text, it is probably unwise to use it in isolation. It seems that research¹⁵ is demonstrating that topic knowledge is sometimes a hindrance instead of an aid and that experts in the field are not always the best persons to ask about text quality. Readers with a high topic of knowledge were found to be very poor in judging how lay readers would understand the topic.

Editorial review

Editorial in-house reviews are typically carried out by editorial staff who check for such matters as style, consistency, specifications and use of conventions. Traditionally, editorial reviews focus on grammar and mechanical issues. Editorial reviews used to be quite mechanical and rule-oriented. However, in recent times their scope seems to have been expanded to cover organisation, presentation, readability, coherence and accuracy. Editors now tend to see their role as a complex hierarchy of skills and perceptual abilities. They have become much more concerned with ways of improving the text than before. It seems that the definition of editorial review is gradually changing from “editing” to “revising”. Rather surprisingly, there does not seem to have been much research into the editorial review process. However, it is commonly believed that experienced editors are much more skilled than some writers in identifying audiences’ needs and in making effective linguistic and rhetorical choices that meet those needs.

External review

In some circumstances it is impractical or even undesirable to use insiders to conduct an internal review. To overcome the problem, external reviews are sometimes conducted for judging text quality. An organisation that wants critical feedback about a particular text may engage a document design or graphic design consulting agency to carry out a review of the text.

One kind of external review, known as *text features evaluation*, evaluates the relative goodness of a text by assessing the design of visual or verbal features. Text is analysed in terms of key features, such

Document Design Project Technical Report No. 8, Pittsburgh: Carnegie-Mellon University, Communication Design Centre, April 1980.

¹⁵ Hayes, J.R., Schriver, K.A., Baustein, A., and Spilka, “If it’s clear to me, it must be clear to them: How knowledge makes it difficult to judge”, paper presented to the American Educational Research Association Conference, San Fransisco, April 1986.

as style, tone, content, format, and so on.

Another kind of external review uses holistic rating methods to judge text quality. This method “is a quick, impressionistic, qualitative procedure for sorting or ranking samples of writing. It is not designed to correct or edit a piece, or to diagnose its weaknesses. Instead it is a set of procedures for assigning a value to a writing sample according to previously established criteria”¹⁶. Holistic rating refers to the set of methodologies used to arrive at a total impression of a text.

Further kinds of external review are *general impression marking* and *primary trait scoring*. General impression marking is a method in which the evaluator fits a writing sample into an ordered ranking on the basis of the total impression created by the document. The defining characteristic of this approach is that sample documents are weighed against each other rather than against a predetermined set of criteria¹⁷. The relevant criteria are arrived at inductively either by the test organisers or by the evaluators themselves. Often test organisers will select a set of “anchor” documents that represent the range of good to poor texts that the judges can expect to see. Evaluators are then trained to judge a text against the anchor documents.

Primary trait scoring¹⁸ is different in so far as it provides testers with a scoring guide carefully adapted for the judging task. It thus uses a set of explicit criteria to judge text quality. Testers are then trained to evaluate texts using the agreed set of text features, such as style, organisation or coherence. Although the procedure seems quite straightforward, studies¹⁹ show that it is very difficult, and sometimes impossible, for a group of evaluators to agree on a set of criteria and to invoke such criteria consistently and reliably. According to Charney, “in spite of training, readers’ judgements are strongly influenced by salient, though superficial, characteristics of writing” (such as spelling, length and unusual words). Consequently, there are serious concerns about the reliability of holistic scoring procedures.

Yet another kind of external review is the consumer advocate review conducted by people who are concerned with judging the quality of text from the perspective of consumers. The evaluators are concerned with legal and other implications of poorly designed text. Consumer advocate reviews usually use weighted scoring models or scaled surveys.

A further kind of external review is what is known as *the gatekeeper review*. With this kind of review a text is evaluated by a group of individuals who are responsible for disseminating a text (such as health professionals for example). Gatekeeper reviews can be useful both when planning and revising text.

¹⁶ Charney, D., “The validity of using holistic scoring to evaluate writing: A critical overview”, *Research in the Teaching of English*, vol. 1(1984), pp.65-81.

¹⁷ See Charney, *ibid*.

¹⁸ Developed by Lloyd-Jones, R., “Primary trait scoring”, in *Evaluating Writing*, C. Cooper and L. Odell (ed.), Urbana II: National Council of Teachers of English, 1977.

¹⁹ e.g. see Grobe, C., “Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings”, *Research in the Teaching of English* (1981), pp. 75-86.

The document design process critique is yet a further method involving expert-judgement. The procedure focuses on identifying predictors of poor writing quality and is designed to help show up weaknesses in the ways in which text is created. The idea is to try to predict (and prevent) poor writing before it occurs. Evaluators examine the approach to planning, generating, revising and evaluating text. They consider the way people collaborate, the guidelines that writers follow and the kinds of feedback that is involved in shaping a document. The aim is to identify the strengths and weakness in the drafting process along with recommending education or research that will help to remedy the weaknesses.

(3) Reader-focused evaluation

Reader-focused methods are procedures that rely on feedback from audiences. There are two classes of testing that involve feedback from audiences: concurrent tests and retrospective tests.

(a) Concurrent testing

Concurrent tests evaluate problem-solving behaviours of readers while they are actively engaged in comprehending and using the text for its intended purpose. Concurrent testing methods include:

- cloze testing,
- behaviour protocols (which are sometimes called motor protocols),
- performance testing, and
- thinking-aloud verbal protocols.

The basic essentials for reader-focused evaluation are that:

- those tested are members of the target audiences and perform real tasks; and
- each test is conducted in the same way.

Cloze testing

The cloze test uses text that has had words systematically deleted. Readers are asked to try to fill in the gaps. The theory is that a good quality text should have a high degree of predictability and readers should be able to fill in the gaps if the text is of good quality. Cloze testing takes readers into account and in fact filling in the gaps seems to involve many aspects of the reading process, including word recognition, knowledge of syntax and semantics. The method seems to be best suited to narrative and expository texts and unsuited to reference and procedural texts, which means that it has limited value for testing the comprehensibility of legislation. A major drawback with this method is that it fails to provide any feedback about how text is functioning at the visual level.

Behaviour protocols

This method involves recording readers' actions and behaviours during the reading process. The main feature of behaviour protocols is that participants do not talk aloud while performing a task: they

simply perform the task while their actions are recorded, either by a human evaluator or a computer program, or both. For example, an evaluator may observe where readers look for information in a lengthy document, such as an index, table of contents or list of definitions. Behaviour protocols include *keystroke logs*, *eye movement studies* and *user edits*. Keystroke logs provide detailed information about users' errors and error recovery patterns and can be used to develop models of users' behaviour. Eye movement protocols have been used to find out how people read scientific texts involving prose and diagrams²⁰. Another kind of behaviour protocol, the user edit, involves observing readers directly while they work and interact with a machine, using only its operations manual as a guide. The observer closely watches how readers use the text, when they use it and how the text helps or hinders understanding.

Performance testing

In performance testing, evaluators monitor factors such as readers' task performance, retrieval and access behaviours, error recovery strategies, cognitive load, and the general ability to use the text. Evaluators using performance testing are mostly concerned with obtaining benchmark information about speed and accuracy in coming to grips with the text. Performance testing has played a major role in text evaluation and is likely to continue to do so in the future. Talking or thinking aloud is not encouraged because it adds time to performing the task. However, it is often hazardous to infer problem-solving strategies without more explicit indicators of thinking, so evaluators therefore often supplement performance testing with think-aloud protocols.

Thinking aloud protocols

With the think-aloud protocol, an observer tests each participant individually and records the participant's responses, comment and behaviour. The participant is given one or more tasks to perform using the draft, such as finding information in it or solving a potential problem that the draft is designed to address.

Throughout the testing, participants are asked to think aloud. This provides information not only on what they do, but also why they do it, thus revealing the thought processes leading to their actions, the terms they find difficult to understand and the directions they find inadequate or confusing. As well as recording responses, comments and behaviour, the observer, in order to identify those problems that participants may find difficult to articulate, prompts participants to speak whenever they hesitate and asks them about any difficulties they encounter.

Whenever possible, testing involving this method should be conducted in participants' own surroundings, thus assisting legislative counsel to identify the steps that participants have to take to carry out the task and any constraints.

²⁰ See Hegarty, M., Carpenter, P.A., and Just, M., "Diagrams in the comprehension of scientific text", in Handbook of Reading Research, Vol. II, R. Bar, M. Kamil, P. Mosenthal, and P.D. Pearson (ed), New York: Longman.

(b) Retrospective methods

Retrospective methods are the most commonly used of the reader-focused methods. They are designed to elicit feedback after readers have finished with reading and using the text. Retrospective testing methods include:

- comprehension testing,
- surveys,
- structured interviews,
- focus groups,
- critical incidents, and
- reader feedback cards.

The main disadvantage of retrospective testing is that it does not always reveal specific text features that need to be revised.

Comprehension testing

Comprehension testing has for some time been a widely used retrospective measure in evaluating text quality. This form of testing usually asks readers to paraphrase, recall, summarise, recognise or draw inferences about particular text items or textual features by having them engage in activities such as answering true or false or multiple choice questions or completing blank spaces. The value of comprehension methods lies in the quality of the test. Poorly constructed questions are likely to produce trivial results.

In assessing participants' performance on comprehension tests, evaluators typically use either criterion-referenced or norm-referenced tests. In criterion-referenced tests, the performance of all participants is compared to a pre-established criterion for success²¹. Norm-referenced testing, on the other hand, compares the performance of participants with each other, so that the relative quality of the text is judged by considering readers' performance in comparison to each other.

Surveys

With surveys, participants typically reply to a mixture of open-ended and close-ended questions that are designed to elicit opinions about the use of visual and verbal text features. The advantages of surveys are that

- they are relatively inexpensive,
- they do not require much time, and
- participants can remain anonymous.

However, a major disadvantage is that often the participants are self-selected, thus producing biased

²¹ e.g. that readers should be able to perform all the tests with 85% accuracy.

results. Moreover, if the participants rate the surveyed document poorly, evaluators have to carry out further tests to ascertain which text features caused the problems for the participants. The response rate may be low and participants tend to ignore open-ended questions.

The structured interview

With a structured interview, each participant is asked the same questions about the draft. The method is particularly valuable for ascertaining whether participants understand the language used in the draft. By asking participants to define terms or to explain a phrase in their own words, legislative counsel can determine whether users are interpreting the draft correctly and whether important information is being overlooked.

The structured interview seems to work well in conjunction with the think-aloud protocol and thus ensure that all potential problems have been identified²². If, for example, participants were able to carry out the task correctly, the think-aloud protocol might fail to show that they had not understood one of the terms. The structured interview on the other hand may well reveal this lack of understanding.

Interviews are an extremely valuable way of finding out how a text is working. This is because participants tend to feel more comfortable answering interview questions than objective test items. Disadvantages are that interviews are time consuming to conduct and the data are often difficult to analyse, so that it may be hard to generalise from them.

Focus groups method

With focus groups, open-ended interviews are used to ascertain people's attitudes, perceptions and opinions about a particular text or group of texts. A focus group for testing a draft Bill could comprise a group of lawyers who are interested in the subject-matter of the Bill or a group of lay-people who are similarly interested.

The following are the advantages of focus groups:

- the method is a socially oriented research one that captures real-life data in a social environment;
- the method has flexibility;
- the method provides speedy results;
- the method is cheap to operate.

However, the use of this method has a number of limitations that affect the quality of the results. Limitations include:

²² See Dumas, J. and Redish, J. *A Practical Guide to Usability Testing* Norwood, NJ: Ablex Publishing Corp., 1993.

- the method gives the evaluator less control than the structured interview,
- data are difficult to analyse,
- moderators need special skills,
- differences between groups can be troublesome,
- groups can be difficult to assemble, and
- discussion has to be held in a conducive environment.

Critical incidents method

This method involves asking participants to recall salient aspects of their interaction with a text. It is designed to elicit readers' memories of positive or negative experiences associated with reading or using the text. Sometimes, participants are given a scenario and asked to complete the "story" discussing how and when they might use the text.

A major disadvantage of the method is that it imposes a heavy burden on memory and may predispose participants to exaggerate, so that the resulting data may not be very accurate.

(4) A hybrid method

Prediction measures draw on a document's text features to predict readers' abilities to comprehend them. Two kinds of prediction measures are currently available. One is the readability formula, discussed under "text-focused methods". The other is what have been called "quality metrics". The steps involved in quality metrics are as follows:

- A number of expert document designers are asked to list the features that they regard as essential in reviewing the usability of the relevant document. (For example, "Is information easy to find?", "Does the document have a useful table of contents?", "Do headings comprise verb phrases relating to tasks that users would want to do?")
- Based on this information, the evaluators develop a list of factors that they hypothesise would make a difference to users. This list may include a number of factors (say 4) with several features for each one.
- Several versions of a sample document (say 5) are then created which systematically vary the treatment (good or poor) of the factors and features.
- A number of people who are likely to be users of the final document (say 50) are selected to participate in the study. Each participant (one at a time) does the same set of tasks. Each participant has only one version of the document, so that if there are 5 versions of the document and 50 participants, each version is used by 10 of the 50 participants.
- In analysing the results, statistical techniques such as regression analysis are used. (Regression analysis is a statistical technique for ascertaining the extent to which each of the different features contributes to the final result.)
- Based on the results, a multi-item questionnaire is prepared that operates like a checklist. A group of reviewers takes the draft of a document and answers the items in the

questionnaire. Some items are binary: e.g. “In the table of contents, the headings are verb phrases that match tasks that users would want to do (practically all of the time [90%-100%]; most of the time [65-89%]; sometimes [20-64%]; or seldom [0-19%]”.

- The questionnaire can be tested and refined with further documents so that evaluators can be sure that it is usable and that it will provide results that matched the review that experts gave of the same documents.

The practical goal of “quality metrics” is to predict the effectiveness of documents without the need to use costly criterion-reference measures (such as thinking-aloud protocols) for each document.

However, it should be emphasised that “quality metrics” does not replace reader-focused usability testing. Rather the method is designed to help writers (such as legislative counsel) to prepare better documents for usability testing, documents that already incorporate the features that research shows makes a difference to users. The only true measure of usability (i.e. of quality in a document) is whether real users can use it for real situations.

The method has a number of advantages over such measures as readability formulae. For one thing, it takes into account a broader range of text features, many of which must be assessed through human judgement rather than through simple counts. In addition, unlike readability formulae that take into account only a very narrow range of verbal features, quality metrics have the advantage of including a wider range of “beyond the sentence” verbal features as well as some visual features. Furthermore, the method provides information about the strengths and weaknesses of the document by giving a profile of different dimensions of the text (such as whether the layout helps readers to find information quickly or whether the text has a high number of noun strings).

In developing this method, tricky problems could arise as to the choice of documents on which to construct a model of comprehensibility or usability, and the choice of participants in the usability phase of the design on which to construct a model of readers’ performance. A potential threat to the adequacy of this method lies in the extent to which it is generalisable.

(5) Summary

Advantages of text-focused methods are that:

- they are inexpensive to use,
- some can be automated, and
- they can be helpful in detecting obvious classes of error.

However, the advantages of those methods are outweighed by their inherent weaknesses. Those weaknesses are that:

- the predominant focus of those methods is on the word and sentence level features of the text;
- their output typically provides little, if any, useful information about how the document is working at the macro level; and

- they provide no information about readers' needs.

Research²³ shows that, when text-focused methods are used as the only guide for revision, the revised text may actually become worse.

Although expert-focused evaluations are useful and can provide a wealth of information for the counsel, they often suffer from the evaluators' being too close to the text or product the text describes. Because the only readers who participate in evaluating a draft legislative document may be other legislative counsel, the resulting text may work well for lawyers and those who are experts in the relevant field but may fail for the average reader. External reviews may be quite helpful in supplementing in-house evaluation methods. However, expert judgement focused methods should not be used alone: they should be supplemented by other evaluation procedures, in particular those that are reader-focused.

Retrospective testing can provide very useful data for revising text. However, most researchers²⁴ agree that concurrent methods provide the most reliable data. Retrospective methods should, therefore, be used in conjunction with concurrent methods so that greater reliability is achieved.

What is the best approach for legislative counsel?

The research literature strongly suggests that reader-focused methods yield the best results. By eliciting information about a draft legislative document from representatives of the various audiences who have read the document, the legislative counsel concerned can find out whether or not it is truly "usable". Having found out what problems those representatives have, the counsel can address the problems before the document takes effect. As Martin Cutts has pointed out, "revised documents are never better or clearer until users' performance proves it". However, there are two disadvantages of reader-focused testing: it is expensive and it is time consuming. Nevertheless, there are at least three reasons why reader-focused testing might be seen as being cost-effective.

These are as follows:

- (1) Testing a draft legislative document can help the drafter to work efficiently –
 - by identifying the real problems of the proposed legislation, rather than the perceived problems,
 - by helping to determine the extent of each problem, and
 - by showing the legislative counsel possible ways of solving the problems.
- (2) Testing can ensure that the document will meet the needs of intended audiences before it is

²³ Swaney, J.H., Janik, C., Bond, S.J., and Hayes, J.R., "Editing for comprehension: Improving the process through reading protocols". Document Design Project Report No. 14, Pittsburgh: Carnegie-Mellon University, Communications Design Centre, 1981.

²⁴ Schriver, K.A. "Evaluating text quality: The continuum from text-focused to reader focused methods", IEEE Transactions on Professional Communication, vol. 32, no.4, 238-255.

enacted or promulgated.

- (3) Testing can provide measurable proof that the enacted or promulgated document works, thus saving subsequent amendment.

Legislative counsel should not wait until a document is completed before submitting it to testing²⁵. Dumas and Redish²⁶ claim that usability testing should be undertaken “early and often” not just at the end when it is often too late to consider making changes. What they say applies to draft legislation just as much as it applies to other kinds of documents.

Drafting legislation should be an iterative process in which the various parts are drafted, tested, revised and retested. By including testing early on in the drafting process, potential problems are discovered at an early enough stage to ensure that there is time to rectify them. It is important that the revised draft should be retested, because solving one problem can often create another one. And of course it is essential that all difficulties are identified and addressed if the draft is to end up as fully workable document.

In the early drafting stages, legislative counsel can ask any typical user (colleagues possibly) to test the draft text and provide feedback. However, it is important that the people selected for the test represent actual users, that they perform real tasks and that each test is conducted in the same way. So far very little draft legislation has been subjected to reader-focused testing. One piece of legislation that has is the Canadian *Fireworks Regulations*. An outline of how the testing was carried out and the result of the testing is set out in Appendix A below. Unfortunately, the testing of the Regulations was limited. But, although it would have been better if successive drafts of the Regulations had been tested iteratively and if testing procedures had been uniform in the different locations where testing was carried out, some testing is better than no testing at all. I understand that the current revisions of the Australian Income Tax Assessment Act and the Corporations Law are being subjected to reader-focused usability testing but I have not been able to obtain details of this as yet.

²⁵ See Dumas and Redish, *ibid.*

²⁶ See note 23.

APPENDIX A

TESTING THE CANADIAN CONSUMER FIREWORKS REGULATIONS

Why the testing was undertaken

The need for usability testing of the Regulations was based on the assumption that the vast majority of people in a sector to be regulated will comply with the law if they can read and understand it easily. Thus, non-compliance with a law can and no doubt often does arise because the people affected do not know and understand the law.

The starting point was to prepare draft Regulations using a plain language drafting approach and then to examine the effectiveness of the approach. The key features of the approach were –

- to identify the intended users;
- to know what information needed to be communicated;
- to choose words that intended readers could understand;
- to present the text clearly; and
- to test to find out if the legislative purpose was being achieved.

How the testing was carried out

At the outset, the drafting team consulted with representatives of the various stakeholders in the Canadian fireworks industry (e.g. importers, distributors, retailers, and explosives inspectors etc.) with a view to identifying compliance problems and communication problems. As a result of the consultations, the team—

- gained an understanding of the fireworks distribution process and the primary players in the industry;
- discovered that there was little data on fireworks accidents in Canada;
- discovered that, although importers and distributors understood the existing Regulations reasonably well, retailers and others further down the distribution chain did not;
- discovered that interpretation of the existing Regulations varied significantly;
- came to realise the need to conduct education campaigns about the Regulations directed at some of the stakeholders; and
- gained an appreciation of the difficulties of enforcing the legislation because of the small number of inspectors and regional variations arising from differing requirements of the various enforcement authorities.

In preparing the draft Regulations, the drafting team initially relied on academic texts on plain language and consultations with proponents of plain language and experts on usability testing. The plain language draft of the draft Regulations contained the following features to help users find

information:

- a table of contents (but not forming part of the Regulations);
- marginal notes;
- additional headings;
- bolding of defined terms;
- the use of graphics and icons.

It also used or contained the following features that affected the structure and content of the text:

- elimination of unnecessary cross references;
- examples showing how the legislation was intended to operate;
- clear and concise expression and no legalese or unduly technical language;
- re-organization of the text to avoid unduly long or dense blocks of text;
- the use of “must” rather than “shall”;
- elimination of the need for two titles;
- Schedules designed as ready-made instructions to facilitate use and ease of compliance.

Annexed to the draft Regulations were a background information note and summary designed –

- to accompany the Regulations when published;
- to establish where the Regulations fitted into the context of the Explosives Act and other subordinate legislation under that Act;
- to specify key definitions in that Act that could not be repeated in the Regulations; and
- to provide an address to enable interested people to get more information.

After the final draft of the Regulations was finished, interviewers tested the draft. It should be emphasised that the testing was designed not to find out whether users “liked” the Regulations, but to find out whether they were easier to use than the existing Regulations. The testing was carried out with four user groups (consumers, retailers, distributors and officials - including explosives inspectors, police officers and firefighters). Each group was tested on the Part of the Regulations that related to the group. Consumers were asked to explain in their own words what they thought each instruction meant and to match pictograms to each instruction. Retailers were asked questions based on Schedule 1 to the Regulations (which was relevant to them).

Distributors and officials received a copy of the whole draft and were asked questions on the draft based on situations relevant to their work. Each of those two groups also took part in group discussions, which were video taped. Members of the groups discussed how easy it was to find information, the wording of the draft Regulation and their opinions as to how difficult it was to follow the revised Regulations.

So what did the testing show? It showed that potential consumers found the instructions applicable to them were clear and fairly easy to understand. They understood the pictograms and matched them

fairly well, but had some difficulty with the way the instructions were ordered. Some words caused them difficulties.

Retailers did not understand the purpose of Schedule I (Safe display and storage instructions) and were unable to differentiate satisfactorily between display instructions and storage instructions. They also found it difficult to understand some of the terms used, such as “storage unit”, and in understanding the storage requirements.

Distributors and officials found the wording of the draft Regulations reasonably clear, but found they had to switch around the document to find the answers to the questions put to them. They also found the Schedules, table of contents and definitions very useful. However, they did not notice the background information note and had some difficulty with the storage information, not being able to find the correct sections. Issues that particularly concerned the distributors were the possibility of liability for something that a retailer failed to do and that the strictness of the requirements of the Regulations might result in more “trailer” sales. The major concerns of officials were that they did not have a power to search premises for illegal fireworks and to ensure that the Regulations were being complied with. They also thought that the lack of precision might make the Regulations difficult to enforce.

What the testing showed

The testing process showed that generally speaking interviewees found the draft Regulations reasonably clear. They liked the Schedules, table of contents and definitions. However, they preferred the background information to be located near the beginning of the document. They also thought that the organisation of the document could be improved.

How the document was modified as a result of the testing

As a result of the testing, the following changes were made to the draft Regulations:

- the second person “you” was replaced by the more familiar third person;
- the summary and background information were combined into a single note headed “Important Information” placed immediately before the Regulations so that it was not overlooked; and
- the text was reorganized so that the storage of fireworks was clearly distinguished from their sale. In other words, provisions relating to those subjects were grouped together for ease of reference and clarity, even though doing this made the document, more repetitious.

Evaluation of the testing process

The initial consultations with stakeholders in the fireworks industry provided a lot of information. Such preliminary consultations seem to be an essential ingredient of any process designed to ensure optimum usability of proposed legislation. The inclusion of the drafters at this stage of the process

provided them with an opportunity to familiarise themselves with the realities that had to be addressed by the legislation.

Apart from giving them more information, the drafters were able to put more pertinent questions about the matters with which the legislation was concerned, thus contributing to a better end product.

The proponents of the testing process claimed that the revision of the draft Regulations, using the plain language process, resulted in the final Regulations being easier to use and more realistic. They also claimed that the testing process went very well and was an important step in successfully revising the Regulations. One problem however was getting groups together for group sessions. Another problem was that the testing process took rather longer than was originally anticipated. It seems that more time should have been spent on preparing a test strategy, developing test questions, pretesting the questions, contacting participants in the testing process, conducting the testing and writing up the results. Nevertheless, the team considered that the research component of the project was most cost-effective and the testing should become a standard part of the drafting process.
