Readability Factors and Assessment Tasks

David Wray (University of Warwick, United Kingdom)
and
Dahlia Janan (Sultan Idris University of Education, Malaysia)

Corresponding author:
David Wray
Institute of Education
University of Warwick
Coventry
CV4 7AL
United Kingdom
Email: d.j.wray@warwick.ac.uk

28th September, 2012

Readability Factors and Assessment Tasks

Abstract

Readability is the study of matching a reader to a text. The factors influencing readability, both text factors and reader factors, have been widely researched from the standpoint of attempts to maximise reader understanding of texts. The application of understandings in the area has not always been systematically applied to the design and writing of assessment tasks, however.

This paper is an attempt to provide a wide ranging review of literature which bears on the task of the assessment designer in ensuring that assessment items measure what they are supposed to measure, and not just the reading abilities of the test takers.

Introduction

Crisp (2011) has argued that, whilst part of the difficulty of an assessment task will, of course, be due to the intrinsic demands of the subject content of that task, the actual difficulty can be affected, sometimes in unexpected or unfair ways, by features of the way that questions are asked (Pollitt et al., 1985; Fisher-Hoch, et al, 1997). Ahmed & Pollitt (2007) argue that, "Putting questions into context inevitably involves using extra words to ask the question. If pupils have to read more text in order to answer a question then their reading ability is being tested as well as their understanding of concepts" (p. 203).

Research into the effects on test takers of the contextual variables of assessment questions has a substantial history. The Assessment of Performance Unit (1985), for example, claimed that context, that is the material surrounding a mathematics assessment task such as accompanying pictures and/or the embedding of the task in a real-life situation, could affect success rate on that task from a few percentage points up to 20%. Nickson & Green (1996) later found that the degree of context in which a mathematical question was set could affect pupils' choice of the correct mathematical operator with which to answer the question.

Schagen & Sainsbury (1996) have confirmed that reading ability can make a significant contribution to pupils' scores on Mathematics assessments, and the same conclusion can be drawn from the study by Shorrocks-Taylor et al (2003) which found that the substitution of what the authors refer to as Contextual Number questions (problem solving) by number-focused data-handling questions improved the assessment scores of a number of the children taking this assessment. The suggestion is that the embedding of number questions within a heavily language reliant context had made certain questions more difficult for certain (but not all) children to answer successfully.

A similar picture emerges from research into the effects of language on learning, and hence assessment of learning, in science. Fang (2006), for example, has investigated the linguistic demands of school science texts and concluded that these can make a significant difference to pupil understanding of these texts. If this is the case, then it is likely that this language may remain a barrier to pupils performing their best in assessments of their science knowledge and

understanding, with this assessment often, necessarily, being carried out through the medium of language.

The accessibility, therefore, of the language through which assessments are made is a crucially important consideration for the designers of assessment instruments. The language used needs to be *readable* in the broadest sense, and the principles at work here are those underpinning the concept of *readability*. It appears that a number of factors can influence the readability of any text, and designers of texts through which assessments are made need to be alert to the influence of these factors.

The nature of readability

Readability is the study of matching a reader and a text (Gilliland, 1975). Arguably the most important pedagogic decision that teachers make is "making the match" (Fry 1977), that is, ensuring that learners are supplied with reading materials of an appropriate level of difficulty. Learners given reading materials that are too easy are not challenged and their learning growth can be stunted (Chall & Conard, 1991). Learners given reading materials that are too difficult can fail to make progress (Gambrell, et al, 1981), are frequently off task and may exhibit behaviour problems (Anderson, Wilkinson & Mason, 1987). Making the match is therefore a crucial skill for teachers, and its successful exercise requires knowledge of the readability level of materials.

Similarly, without understanding the readability of assessment questions, the test developer risks producing items that do not correctly match to the reading abilities of the learners for whom the assessment is planned. If the readability level of a test item is higher than the reading ability of the test takers, then it is likely that the item is not assessing the construct of interest (the subject matter) but rather the test taker's reading ability.

Explorations of readability gave rise to a significant body of research from the 1920s to the early 1990s, one of the major outcomes of which was the production of "readability formulae", that is, analyses of texts designed to give a quantitative measure of the "level" a reader would need to be at in order to read and understand them. Various definitions of the concept of readability have emphasised elements in a text associated with comprehension (or lack of it) on the part of the reader. Parts of the concept also referred to a person's ability to read a given text at an optimum speed. Finally, the concept also included motivational factors which affected a reader's interest in reading a text. According to Dale & Chall (1948) these three elements of the definition of readability were not separate, but interacted with each other. Thus, definitions of readability have never been entirely text-centric. However, despite the claim of Harris & Hodges (1995, p. 203) that, "Text and reader variables interact in determining the readability of any piece of material for any individual reader", approaches to the measurement of readability have usually involved objective estimates of the difficulty level of reading material derived from the application of formulae which generally took into account sentence and vocabulary difficulty.

Most studies of readability were carried out within a positivist paradigm (Janan et al, 2010) which saw text difficulty as determined by factors within the text itself, and reading as a matter of getting meaning from the page. However, views about the nature of the reading process have

changed over the past 20 years towards a more interpretive definition which emphasises that making meaning through reading comes from a process in which the readers interact with texts. This new paradigm of reading has meant that research into readability has also changed (Janan et al, 2010). In this review, we will explore these two dimensions of readability by focusing firstly upon factors within the text itself, and secondly upon characteristics of readers.

Readability: looking at text features

The effects upon reader understanding of number of text features have been well researched.

a) Word difficulty

Word difficulty has to do with the reader's/test taker's understanding of individual words. It has traditionally been measured by word length, with the assumption that longer words are harder to read than short ones. It is often suggested that short words are perceived as more familiar and long words as more formal or technical and there is research that suggests that readers pause longer on longer words (Just, et. al, 1982).

Nevertheless, there have also been findings questioning the assumption that short words are always easier than long ones. There are, for instance, examples of monosyllabic words (e.g. *adze, gneiss)* found in lower secondary school text-books which are unlikely to be easy words for the pupils who read such books (Perera, 1980).

Neither is it always the case that longer words are harder to read. There are very few seven to eleven year olds, for example, who will not be able to read and understand words such as *tyrannosaurus* and *diplodocus*. Such examples suggest that the length of a word is not the crucial feature in whether it can be read easily or not. Children's motivation to read a word and their existing familiarity with it are much more significant indicators of reading ease.

b) Word familiarity

Word difficulty is affected by word familiarity. In previous readability research word familiarity has referred to those words that appear in word lists such as the Dale-Chall (1948) list (revised in 1995 – see Chall & Dale, 1995). It is presumed that words which appear on this list will be relatively easy for children to read and that words which do not appear will be unfamiliar and more difficult to read.

It is certainly the case that, from analyses of English word usage, a fairly small number of words make up a substantial proportion of words in common use. Nation & Waring (1997), for example, using data derived from the Brown University corpus of present-day English (Francis & Kucera, 1979) show that over 70% of English text is composed of just 1000 words. One implication of this may be that, if the text in written assessment tasks was limited to the first 1000 of these words, this would maximise the readability of these tasks.

Unfortunately, there are some questions about the validity of the means used to determine lists of familiar words such as this. Perera (1980) noted that many such lists, especially those used in

readability formulae, were based on frequency counts done in the USA, although the formulae were still used in Britain, where patterns of vocabulary use were different. A comparison of the revised Spache (1974) list (American) with a British frequency count of children's written vocabulary (Edwards & Gibbon 1973) reveals some discrepancies. Words such as *bonfire*, *doll*, *fairy* and *mummy* are listed as familiar words in the British list but not the American, whereas words like *cabin*, *candy*, *parade* and *neighborhood* are listed as familiar words to American children but not to British. It has also been suggested that, 'average word frequency is not a good predictor because many words are common at certain ages, but then become uncommon – such as "kitten". But … infrequency at higher grade levels does not make them difficult words' (Milone, 2008, 6).

c) Sentence difficulty

The common belief regarding sentence difficulty is that the longer its sentences, the harder a text is to read. Hence, the average sentence length of a text has often been used to measure its difficulty. Most readability formulae have calculated this by dividing the number of words by the number of sentences in a text.

Care needs to be taken, however, in using sentence length as an absolute measure of reading difficulty. Short sentences may well convey conceptually difficult ideas. Also Perera (1980) argues that, at times, longer sentences are easier because they provide more clues to meaning and to the relationship between sentence elements.

Nevertheless, there is evidence that sentence complexity can make a difference to readers' comprehension of a text. In a classic study, Reid (1972) took sentences from a range of reading material produced for 7 to 8 year olds which she judged to be ambiguous in their syntactic structure. These sentences were then rewritten to make them less ambiguous and the two versions shown to 7 year old children who were then asked questions about the sentences. Reid was able to double levels of understanding by modifying the sentences. Reid used her findings to advocate that greater consideration needed to be given to the linguistic structures used in early reading material, but also to suggest that children were disadvantaged in reading texts unless they had had a great deal of prior experience of texts with similar structures.

Thompson & Shapiro (2007) have identified four variables that contribute to sentence complexity: the number of propositions within a sentence, the number of embedded clauses, the order in which major elements appear, from simple, active sentences such as subject-verb-object (SVO) to passive sentences (OVS), and the distance between crucial elements in the sentence.

d) Cohesion and coherence

Language features operating at the level of the word or the sentence may lead to accessibility issues, but one of the key features of a text is that it is not just a group of words and sentences. Instead, there is a structure in a text which glues the various text components together. In reading, the reader needs to construct a coherent, mental representation of the ideas which have been cohesively presented in the text. Louwerse & Graesser (2004) use the term "coherence" for the way ideas 'hang together' in a text and "cohesion" for the textual links through which

coherent ideas are built up. The effects on readability of the cohesion and coherence of the texts used in assessment questions are often not explicitly considered by test designers.

Connor (1996) defines cohesion as "the use of explicit linguistic devices to signal relations between sentences and parts of texts." These cohesive devices are phrases or words that help the reader associate items or statements in a text with others elsewhere in that text, or outside. Halliday & Hasan (1976) originally identified four general categories of cohesive devices in texts: Reference, Substitution and ellipsis, Lexical, Conjunction.

Studies of cohesion in reading show that it can make a substantial contribution to readability. Chapman (1987) demonstrated that readers between the ages of eight and fifteen showed growth in their ability to perceive cohesion in text and to use it to support their comprehension. This suggests that readers develop an awareness of cohesion over time and make increasing use of it to get meaning from print. However, if they lack sufficient experience and knowledge of the ways in which texts are cohesive and coherent, this can be a major hindrance to their comprehension. Other studies (e.g. Fulcher, 1989) have suggested that readers' failure to comprehend a text can result from their inability to follow the flow of cohesive ties within the text. A more complex picture, though, is provided by the research of Ozuru et al (2009). They compared the reading of science texts which were deliberately written to have either high or low cohesion between sentences. They found that the effect of text cohesion depended both on the reading skill and the level of prior knowledge of the reader. Higher text cohesion seemed to benefit readers with poorer levels of prior knowledge. However, readers with lower levels of reading skill but higher levels of prior knowledge of the topic of a text tended to process the text more shallowly and actually perform less well on a subsequent assessment of their understanding. This finding replicates that of O'Reilly & McNamara (2007) and suggests that readers' difficulty in learning new concepts can be alleviated to some extent by making text more cohesive which makes readers less dependent on pre-existing knowledge. Yet, it seems that readers are not able to take advantage of increased cohesion unless they have sufficient reading skill. It seems important for teachers not only to work on improving learners' understanding of content, but also on their abilities to read to learn from texts (and their abilities to read effectively the ways in which assessment questions are typically written). A further implication relates to the need to improve the texts learners are asked to read, for both learning and assessment purposes (Beck et al., 1991; Graesser et al., 2003). Such texts need to be evaluated for their levels of cohesion but it should not be taken for granted that increasing the levels of cohesion, for example by spelling out all the cohesive links within a text, will benefit readers in the way it might be thought.

e) Content structure and complexity

Well-written text requires, in addition to coherence and cohesion, a structure that readers can easily use to find the information they need and then to understand it correctly. Text can become confusing when information is inappropriately presented. When we read text, we build a collection of the concepts described therein, deducing these concepts from the words and phrases used within the text. We build certain interpretations out of these blocks of words which are not randomly organised, but obey quite strict rules of association.

When linguistic expressions combine into units for processing, many of the individual linguistic elements are ignored and the whole chunk is treated as one semantic unit. When a significant amount of information is conveyed in a relatively small amount of text, the reader can easily become confused. This problem is known as 'Propositional Density' (Kintsch, 1974). The greater the number of ideas expressed in a text, the more work is required of the reader to interpret the text correctly (Newbold & Gillam, 2010).

Kintsch & Keenan (1973) presented readers with sentences of constant length but varying propositional density. They found that, as the number of propositions in a text increased, so did both the time taken by readers to read the text, and the number of propositions they were able to recall from the text. This suggests that the unit of meaning that readers deal with in reading is the proposition. There is now quite robust evidence that high propositional density in a text adversely affects readers' understanding of that text (e.g. Barshi & Healy, 2002; Sonnleitner, 2008). The implication of this for assessment designers links with the earlier recommendations about sentence complexity in assessment questions. The more complex, and more propositionally dense the text of a question, the harder will that question be to answer, no matter what the test taker's actual content knowledge.

f) Legibility and print issues

Studies of legibility have researched factors such as character size, thickness of strokes, white space between strokes, dissimilarity of characters, leading, line length, quality of paper, colour of paper, and colour of ink (Waller, 1991, p. 342). Research has shown that legibility issues such as the size of font and typeface can affect reading and reading speed (Hughes & Wilkins, 2000 and Wilkins, et. al, 2009). Eyles, Skelly & Lou Schmuck (2003) found, that a san serif font generally improved readability, although it has sometimes been argued that serif fonts ease reading because the serifs draw the eye along the line.

Text legibility is also influenced by the size of the font (see for example, Feely, et al, 2005; Wilkins, et al, 2009). Studies have shown that by increasing the font size the percentage of fluent reading is also increased (Feely, et al, 2005) but small font sizes (below 12 point) are thought to make reading increasingly difficult, and are more stressful to the visual system (Wilkins, et al, 2009).

Readability: looking at the characteristics of readers

It is unlikely that two test-takers are going to perform exactly the same when faced with a test, especially one which involves the extensive interpretation of written language. Test-takers will each have individual characteristics which will affect, however slightly, their responses to assessment questions. As mentioned previously, an important omission in most research into readability has been the effect of various reader characteristics. Readability, we now recognise, is the products of the features in a text **and** the characteristics of a reader. It is, therefore, important for writers of assessment questions to take into account these reader characteristics if they are to work towards what Cole & Zieky (2001) have termed, "the new faces of fairness".

Understandings of the ways in which reader characteristics can affect the readability of assessment questions have been developed over a number of years through the use of Differential Item Functioning (DIF) analysis (Gierl, 2005; Zumbo, 2007). This is a well-established statistical procedure that has been used to identify individual questions in assessments that may be biased against particular groups of test-takers. Bias occurs when assessments produce different scores for members of different groups (e.g., groups with differences in racial, ethnic, language, cultural, gender, disability, or socio-economic status).

Although DIF analysis now has a substantial research history, it is still the case that we lack a full understanding of just WHY DIF occurs in educational assessments (Gierl et al, 2003). To develop such an understanding requires an appreciation of the kinds of test-taker characteristics which have been shown to affect the readability of the texts used in assessments. These characteristics will be examined in the following sections.

a) Physical capabilities

Having a disability or impairment can clearly influence a child's reading ability. Some examples of impairment that can affect reading include autism, dyslexia and ADD (Attention Deficit Disorder) and ADHD (Attention Deficit Hyperactivity Disorder). Children with ADD and ADHD have difficulties in concentrating on a task for any lengthy period of time (NINDS, 2011). Autistic children need special teaching techniques as they are often unable to interact with others. Dyslexia affects a child's reading ability in that it might be difficult for them to translate images to language and this may cause difficulty in spelling and reading (Just & Carpenter, 1987). There is also a range of physical capability issues which may affect readers, that is, readers who have specific learning difficulties, or hearing or visual impairments. Such issues are likely to have an even greater impact upon the accessibility of assessment texts for younger readers.

Abedi et al (2008) have confirmed that pupils with disabilities tend to perform in assessments at lower levels than those without disabilities. While their lower performance can be partly explained by their specific disability, there may be other factors that potentially interfere with this performance.

b) Reading abilities

Reading abilities enable the reader to read meaningful language, to read any written form with independence, comprehension and fluency, and to mentally interact with the message from the written form (Just & Carpenter, 1987; Downing & Leong, 1982). Hence, the reader needs to master skills such as word attack and comprehension. It is obvious that, if a test taker is handicapped by lack of reading ability, then he/she will be much less likely to succeed in any form of text which involves reading, whatever the level of content knowledge he/she may have.

Research by Morgan et al (2008) suggests that what has become known as the "Matthew effect" (Stanovich, 1986), that is, a pattern of increasing advantage or disadvantage in reading skill development following an initial advantage or disadvantage ("the rich get richer, the poor get poorer"), is very evident in test-taking situations. Poorer readers are less able to access

effectively the written language through which they are tested, and thus demonstrate lower abilities, causing expectations about their achievements, and perhaps also the level of material upon which they are tested, to be depressed even further.

c) Engagement/motivation

Engagement or motivation in reading refers to the intrinsic drive to read for the knowledge and the enjoyment that it provides (Guthrie & Cox, 2001). Engagement is important as it drives the reader to use their best strategies for understanding and interpreting the text (Guthrie, at. al., 1997). There are many examples in the literature and in common experience of readers who can read beyond their normal levels when they are engaged and motivated by particular texts. It also seems to be the case (Martin et al, 2007) that, while pupils in all countries have generally positive attitudes toward reading, those with the most positive attitudes tend to have the highest average achievement. In addition, motivational factors become more and more important as predictors of the ability to read for understanding as readers get older and develop their skills (Saarnio et al, 1990).

Research on test motivation suggests that this could be a crucial factor in obtaining high quality and accurate information from assessments in a range of subjects. One study found that test-taking motivation was positively related to subsequent performance on a cognitive ability test even after the effects of race and performance on the first test were controlled (Chan et al, 1997). Another study found that the validity of a particular test was much higher for a group with more positive motivation towards test-taking than for a group with less positive motivation (Schmit & Ryan, 1992).

Although, as discussed above, the underlying reasons for the differential functioning of some items in assessments are still speculative (Roussos & Stout, 1996), one of the most widely discussed explanations is test takers' interest in the content of assessments and/or their emotional reaction to this content. Stricker & Emmerich (1999) suggested that both of these explanations could account for the different levels of responses to assessment questions.

d) Prior knowledge

Prior knowledge is an integral part of the comprehending process (Johnston, 1984). Hence, prior knowledge influences what is understood from text. Not surprisingly, pupils who know more about a topic understand and remember content better than those who have a limited background in the domain (Chi, 1985) This factor also comes into play during test-taking. Ozuru et al (2009) found, for example, that, while understanding of a science text, as measured by performance on a set of assessment questions, was positively affected by reading skill, it was prior knowledge that was a much more significant predictor of success. This finding supports that of Bugel & Buunk (1996) who found that the differences often found between male and female success in assessments involving reading comprehension could largely be accounted for by differences in the prior knowledge that each gender tended to bring to the assessment.

One aspect of prior knowledge which has been extensively investigated is knowledge of the language of the assessment. Research conducted by Abedi and his colleagues has demonstrated

that there is a substantial link between pupils' English language proficiency and their performance on assessments (in English) in mathematics, science, and social studies (e.g., Abedi, et al, 2003; Bailey, 2000). Furthermore, several studies have found that assessments that are more linguistically complex produce larger performance gaps between learners of English as an additional language (EAL) and native English speakers (e.g., Abedi, et al, 2003; Abedi, et al, 2000). These findings suggest that assessments in all subjects assess language skills as well as content knowledge and skills.

Butler & Stevens (1997) have suggested a number of possible responses to the problems caused by the language of assessments to EAL learners. These range from modifications of the assessment for these particular learners (e.g. carrying out assessments in learners' native languages, or modification of the language used in test directions), to modifications in assessment procedures for this group (including, for example, extra assessment time or oral directions given in the native language). A meta-analysis (Kieffer et al, 2009) of studies of the effects of several of these 'accommodations' has, however, proved disappointing in finding little evidence that the assessment performance of EAL learners was much improved by them. What seems more important is to provide EAL learners with "targeted, explicit, and intensive instruction in the complex and specialized language that lies at the heart of each content area" (Kieffer et al, 2009, 1190).

Prior knowledge also includes the social and cultural backgrounds of test takers. The schema theory of reading comprehension proposes that the organisation of prior knowledge in a learner's mind provides a framework which enables understanding of the setting, mood, characters, and chain of events in a text. Readers acquire meaning from a text by analysing the words and sentences against the backdrop of their own personal knowledge of the world. Readers who share the knowledge background of the writer of a text 'come equipped' with the appropriate schemas for making sense of this text. The absence of an appropriate schema might be expected to lead to misunderstandings, which could be very significant in a test situation.

e) Gender

Gender differences in test responses have been commonly found in assessment research. Hamilton (1998), for example, found that, while male pupils were advantaged by the content of Science tests, particularly where they were required to bring to bear their existing, out-of-school knowledge, it was the format of the assessments which gave them the greatest advantage, with the use of diagrams being particularly salient. Gierl et al (2003) produced similar findings in mathematics tests, where males did much better than females on questions requiring spatial processing, rather than simple memorisation.

However, although there is documented evidence of gendered differences in reading achievement, as well as attitude, choice, and response for some boys (e.g., Millard, 1997), considerable evidence also suggests that this is not the case for all boys. Maccoby's (1990, p. 513) synthesis of decades of research on gender differences led her to claim that even when consistent differences between males and females were found, the amount of variance accounted for by gender was small, relative to the amount of variation within each gender. It has been repeatedly pointed out that boys are more different than alike, and that statistics lose sight of

individual differences. We need, therefore, to be very wary of assuming that all individuals fit the characteristics of the groups they belong to.

Accommodations, modifications and universal design

There is, therefore, a burgeoning range of reader characteristics which seem to have a significant effect on test takers' demonstrations of their capabilities in assessments. The traditional response of test development agencies has been to explore various assessment accommodations, including modifications of assessments and procedures for particular learners and groups.

Research has explored the effects of such accommodations but has rarely provided conclusive evidence (e.g. Stone et al, 2010). One example of this follows the review of literature by Rasinski (1990) that suggested that organizing text into smaller units could facilitate memory recall and improve comprehension for certain readers. Abedi et al (2010), however, found that doing this made no difference at all to the assessment scores of the pupils with disabilities that they studied.

Thompson et al (2004) have argued for a more global approach to the issue and a move towards universal design in assessments, that is, the design and development of assessments that allow the participation of the widest range of test takers, and produce valid outcomes reflecting the true capabilities of everyone who takes them.

Thompson et al (2004) outline seven key elements which underpin the concept of universally designed assessments.

1) ***Inclusive Assessment Population***. Assessments designed for national use must try to include every pupil. They need to be responsive to growing demands – increased diversity, increased inclusion of all types of pupils in the general curriculum, and increased emphasis and commitment to accountability for all pupils.

2) ***Precise*****ly Defined Concepts**. The specific constructs tested must be clearly defined so that all irrelevant barriers can be removed. An important function of well-designed assessments is that they actually measure what they are intended to measure. Test developers need to examine carefully what is to be tested and design items that offer the greatest opportunity for success within those constructs.

3) ***Accessible, Non-Biased Assessment questions***. Accessibility should be built into assessment questions from the beginning, and bias review procedures need to ensure quality in all items. Most importantly, items are developed by individuals who understand the varied characteristics of the pupils they are aimed at, and the characteristics of items that might create difficulties for any group of pupils.

4) ***Amenable to Acc*****ommodations**. The assessment design should facilitate the use of essential accommodations. Even though items on universally designed assessments will be accessible for most pupils, there will always be some who continue to need accommodations. For

example, the use of Braille as an accommodation will be facilitated if the following features are avoided in the design of the assessment:

- Use of irrelevant graphics or pictures
- Use of vertical or diagonal text
- Items that include distracting or purely decorative pictures, which draw attention away from the item content

These features are also relevant for pupils with visual disabilities who do not use Braille, and possibly also for the many for whom visual features may create distractions.

5) ***Simple, Clear, and Intuitive Instructions and Procedures***. All instructions and procedures should be simple, clear, and presented in understandable language. Assessment instructions should be easy to understand, regardless of a pupil's experience, knowledge, language skills, or current concentration level.

6) ***Maximum Readability*** a***nd Comprehensibility***. Plain language guidelines should be used to produce readable and comprehensible text. Plain language has been defined as language that is straightforward and concise. Several strategies that have been identified for editing text to produce plain language are listed below:
    - Reduce excessive length by reducing wordiness and removing irrelevant material.
    - Avoid unusual or low frequency words and replace these with common words (e.g., replace "utilize" with "use").
    - Avoid ambiguous words - for example, "crane" should be avoided because it could be a bird or a piece of heavy machinery.
    - Avoid words with particularly unusual or irregular spelling patterns, e.g. "trough", "feign".
    - Avoid proper names and replace with simple common names such as first names.
    - Avoid inconsistent naming and graphic conventions, by avoiding multiple names for the same concept and inconsistencies in the use of font.
    - Avoid unclear signals about where test takers' attention should be directed by using well-designed headings and other graphic features (bold, italic fonts) to convey information about the relative importance of information and order in which it should be considered.
    - Mark all questions clearly by the use of an obvious graphic signal (e.g., bullet, letter, number) to indicate separate questions.

7) ***Maximum Legibility***. Legibility is the physical appearance of text, the way that the shapes of letters and numbers enable people to read text easily. Bias results when assessments contain physical features that interfere with a pupil's focus on or understanding of the constructs that the questions are intended to assess.

Conclusion

As we argued earlier, the concept of readability has developed over the past twenty years, in line with theories about the reading process. Traditionally, studies of readability have focused largely on features in the text itself. We have reviewed in this paper the major conclusions which can be drawn from this research, i.e. that the readability of a text is influenced by issues such as word

and sentence difficulty, by cohesion and coherence, by conceptual difficulty, by legibility and print issues.

More recently, as more attention has been given to the role of the reader in reading, so attention in readability has focused more on the reader factors which may affect understanding. In this paper we have reviewed the influence of such factors as readers' physical capabilities, reading abilities, engagement/motivation, prior knowledge and gender.

A strong likelihood, of course, is that a modern concept of readability would need to take into account both these sets of factors and, indeed, a major interest for researchers is the way in which such factors might interact with each other. Such interactions are of importance as factors in the language accessibility of assessments. Test developers and designers need to understand the principles explored here if they are to produce "fair access by design" for all test takers.

References

Abedi, J., Kao, J., Leon, S., Mastergeorge, A., Sullivan, L., Herman, J. & Pope, R. (2010): Accessibility of Segmented Reading Comprehension Passages for Students with Disabilities, *Applied Measurement in Education*, 23:2, 168-186

Abedi, J., Leon, S., & Kao, J. C. (2008). *Examining differential item functioning in reading assessments for students with disabilities. (CRESST Report 744)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data (CSE Technical Report 603)*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurements: Issues and Practice*, 19: 3, 16–26

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: and experimental investigation of focus. *Assessment in Education: Principles, Policy and Practice* 14: 2, 201-32.

Anderson, R., Wilkinson, I. & Mason, M. (1987). *Do errors on classroom reading tasks slow growth in reading?* Center for the Study of Reading Technical Report No. 404. Champaign: University of Illinois at Urbana-Champaign

Assessment of Performance Unit (1985) *A Review of Monitoring in Mathematics 1978-82.* Slough: NFER

Bailey, A. (2000) Language Analysis of Standardized Achievement Tests: Considerations in the Assessment of English Language Learners. In Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S. & Mirocha, J. (Eds) *The validity of administering large-scale content*

assessments to English language learners: An investigation from three perspectives (CSE Report 663) 85–105. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Barshi, I., & Healy, A. (2002). The effects of mental representation on performance in a navigation task. *Memory and Cognition*, 30, 1189-1203

Beck, I., McKeown, M., Sinatra, G. & Loxterman, J. (1991). Revising social studies text from a text-processing perspective: evidence of improved comprehensibility. *Reading Research Quarterly*, 26, 251-276.

Bügel, K. & Buunk, B. (1996) Sex Differences in Foreign Language Text Comprehension: The Role of Interests Knowledge. *The Modern Language Journal*, 80: 1, 15-31

Butler, F. & Stevens, R. (1997) *Accommodation Strategies for English Language Learners on Large-Scale Assessments: Student Characteristics and Other Considerations. (CSE Technical Report 448)*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) University of California, Los Angeles

Chall, J. S. & Conard, S. (1991). *Should textbooks challenge students? The case for easier or harder books*. New York: Teachers College Press.

Chall, J. S. & Dale, E. (1995) *Readability revisited: the new Dale-Chall readability formula*. Cambridge, Mass.: Brookline Books.

Chan, D., Schmitt, N., DeShon, R., Clause, C. & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test taking motivation. *Journal of Applied Psychology*, 82, 300-310

Chapman, L.J. (1987) *Reading from 5-11 Years*. London: Open University Press.

Chi, M. T. H. (1985). Interactive roles of knowledge and strategies in the development of organized sorting and recall. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills* (Vol. 2, pp. 457-484). Hillsdale, NJ: Erlbaum.

Cole, N. & Zieky, M. (2001) The New Faces of Fairness. *Journal of Educational Measurement*, 38: 4, 369-382

Connor, U. (1996) *Contrastive Rhetoric: Cross-cultural aspects on second-language writing*. Cambridge: Cambridge University Press

Crisp, V. (2011): Exploring features that affect the difficulty and functioning of science exam questions for those with reading difficulties, *Irish Educational Studies*, 30:3, 323-343

Dale, E., & Chall, J.S. (1948) A formula for predicting readability. *Educational Research Bulletin*, 27, 37-54.

Downing, J. & Leong, C. (1982) *Psychology of Reading*. New York: Macmillan.

Edwards, R. & Gibbon, V. (1973) *Words Your Children Use: A Survey of the Words Used by Children in Infants' Schools with the Resultant Graded Vocabulary*. London: Burke

Eyles, P., Skelly, J., & Lou Schmuck, M. (2003). Evaluating patient choice of typeface style and font size for written health information in an outpatient setting. *Clinical Effectiveness in Nursing*, 7: 2, 94-98

Fang, Z. (2006): The Language Demands of Science Reading in Middle School, *International Journal of Science Education*, 28:5, 491-520

Feely, M., Rubin, G. S., Ekstrom, K., & Perera, S. (2005). Investigation into font characteristics for optimum reading fluency in readers with sight problems. *International Congress Series*, 1282, 530-533.

Fisher-Hoch, H., S. Hughes, and T. Bramley (1997). *What makes GCSE exam questions difficult? Outcomes of manipulating difficulty of GCSE questions*. Paper presented at the British Educational Research Association Annual Conference, September 11-14, University of York.

Francis, W. & Kucera, H. (1979). *Brown Corpus Manual*. Providence, Rhode Island: Brown University. Available at: http://icame.uib.no/brown/bcm.html

Fry, E. (1977) *Elementary reading instruction*. New York: McGraw-Hill.

Fulcher, G. (1989) Cohesion and coherence in theory and reading research. *Journal of Research in Reading*, 12: 2, 146-163.

Gambrell, L., Wilson, R. & Gantt, W. (1981) Classroom observations of task-attending behaviors of good and poor readers. *Journal of Educational Research*, 74: 400–404.

Gierl, M., Bisanz, J., Bisanz, G. & Boughton, K. (2003) Identifying Content and Cognitive Skills That Produce Gender Differences in Mathematics: A Demonstration of the Multidimensionality-Based DIF Analysis Paradigm. *Journal of Educational Measurement*, 40: 4, 281-306

Gierl, M. (2005) Using Dimensionality-Based DIF Analyses to Identify and Interpret Constructs That Elicit Group Differences. *Educational Measurement: Issues and Practice*, 24: 3-14.

Gilliland, J. (1975). *Readability*. London: Hodder and Stoughton.

Graesser, A., McNamara, D., & Louwerse, M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet, & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York: Guilford.

Guthrie, J., & Cox, K. (2001) Classroom Conditions for Motivation and Engagement in Reading. *Educational Psychology Review*, 13: 3, 283-302.

Guthrie, J. T., Solomon, A., & Rinehart, J. M. (1997) Literacy Issues in Focus: Engagement in Reading for Young Adolescents. *Journal of Adolescent & Adult Literacy*, 40: 6, 438-446.

Halliday, M., & Hasan, R. (1976) *Cohesion in English*. New York: Longman.

Hamilton, L. (1998) Gender Differences on High School Science Achievement Tests: Do Format and Content Matter? *Educational Evaluation and Policy Analysis*, 20: 3, 179-195

Harris, T. L. & Hodges, R. (1995) *The literacy dictionary: the vocabulary of reading and writing*. Newark, Del.: International Reading Association.

Hughes, L.E. & Wilkins, A.J. (2000) Typography in children's reading schemes may be suboptimal: Evidence from measures of reading rate. *Journal of Research in Reading*, 23: 3, 314-324.

Janan, D., Wray, D. & Pope, M. (2010) Paradigms in Readability Research, *International Journal of Arts and Sciences*, 3: 17, 19-29.

Johnston, P. H. (1984) Prior knowledge and reading comprehension test bias. *Reading Research Quarterly*, 19: 3, 219-239.

Just, M. & Carpenter,P.A. (1987). *The psychology of reading and language comprehension*. Newton, MA: Allyn & Bacon.

Just, M.A., Carpenter, P.A., and Wooley, J.D. (1982) Paradigms and Processes in Reading Comprehension. *Journal of Experimental Psychology*, 111: 2, 228-238.

Kieffer, M., Lesaux, N., Rivera, M. & Francis, D. (2009) Accommodations for English Language Learners Taking Large-Scale Assessments: A Meta-Analysis on Effectiveness and Validity. *Review of Educational Research*, 79: 3, 1168–1201

Kintsch, W. & Keenan, J. (1973) Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology* 5: 257-274.

Kintsch, Walter (1974). *The representation of meaning in memory*. Hillsdale: Lawrence Erlbaum.

Louwerse, M.M. & Graesser, A.C. (2004) Coherence in discourse. In Strazny, P. (ed.) *Encyclopedia of Linguistics*. Chicago: Fitzroy Dearborn.

Maccoby, E. (1990). Gender and relationships: A developmental account. *American Psychologist*, 45; 4, 513-20.

Martin, M., Mullis, I & Kennedy, A. (2007) (Eds). *PIRLS 2006 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Millard, E. (1997) *Differently Literate: boys, girls and the schooling of literacy*. London: Falmer Press.

Milone, M. 2008. *The development of ATOS: The Renaissance readability formula*. Wisconsin Rapids, WI: Renaissance Learning.

Morgan, P., Farkas, G. & Hibel, J. (2008) Matthew Effects for Whom? *Learning Disability Quarterly*, 31: 4, 187-198

Nation, P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In Schmitt, N. and McCarthy, M. (Eds.) *Vocabulary: Description, Acquisition and Pedagogy* (6-19). Cambridge: Cambridge University Press.

Newbold, N., & Gillam, L. (2010). The Linguistics of Readability: The Next Step for Word Processing. *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing*, (June), 65-72. http://aclweb.org/anthology/W/W10/W10-0409.pdf

Nickson, M. & Green, S. (1996). *A study of the Effects of Context in the Assessment of the Mathematical Learning of 10/11 year olds*. Paper presented to the 1996 British Educational Research Association.
http://www.cambridgeassessment.org.uk/ca/digitalAssets/113794_A_Study_of_the_Effects_of_Context_in_the_Assessment_of_Mathe.pdf

National Institute of Neurological Disorders and Stroke (NINDS) (2011). *NINDS Attention Deficit-Hyperactivity Disorder Information Page.* Bethesda, MD: National Institutes of Health http://www.ninds.nih.gov/disorders/adhd/adhd.htm

O'Reilly, T. & McNamara, D. (2007). Reversing the reverse cohesion effect: good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, 43, 121-152.

Ozuru, Y., Dempsey, K. & McNamara, D. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19, 228-242

Perera, K. (1980) The assessment of linguistic difficulty in reading material, *Educational Review*, 32: 2, 151-161.

Pollitt A (1985) *What makes examination questions difficult?* Edinburgh: Scottish Academic Press.

Rasinski, T. V. (1990). *The effects of cued phrase boundaries on reading performance: A review*. Kent, Ohio: Kent State University, Department of Teacher Development and Curriculum Studies.

Reid, J. (1972) Children‘s comprehension of syntactic features found in extension readers. In Reid, J. (ed) *Reading: Problems and Practices* London: Ward Lock, 394-403

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, 20, 355-371.

Saarnio, D. A., Oka, E. R., & Paris, S. G. (1990). Developmental predictors of children's reading comprehension. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: components skills approaches* (pp. 57–79). San Diego, CA: Academic Press.

Schagen, I. & Sainsbury, M. (1996). Multilevel Analysis of the Key Stage 1 National Curriculum Assessment Data in 1995. *Oxford Review of Education*, 22: 3, 265-272

Schmit, M. & Ryan, A. (1992). Test-taking dispositions: a missing link? *Journal of Applied Psychology*, 77, 629-637

Shorrocks-Taylor, D., Curry, J., Swinnerton, B. & Nelson, N. (2003) National Curriculum Mathematics Tests in England at Key Stage 2: Weights and measures? *Oxford Review of Education*, 29: 1, 51-66

Sonnleitner, P. (2008). Using the LLTM to evaluate an item generating system for reading comprehension. *Psychology Science Quarterly*, 50, 345-362

Spache, G. (1974) *Good reading for poor readers*. Champaign, IL: Garrard.

Stanovich, K. (1986) Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21: 4, 360-407

Stone, E., Cook, L, Laitusis, C.C., & Cline, F. (2010). Using differential item functioning to investigate the impact of testing accommodations on an English-language arts assessment for students who are blind or visually impaired. *Applied Measurement in Education*, 23: 2, 132–152.

Stricker. L. & Emmerich, W. (1999) Possible Determinants of Differential Item Functioning: Familiarity, Interest, and Emotional Reaction *Journal of Educational Measurement*, 36: 4, 347-366

Thompson, C., & Shapiro, L. (2007). Complexity in treatment of syntactic deficits. *American Journal of Speech-Language Pathology,* 18, 30–42.

Thompson, S. J., Thurlow, M. L., & Malouf, D. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology*. 6: 1. Available at www.testpublishers.org/atp.journal.htm

Waller, RHW. (1991) *Designing reports and presentations*. Redhill: Monotype Desktop Solutions.

Wilkins, A., Cleave, R., Grayson, N. and Wilson, L. (2009) Typography for children may be inappropriately designed. *Journal of Research in Reading*, 32: 4, 402-412.

Zumbo, B. (2007) Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, 4: 2, 223–233