# The Readability of Sample Stories for Eye Movement Recording

**Paul Harris, OD**
**Memphis, Tennessee**

## Abstract

***Background:*** *Readability is a measure of the relative difficulty or ease of a particular reading passage. Since the late 1950's, the profession of optometry has been using a series of stories with various eye movement recording equipment. This study aims to assess the homogeneity of passages by grade level and the appropriateness of the grade level assigned to each.*

***Methods:*** *A study of readability statistics was conducted on 97 passages using a number of different formulae. The readability analysis engine in Microsoft and Word Readability Calculations by Micro Power & Light Co. were used.*

***Results:*** *Statistical analysis using ANOVA showed significant differences in the various methods for all passage levels (f<.0001). Certain passages were as much as two years different than their labeled readability levels.*

***Conclusions:*** *Contrary to the intentions of the creators of these stories, a great deal of variability in the readability scores of these passages was found. The results call into question the degree of homogeneity of the stories within a specific grade level, as well as the validity of the grade level assignment for many of these paragraphs. In working through the steps needed to ascertain the readability of these passages, it was also determined that no one readability formula exists that will accurately determine the readability of stories at all levels of reading.*

## Key Words

*Dale-Chall, Educational Developmental Laboratories, Flesch, Flesch-Kincaid, grade level, Gunning FOG, Powers-Sumner-Kearl, readability, reading comprehension, reading level, Spache.*

For more than 40 years, the profession of optometry has relied upon an established series of graded reading passages and norms to evaluate the mechanics of reading. Educational Developmental Laboratories (EDL) initially copyrighted these stories in 1958.[1] The passages tested students from first grade to adult levels. These same stories have continued as the gold standard with a number of different eye movement recording devices including the Reading Eye Camera I and II, EyeTrac I and II, the Applied Science Labs Model 110 and 220, and all versions of the Visagraph.

The stories in question have served as the basis for all eye movement reading studies involving text within the profession of optometry and other disciplines such as education and psychology. The Visagraph manual[2] gives some of the history of how these stories were created and how their readability levels were determined. The following is extracted from the section entitled, "Original Study to Determine Grade Level Norms."

*The subject areas for the various grade levels were as follows:*

*Grade 1 – Pets*
*Grade 2 – Community*
*Grade 3 – Hobbies*
*Grade 4 – Pioneers and Indians*
*Grade 5 – Animals*
*Grade 6 – Foreign Lands*
*Junior High – Inventions*
*High School College – Biography*

*Two selections for each level were available: card #1 to be normally used and card #2 to be used for testing.*

*The test selections and quizzes were carefully prepared in terms of reliability, student interest, appropriateness and consistency of these factors throughout each level, as described below:*

*The vocabulary control used in the preparation of selections for grades 1-3 was a list based on a study of the point of introduction of each word per quarter grade for ten leading basal series. In grade 4 and above, the words used in the selections were taken from a study of basal reader vocabularies but with the modifying control of EDL's Basic Vocabulary of 6,310 words (compiled from a study of 150 words lists, frequency counts, studies, etc.).*

*During the preparations, the selection and quizzes were reviewed by a number of test specialists under the direction of Dr. George Spache of the University of Florida. The content was checked with the following readability formulas: Spache for grades 1-3, Large, Yoakam, Dale Chall, and Flesch for Jr. High and above only. The selections were then modified so that their average readability were as follow: first grade, 1-8, second through sixth grade, mid-year difficulty for the grade; junior high 8.0; and high school-college, 10.5.*

It can be seen that EDL took great care, time, effort and energy to ensure that the stories were indeed at the difficulty levels they purported to be. Over the years, the story sets have

## Table 1 Readability Analysis Engine Descriptions

| Readability Analysis Engine | Relevant Information About the Engine |
|---|---|
| Spache | The **Spache** Formula is vocabulary-based, with special emphasis on the percentage of words not present in the formula's own word list. The words from this sample are counted as unfamiliar (difficult) only once, regardless of the number of times they actually occur in the sample. This formula works best, evaluating materials for use in primary and early elementary grades, through perhaps fourth grade. |
| Dale-Chall | The **Dale-Chall** formula uses its own word list, plus other factors including the total number of words and sentences. The formula implemented is designed for use in assessing upper elementary and secondary level materials. |
| Flesch | The **Flesch Grade Level** Formula is most reliable when used to assess upper elementary and secondary materials. It considers the number of words, syllables, and sentences. The formula implemented is also referred to as the "Flesch-Kincaid" formula. This test is a United States Government Department of Defense standard test.[1] The Flesch-Kincaid Grade Level Score gives a grade level for the reading passage.<br>The **Flesch Reading Ease** score grades a passage on a scale from 0 to 100. The higher the score the easier it is to read. |
| Gunning "FOG" | The **Gunning FOG** formula takes into account the total number of words, words of three or more syllables, and sentences. This formula gives a grade level score that is most suitable for secondary and older primary grades. The formula first calculates the average sentence length (L) by dividing the total number of words in a passage by the number of sentences. It then counts the number of words in the passage with three or more syllables (N). The idea is that the denser a passage is packed with words of three syllables or more, the more difficult the passage. The grade level needed to understand the material is (L + N) 0.4. To convert this to an age, add 5. |
| Powers Sumner Kearl | The **Powers-Sumner-Kearl** Formula is most often used in assessing materials for primary grades. The formula exhibits a ceiling effect, hardly ever producing scores above the 7th grade level. The formula takes into consideration the total number of words, syllables, and sentences. |
| SMOG | The **SMOG** formula relies on a single variable – the number of words containing three or more syllables. NOTE: all of these passages are shorter than the optimum to use this evaluation for any one passage. A sample of 30 sentences is the optimum size for this formula. However, when used for an entire grade level the formula is considered accurate. Most formulas are looking to predict a grade level at which comprehension is at the 75-85% level. The SMOG formula focuses on 100% comprehension and reports the lowest grade that 100% of the children are expected to understand the story. Thus, the numbers are highest for this formula. |

grown to enable test-retest data to be collected on subjects without having to use the same stories repeatedly. The author was not able to ascertain which of the current stories were the original two stories for each level. At present, the best that can be said is to awaken the profession and those who are involved in eye movement recording, to be wary about assuming that reading material is normalized and consistent.

When these reading passages were created, calculations were performed by hand, using specific formulae to determine the reading levels of each passage.[1] In the computer age, a myriad of analysis engines can be performed quickly. This paper compares homogeneity within each of the different stated levels of reading of the EDL paragraphs, as well as the grade level assigned to each passage.

## METHODS

The readability analysis engine in Microsoft Word provides the Flesch Reading Ease Score[1] and the Flesch-Kincaid Grade Level Score.[2] These can be used to evaluate an entire passage or a subsection of a passage. An alternate analysis engine, "Readability Calculations" by Micro Power & Light Co,[3] a Windows based program, is capable of performing up to nine different readability calculations including: Dale-Chall, Flesch Grade Level, Gunning FOG, Powers, SMOG, and Spache. This program also counts the number of sentences and the total number of syllables per paragraph. Each of the 97 EDL paragraphs was analyzed one at a time using the Microsoft Word and Micro Power and Light Co. analyses engines. Table 1 summarizes the readability engines used in the analysis in this paper.

## RESULTS

Table 2 summarizes the readability analyses by each of the analysis methods for each of the grade levels. ANOVA was performed for each grade level, as well as for the entire set of reading passages. Significance was met for each grade level ($f<.0001$) and for the analysis of the stories when taken as a whole ($f<.0001$).

Figure 1 shows a scatter plot of all of the stories analyzed by each of the formulae. This gives a graphic representation of the degree of variability both in terms of the absolute average levels and the standard deviations for each of the readability analysis formulas. Table 3 and Figure 2 show the degree of variability within a single grade using just one of the readability indices, the Flesch-Kincaid. It shows the variation within all the stories that are purported to be at the sixth grade level of difficulty. These range from grade 3.8 to 7.0. The average was indeed grade 6.13 but the degree of variation or standard deviation, was +/- 1.02 grades.

## DISCUSSION

There are many factors that affect readability, comprehension, and what has been called "flow" in reading. These factors include familiarity with the subject area of each passage, the motivation of the test subject, and the role of fonts and paragraph spacing as it affects reading speed.

### Familiarity & Motivation

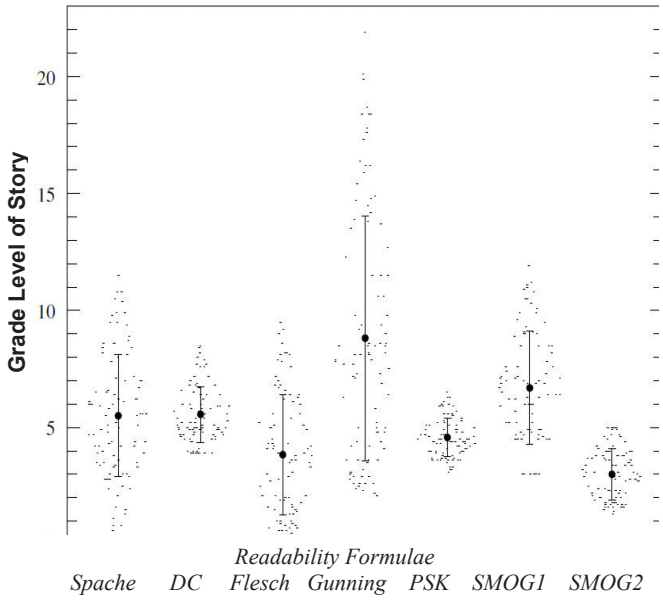
Two very important factors cannot be measured by any objective measure of the text itself. These include familiarity with the subject and the degree of motivation when reading the specific passage. If the reader is familiar with the subject

## Table 2. Readability by Level for Each Analysis Engine

**The SMOG formula relies on a single variable – the number of words containing three or more syllables. A sample of 30 sentences is the optimum size for this formula. The first SMOG column in the table below includes the average for each of the passages with the second number being the standard deviation. **The second column SMOG in the table below is the grade level assigned to the entire grouping of ALL passages within the grade taken together as a whole. This was done to improve the accuracy of the SMOG readability engine which required 30 or more sentences to be accurate.*

| Assigned Grade Level | Spache | Dale-Chall (DC) | Flesch | Gunning "FOG" | Powers Sumner Kearl (PSK) | SMOG* | SMOG** |
|---|---|---|---|---|---|---|---|
| 1 | 1.7 (0.15) | 4.8 (0.48) | 0.9 (0.51) | 3.4 (0.77) | 3.7 (0.30) | 3.92 (0.80) | 4.3 |
| 2 | 1.8 (0.24) | 4.6 (0.51) | 1.6 (0.69) | 5.0 (2.93) | 3.9 (0.26) | 4.58 (1.41) | 5.2 |
| 3 | 2.3 (0.26) | 4.7 (0.51) | 2.0 (0.99) | 5.4 (2.14) | 4.0 (0.39) | 5.16 (1.08) | 5.6 |
| 4 | 2.9 (0.13) | 5.1 (0.44) | 4.2 (0.30) | 8.6 (0.77) | 4.6 (0.15) | 6.66 (0.43) | 6.9 |
| 5 | 3.5 (0.21) | 5.6 (0.40) | 4.8 (0.73) | 9.0 (1.53) | 4.8 (0.24) | 6.89 (0.77) | 7.1 |
| 6 | 3.7 (0.31) | 6.0 (0.52) | 6.1 (0.98) | 12.2 (2.85) | 5.2 (0.35) | 8.08 (0.98) | 8.5 |
| Jr. High (8) | 4.2 (0.42) | 7.0 (0.55) | 7.2 (1.01) | 14.8 2.83) | 5.6 (0.35) | 9.36 (1.01) | 9.8 |
| High School (10) | 4.7 (0.37) | 7.9 (0.49) | 8.6 (1.02) | 18.8 (1.90) | 6.1 (0.34) | 10.61 (0.69) | 11.1 |

*Figure 1. Readability Scatter Plot of Story Grading by Formula*
*This figure shows a scatter plot of how each of the seven readability formulae categorized each of the reading passages. Each dot represents a single story. The center dot is the average readability for all stories as analyzed by a single readability formula and the vertical bar is the standard deviation for that formula.*




matter of a particular passage, they may find the passage much easier to read than if it is filled with new ideas, concepts, or word usage. Although the passage may have shorter words with fewer syllables, the subject may be new to the reader and therefore the reader does not have enough prior knowledge to easily understand the topic. A topic that the reader has prior knowledge of or has some affinity with, may be densely filled with multi-syllabic words but easily be understood.

Motivation also plays a large part in attention devoted to a task. All readability formulae must assume a standard reader with a standard diet, standard levels of rest, and standard desire to take tests and please authority figures, etc. While this may be true for most participants in normative studies, it most likely is not true on an individual basis. The same individual under slightly different circumstances may display a radically different reading ability. The final output can be affected by many variables that are beyond one's ability to control or comprehend. Something as simple as not liking the examiner might significantly change the results. When feeling tired or if distracted, reading a difficult text might be discarded for some lighter fare. At a later time, flow might be achieved.

## Fonts & Spacing

According to the original EDL study:

> *The test cards were printed so as to fulfill the criteria for optimum legibility established by Paterson and Tinker. Letter spacing as well as word spacing was used to insure an even visual density. The selections were photographically enlarged and reduced to allow a constant number of words per line and to maintain a 21-pica line length in spite of the fact that the words grow longer through the grades. Caledonia type was used, in what would approximate 13.7 point type with 4 point leading to 9.3 point type with 2 point leading.*[1,2]

The Reading Eye Camera used a Purkinje image reflection system focused on a moving film plane to record the eye movements. When the shift was made to the infrared eye movement recording systems, the stories were enlarged significantly. Figures 3 and 4 contain examples of these two font sizes.

Figure 3 shows the relative size differences between the two different printings of the same first grade story. The upper part of the figure is the card that was used in the original EDL Eye Movement Camera. The card itself measures 12.7 cm x 8.9 cm. The printed area measures 8.8 cm x 5.3 cm. The lower picture is printed on paper that measures 21.6 cm x 13.8 cm, with the printed area measuring 11.4 cm x 7.0 cm. This amounts to a 29.5% increase in horizontal size and a 32.1% increase in vertical size. It can also be seen that the line spacing was changed to be uniform in the lower larger version.

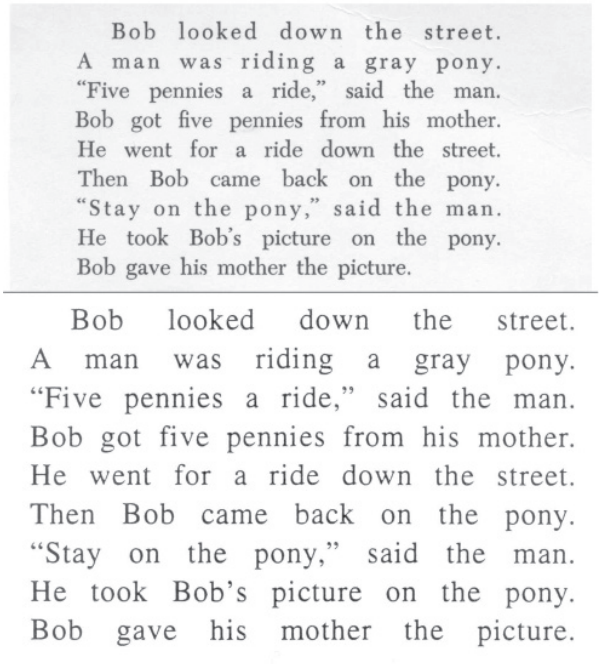
Figure 4 shows similar changes for one of the fourth grade stories. The card or paper sizes were the same as in Figure 3 but the printed areas differ. In the upper original card the story is 8.8 cm wide by 4.9 cm tall. The lower story is 12.3

## Table 3. Readability of Grade 6 Stories

*NOTE: The story number was assigned to the stories by EDL after their creation and are still associated with the same stories currently distributed with the Visagraph III.*

| Grade Level Assigned | Story Number | Flesch-Kincaid |
|---|---|---|
| 6 | 65 | 5.0 |
| 6 | 66 | 5.4 |
| 6 | 67 | 6.4 |
| 6 | 68 | 5.1 |
| 6 | 69 | 4.1 |
| 6 | 70 | 3.8 |
| 6 | 71 | 6.2 |
| 6 | 72 | 6.7 |
| 6 | 73 | 5.6 |
| 6 | 74 | 6.0 |
| 6 | 75 | 7.0 |

*Figure 3: The first story in the group is shown here in the two sizes to scale for comparison. The upper version was to be used with the Reading Eye Camera by EDL. The cards were placed on the lighted holder during testing. The lower version is from the printed book that is part of the Visagraph distribution kit. The scanned images are actual size no manipulation of the image has occurred.*




cm wide by 7.5 cm tall. The magnification is 39.7% in the horizontal and 53.1% in the vertical dimension.

Why were these stories made bigger? In evaluating the Reading Eye Camera, the events, fixations, regressions, and return sweeps were hand counted by viewing the recording on the film directly. A trained observer was able to distinguish between each of these events without difficulty. When the Reading Eye Camera was used, the head was held steady by a headrest and two metal arms that touched the side of the head. Most movement on the traces was attributable to eye movement and not to head movement or other recording artifacts.

*Figure 2: Readability Scatter Plot of 6th Grade Stories by Formula*

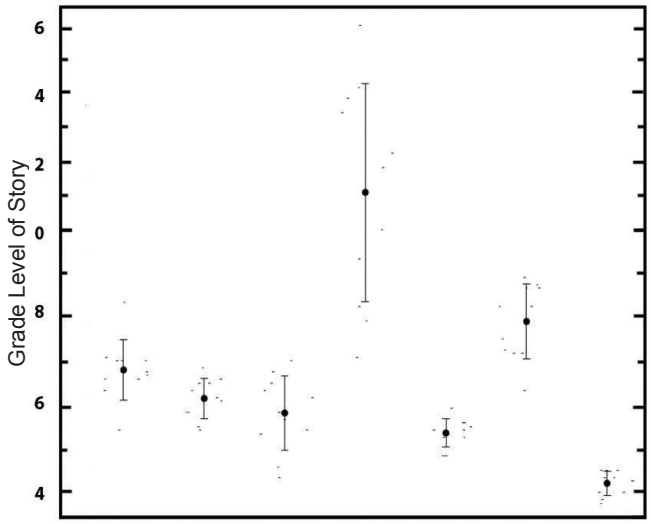
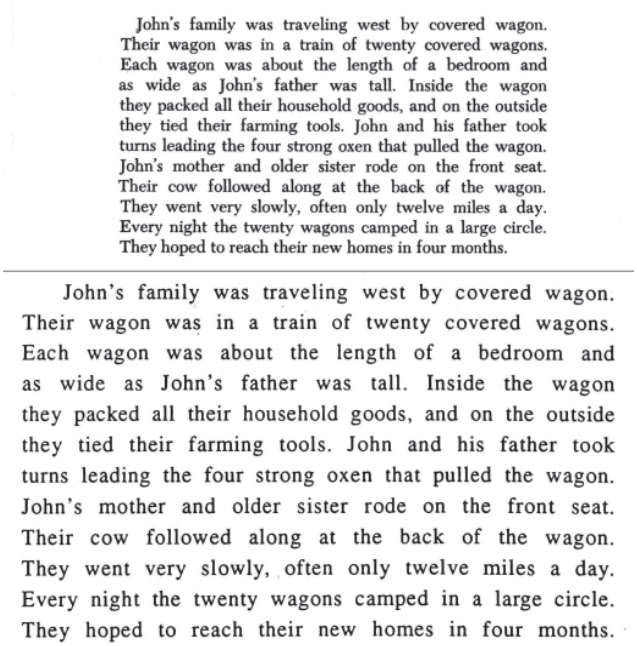



*Figure 4: This figure shows the relative size differences in one of the fourth grade reading passages. Both were scanned at the same time and kept as a single graphic to show the relative differences regardless of the presentation. The upper part of the figure is the card that was used as part of the Reading Eye Camera from EDL and the lower part of the figure is the same passage as distributed as part of the Visagraph reading passages.*




In the electronic form, the analysis of the data is performed automatically with software that is part of the recording system. In most current forms, the subject wears a pair of goggles that house the infrared source, as well as the recording sensors and processors. The subject is free from the headrest and the metal arms that held their head steady. This change could be considered both a positive and a negative. It is positive from the point of view that the technician could now observe the

subject more naturally. Granted it isn't very natural to read with the goggles, but it is better than previous incantations. Early goggles, such as the ones with the OBER2 by Permobil in particular, had very small eye openings and tended to ride high, causing the subject to tip his head forward more than usual. This forced the patient to move his head as they tracked across the text. The goggles were supposed to shield the ambient light sources. Future models solved the ambient light problem and were able to use clear plastic and miniaturize the electronics. So, as close as this is to being natural, it still is reading using strange goggles.

The increase in freedom of movement came at a cost. When reading with significant head movement, the traces move in parallel diagonals during a fixation as the head continues to swing and the subject continues a slow pursuit movement during the fixation. A small saccade now becomes more difficult for the software to identify. During the fixation where the head movement continued, the horizontal position recorded for the eyes continues to change but at a rate slower than the saccade. The saccade itself is then a smaller angle of movement. This has to be considered and overlaid on the diagonally moving eye movements of an overall slow pursuit movement. It becomes harder for the software to decide exactly when or if a saccade has actually taken place. By enlarging the size of the text, the amplitude of the eye movement is increased. In theory, analysis should then be easier for the computer to calculate a saccade versus an inconsistent fixation pattern.

Interestingly, the problems discussed above become magnified as the degree of head movement increases and the maturity of the scan pattern employed by the subject decreases. Early readers make more saccades in comparison to mature readers. Many of the small regressions are of a smaller angular size than a typical forward fixation. These problems conspired in the early electronic versions of the analysis software to cause some graphs either not to be analyzed or to report erroneous counts of fixations and saccades. For the experienced eye movement recorder this posed no problem; the electronic versions allowed for printing of the raw waves that could be counted manually just as with the old Reading Eye Camera. For those with no experience counting events from the films, this posed a significant hurdle in using some of the earlier infrared computerized machines.

One last critical difference was the overall lighting levels when using the different devices. The Reading Eye Camera was used in a nearly dark room. The device had a light that shone directly on the story. However, for the subject, the story was set against a very dark background providing contrast levels well in excess of that normally found in the normal near-work environment. The modern infra-red eye movement recording devices can all be used in full room illumination.

When a series of before and after recordings is performed we need to know with some degree of certainty how difficult the story was for that reader relative to their reading abilities. It may be that the simple "Yes/No" types of quizzes on short samples of reading are not a valid method of assessment. Possibly, we might look into moving away from keeping comprehension within a narrow band for our measures (70-90%) and moving to a different type of testing where comprehension is allowed to vary through a wider range. This comprehension can then be factored back into the performance scores. We may have to assess subjects/patients over a wide range of reading samples at varying levels to titrate the difficulty and do so on an individual basis. Then we may be able to use nearly any reading passage of reasonable length and have an analysis system that accounts for these factors, without having to "grade" the passages beforehand.

## Conclusion

The main question is whether in spite of these differences, can or should the norms from the late 1950's continue to be consulted with the modern day equipment and with the students of the information age. In light of the massive variations from one formula to the other, it must be concluded that no good measure exists that can objectively identify, with the required degree of exactness, a specific grade level for reading passages. These readability measures may best be used to identify relative differences between paragraphs but even that can be questioned. Certainly one could take an ostrich approach to the problem, bury their head in the sand, and pretend the problem of what constitutes a specific grade-level story does not exist. The problem does exist and it calls into question much research done with excellent intentions, that uses these stories.

The main purpose of this paper is to serve as a wake-up call to the profession that assessment of the relationships of eye movements to reading may be much more complex than previously thought. Additional work is necessary to ensure that the stories used are at the levels claimed. By creating an awareness of these difficulties, new stories that are more homogenous in their difficulty levels should be considered. These new stories could then become the basis for new norms for reading mechanics as measured by eye movement devices. This will allow clinicians working with their patients and researchers working with their subjects the opportunity to obtain results without readability contributing to possible variability.

## References


1. Reading Eye Test Selections, Huntington, NY: Educational Developmental Laboratories, 1958.
2. Taylor S,, Nystrom K. User's Guide Visagraph II Eye-Movement Recording System Appendix D, Original Study to Determine Grade Level Norms, 2000.
3. Frantz E. Readability Calculations, Micro Power and Light Co. Dallas, Texas.www.micropowerandlight.com. Last accessed 2 March 2011.
4. Johnson C, Johnson K. Readability and reading ages of school science text-books.www.timetabler.com/reading.html. Last accessed 2 March 2011.



Acknowledment: I would like to thank Meggan Heinz, OD for assistance in the statistical analysis.



Corresponding Author:
Paul Harris, OD
Southern College of Optometry
1245 Madison Avenue
Memphis, TN 38104
pharris@sco.edu



Date accepted for publication: 3 March 2011