

# Learning from previous experiments: readability in national mathematics assessments

First submission: 11 January 2012

Acceptance: 20 June 2012

This article aims to investigate the challenges associated with understanding the instructions in the Common Tasks for Assessment (CTA) by a Grade 9 mathematics class of second-language English speakers. The design of the CTA was such that a series of tasks was set using an extended context. The Flesch-Kincaid readability test judged the instructions of some tasks to be beyond the readability levels of an average Grade 9 learner. Some instructions had high lexical density, showing that these were difficult to understand. It is recommended that education authorities note the challenges associated with contextualised assessment such as the CTA, and ensure that the challenges associated with understanding instructions in national assessment should not outweigh the benefits of accessing the mathematics in different settings.

## Leer vanuit vorige ervarings: die leesbaarheid van nasionale wiskundetake

Hierdie artikel ondersoek die uitdagings wat die instruksies in die Gemeenskaplike Asseseringstaak (CTA) vir Wiskunde bied aan Graad 9-leerders met Engels as addisionele taal. Die asseseringstaak is so ontwerp dat leerders 'n aantal take binne 'n uitgebreide konteks moes uitvoer. Volgens die Flesch-Kincaid leesbaarheidstoets, is sommige instruksies by die take bokant die vermoë van die gemiddelde Graad 9-leerder. Die leksikale digtheid van sommige instruksies het getoon dat die instruksie baie moeilik was om te begryp. Daar word aanbeveel dat die onderwysowerhede kennis moet neem van die uitdagings verbonde aan sulke gekontekstualiseerde assesering en moet verseker dat die uitdaging om die instruksies in nasionale asseseringstake te verstaan nie die voordeel verbonde aan die assesering van wiskundige vaardigheid in verskillende kontekste oorskadu nie.

*Dr S Bansilal & Ms M Khan, School of Education, University of KwaZulu-Natal, Private Bag X03, Ashwood 3605; E-mail: Bansilals@ukzn.ac.za & khammb786@gmail.com.*



*Acta Academica*  
2013 45(1): 79-99  
ISSN 0587-2405  
© UV/UFS  
<<http://www.ufs.ac.za/ActaAcademica>>

SUN MEDIA  
BLOEMFONTEIN

A great deal has been written in South Africa about why Curriculum 2005 (C2005) with its outcomes-based philosophy did not succeed in creating a sound and effective education system. One reason for the failure of the curriculum is that a large sector of the education system was not sufficiently sophisticated to deal with the required demands for self-authority (Jansen 1998, Rogan 2008, Chisholm *et al* 2001). Chisholm & Leyendecker (2008) argue that the failure lies in expectations that education would lead to transformation without paying necessary attention to implementation and capacity. Amidst all the discussion and interrogation of the failings of C2005, hardly any attention has been paid to another innovation introduced in 2002 as part of C2005. The Common Tasks for Assessment (CTA) was an external summative assessment programme which the Department of Education (DoE) introduced at the Grade 9 level in all South African schools in 2002 as part of the curriculum reform process. The CTA was designed to assess whether Grade 9 learners had achieved the outcomes of each learning area (DoE 2002: 9). The design was planned to be of performance-based assessment which the DoE viewed as a vehicle intended to allow learners to demonstrate what they would do in a real-life situation (DoE 2002: 9). However, the CTA was unofficially terminated in 2010, without much interrogation of why it did not work. Many remnants in the form of assessment tasks still exist, some of which have been reconfigured as mathematical literacy assessments. Although there have been several innovative CTA tasks in mathematics that would work well in a classroom, many of them were of poor quality which would not yield valid assessments of learners' capabilities (Bansilal 2008a & 2008b).

We believe that reflection on the mathematics CTA and the reasons why it did not work can provide insight into the design of current and future assessment programmes in mathematics and mathematical literacy. In previous work, challenges have been identified to the validity and fairness of the assessment function of the CTA, and validity issues related to teacher mediation of the CTA assessment tasks are discussed (Bansilal 2008a & 2008b, Khan 2009, Bansilal 2010).

The validity of an assessment rests on the interpretation of the information it provides. Nitko (2001) explains that the concept of validity applies to the ways in which assessment results are interpreted

and used. Nitko (2001) also cautions that the uses one makes of assessment results are valid only to the extent to which one can point to evidence in support of their correctness and appropriateness. This article focuses on one issue that affects the validity of assessments, namely the extent to which learners understand the language used in the task. Accordingly, this article aims to report on the study on the readability levels of the instructions in one task, and to explore the learners' perceptions about the language used in the contextualised task. It is hoped that this article will add knowledge regarding the challenges associated with designing contextualised assessments in mathematics, which will contribute to improved planning of future assessment interventions.

## 1. Literature review

The CTA was designed in such a manner that all the tasks were based on one extended real-life context. This approach of drawing upon real-life contexts in mathematics assessments is a global trend and is also evident in South African policy documents (DoE 1997 & 2003). In recent years, a large body of literature has highlighted both the benefits and limitations of using the everyday in mathematics. Barry Cooper wrote extensively with various colleagues about the use of these real-life contexts in national assessment tasks in Britain (Cooper 1998, Cooper & Dunne 1998 & 2004, Cooper & Harries 2003). Cooper & Dunne (1998) drew upon data from primary schools to show that certain realistic mathematics test items were associated with the underestimation of learners' existing capacities in mathematics. Some learners drew upon their everyday knowledge instead of their mathematics knowledge, because they did not understand the question. When probed later during interviews, these learners revealed that they did know the correct answers but that they had misjudged the demand of the question.

Associated with the use of contextualised tasks is the heavy language load of the tasks, which adds a layer of complexity to the tasks, and may block children from understanding what it is they are required to do. In South Africa this can cause greater problems for the majority of learners who speak English as a second language (ESL) but are taught in English. Hugo *et al* (2010) point out that, in KwaZulu-Natal, 88% of

learners have isiZulu as their home language and only 8% of learners speak English at home. In terms of the language of teaching and learning, only 27% of learners are taught in isiZulu, whereas 70% of learners are taught in English. The authors quantify that 1.5 million isiZulu home-language learners are being taught in English. This is nearly 60% of all learners in the province. Reddy (2006) points to the differentiation in performance between African and non-African learners in the Trends in International Mathematics and Science Study (TIMSS) 2003 as well as in the national exit level examinations in South Africa. African learners attending non-African schools (like the sample in this study) performed better in their Grade 12 examinations than their counterparts in African schools (Kahn 2004). When Clark & Linder (2006) asked a group of Grade 9 learners from a township science classroom to list the words from a newspaper article that they did not understand, the learners listed the following ten words from a single paragraph in a newspaper article: 'blaze', 'confirmed', 'guts', 'arson', 'halt', 'engulfed', 'battled', 'abandoned', 'suggested' and 'cable'. Clark & Linder (2006) noted the importance of background knowledge in reading comprehension. A consequence of the learners' weak language skills was that they found it difficult to cope with questions in text written at a high level of complexity and to make sense of questions requiring more interpretation than those to which they were accustomed.

Another South African study (Vale 2012) carried out with 43 ESL learners investigated student errors related to the linguistic complexity of mathematical literacy test items. Vale (2012) argues that, unlike home-language speakers, ESL learners have to make an effort to decode text and information in English. Vale (2012) found that all the participants lost marks due to low levels of language proficiency. Language proficiency is important both for the reading comprehension required to decode the instructions and for the writing communication required to encode or communicate the results to the reader.

Using data from 8.000 students, Shaftel *et al* (2006) examined the relationship between item linguistic characteristics as independent variables and item difficulty as the dependent variable. The design of the study allowed them to identify and quantify effects of specific language characteristics on the item difficulty.

Prins & Ulijn (1998) carried out a study with 108 students, who were 17-18 years old, by administering three versions of nine tasks, namely original (O), adapted (A) and non-verbal (N). The adapted (A) versions were modified to make the tasks more readable, while the non-verbal (N) versions did not have any references to context. Students were divided into three groups, based on whether their first language was English (E1), Afrikaans (E2) or an African language (E3). An important result was that all three language groups performed equally well on the non-verbal (N) versions, showing that the E3 students had the same computational skills as their E1 and E2 counterparts, but that they performed more poorly on the original (O) items. The results also showed that the two second-language (E2 and E3) groups gained more by improved readability than the E1 (English first language) group.

Abedi & Lord (2001) administered some original items from the National Assessment of Educational Progress (NAEP) along with parallel items of reduced linguistic complexity to 1.174 Grade 8 learners in the US. Linguistic modification resulted in significant differences in mathematics performance, with scores on the modified items being slightly higher. Students with a low or average mathematics background benefited more than those from a high-achieving or advanced mathematics background. Similar to the results of Prins & Ulijn (1998), second-language English speakers also benefited more from the modifications than first-language English speakers.

Dempster & Reddy (2007) analysed three readability factors (sentence complexity, unfamiliar words and long words) to investigate the readability of the TIMSS items in science. They found that sentence complexity influenced the performance of learners on TIMSS items, and the effect was more pronounced in learners with limited proficiency in English than those who were more proficient in English. They acknowledge that readability alone cannot explain the poor performance of the majority of South African learners in the test.

Prins & Ulijn (1998: 141) define readability as “the ability of the text to communicate the intention of the writer to the intended reader”. Rakow & Gee (1987: 28) define readability as “an estimate of probability of comprehension by a particular group”. They

express concern that the readability of tests is especially important for large-scale assessments: “Unless you have worded your test items so students are sure to understand what you are asking them, you may be challenging their reading ability rather than their grasp of scientific concepts”. Thompson *et al* (2002) emphasise the need to use plain language when vocabulary level is not part of the construct being tested. They provide recommendations as part of the principle of universal design in assessment intended to ensure fair and reliable large-scale assessment.

Methods for analysing readability have increased significantly in recent times. While Harrison & Bakker (1998) comment that until 1963 alone, there were 31 readability formulae, Benjamin (2012) estimates that by the 1980s there were up to 200 and since then their number has greatly increased and would be difficult to quantify. Harrison & Bakker (1998) used a combination of two conventional readability formulae with a test for lexical density in order to obtain a fuller picture of the complexity of the readability of text. Shorrocks-Taylor & Hargreaves (2000) used a variety of readability formulae together with a readability formula designed specifically for use with mathematics texts. Their analyses show that, when these formulae were applied to the same texts, the correlations between most of them were highly significant, suggesting close relationships between them. Halliday (1993) suggests that scientific (and mathematical) texts have a very high ‘lexical density’. That is, they have a large number of lexical items (or content words) per clause. Informal spoken language has a lexical density of about two words per clause, and written English has a lexical density of about four to six words per clause.

In this article we use an easily accessible tool, namely the Flesch-Kincaid tool available on any Word programme, to judge the readability of tasks set within real-life contexts. We will show that, according to this test, much of the text in the CTA is beyond the comprehension level of the average Grade 9 learner. We also used Halliday’s (1993) measure of lexical density to gain further insight into the challenges of understanding individual instructions.

## 2. Methodology

The study that informed this article combined elements of quantitative and qualitative analysis. The participants in the study were 44 ESL learners who were attending an English-medium high school. Their class marks and the CTA marks were analysed using quantitative methods. The class marks were generated from continuous assessment procedures, including assignments and tests, throughout the first three school terms, while the CTA assessment was completed in the fourth term. Three learners (above average, average and below average) were interviewed. A qualitative document analysis on the CTA task was also used. Cohen *et al* (2000) state that content analysis could be used in the analysis of educational documents. While content analysis can clarify the content of the document, it can also throw “additional light on the source of communication, its author, and on its intended recipients, those to whom the message is directed” (Cohen *et al* 2000: 165). In this situation we studied the tasks set by the national DoE in order to cast more light on the demands of the real-life tasks themselves.

The purpose of the document analysis was to investigate the readability of instructions and some passages appearing in the tool, while bearing in mind that for these learners English was a second language (ESL). The two corresponding research questions were “What do the readability levels and lexical density measures of the instructions in the CTA suggest about the accessibility of the instructions?” and “What are the learners’ perceptions of the language used in the tasks?”.

As mentioned earlier, there are numerous readability formulae (Benjamin 2012), and we opted to use the Flesch-Kincaid formula, since it is most commonly used and widely available. The formula was devised by Rudolf Flesch in 1975 and can be automatically calculated on any Microsoft document. Allan *et al* (2005) state that the Flesch-Kincaid formula is useful in rating educational materials in the form of extended passages of continuous text. The Flesch-Kincaid Grade Level Index is derived from the former formula and was originally designed to check manuals produced for the US armed services (Allan *et al* 2005). In the Flesch Reading Ease test, higher scores imply that the material is easy to read and lower scores indicate that it is harder

to read. Although the two tests use the same core measures (word length and sentence length), they have different weighting factors. The results of the two tests thus correlate approximately inversely: a text with a comparatively high score on the Reading Ease test should have a lower score on the Grade Level test. Scores can be interpreted as shown in Table 1.

Table 1: Mapping of Flesch Reading Ease score to readability level

Flesch Reading Ease score	Readability level	Estimated school grade completed
0-29	Very difficult	College
30-49	Difficult	High school or some college
50-59	Fairly difficult	Some high school
60-69	Standard	7th or 8th
70-79	Fairly easy	6th
80-89	Easy	5th
90-100	Very easy	4th

An interpretation of these measures is found in Wikipedia’s description. *Reader’s Digest* magazine has a readability index of about 65, *Time* magazine scores about 52, an average 6th grade student’s (an 11-year-old) written assignment has a readability test of 60-70 (and a reading grade level of 6-7), and the *Harvard Law Review* has a general readability score in the low 30s (Wikipedia).

### 3. Results

The first subsection quantitatively analyses learners’ results in their class assessment compared to their results in the CTA. The second subsection is devoted to a discussion of the readability and complexity of a selection of activities from the CTA. Finally, the results of the interviews with three learners are presented.

#### 3.1 Comparison of marks in class

We first analysed the marks of each of the 44 learners in the class. We found that the marks were much lower for each child in the CTA than they were in their class assessments (CAs). The relationship between the marks obtained in the CA and those obtained in the CTA was



investigated using the Pearson correlation coefficient. There was a strong positive correlation between the two variables,  $r = .75$ ,  $n = 44$ ,  $p < .001$ . This strong correlation implies that learners who performed better than others on the class assessments also performed better than others on the CTA tasks. However, the means of the class for the average of the class marks and the CTA were 58 and 34, respectively, a 24 percentage point difference, showing that on average learners only achieved 58% of their term marks in the CTA. Of the students, 64% achieved 50 and above in their term marks, while only 16% of the class obtained 50 and above in the CTA. The box plot (Figure 1) illustrates graphically the differences in percentage points between the two sets of scores.

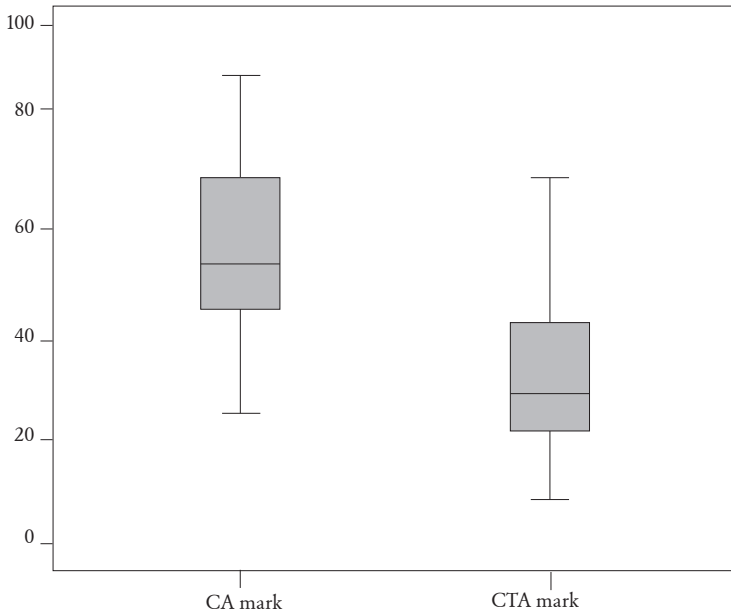


Figure 1: Box plot showing class assessments and CTA marks

This box plot illustrates clearly the wide difference in scores between the two assessments. A paired samples t-test was conducted to evaluate

the changes in the learners' scores on the class marks (CA) compared to the CTA marks. There was a statistically significant decrease in the scores from the class mark ( $M=57.7$ ;  $SD=16.9$ ) to the CTA mark ( $M=33.6$ ;  $SD=16.3$ ),  $t(44)=13.72$ ,  $p<.005$ . The mean decrease in scores was 24.1.

The results demonstrate that the learners in the class performed significantly less well in the CTA than in their CA. These results then prompted the authors to seek reasons to explain the large decline in marks. Accordingly, we studied the tasks in further detail in an effort to understand why the 44 learners performed more poorly in the nationally administered CTA programme. The larger study (Khan 2009) found many other factors that contributed to the poor performance, such as the teachers' explanations, teachers' marking, too much textual information, overload of unnecessary visual signs, unnecessary context information, ambiguous instructions, context-specific terminology that was unfamiliar to learners, and errors in the marking memorandum. In this article we report on one issue, namely the readability of instructions, bearing in mind that other factors also contributed to the poor performance.

### 3.2 Exploring the readability statistics

This section examines certain activities with respect to their readability levels, using easily available measures of readability. Altogether there were four tasks but the participants did not proceed beyond Tasks 1 and 2. We did readability tests on the instructions of Activities 1.1, 1.2, 1.3, 1.4, 2.1, 2.2 and 2.3 from Tasks 1 and 2. In order to increase the reliability of the readability test outcomes, we combined certain activities in order to make up the length requirement of 100 words (Allan *et al* 2005, Shorrocks-Taylor & Hargreaves 2000). Recognising that the combinations would lose focus on individual instructions and sentences, we then used a lexical density test to investigate the sentence complexity of individual instructions. The readability statistics are presented below in Table 2. This is followed by the selected activities (Figure 2).

Table 2: Readability statistics for selected activities

Instructions of activity	Word total	Flesch Reading Ease	Flesch-Kincaid grade level index
1.1 and 1.2	115	78%	5
1.3	117	63%	7
1.4	203	43%	12
2.1 and 2.2	117	34%	11.5
2.3	197	55%	10.1

Consider Activities 1.1 and 1.2 in Figure 2. The instructions for Activities 1.1 and 1.2 (when combined) have a high Flesch readability index of 78% (from Table 1). The readability test results suggest that, in general, a Grade 6 learner should be able to understand the combined instructions which could therefore be considered to be fairly easily understandable by the Grade 9 learners. Activity 1.1, however, was allocated 0 marks, and success in this item would not have added to a learner's overall score.

However, the instruction in the second bullet of 1.2 is complicated by the different instructions or prepositional phrases ('complete the table'; 'convert the distance'; 'measured on the map'; 'choose a scale'; 'describe a conversion'). The sentence complexity is not identified by the readability score which rates the passage as easily understandable. To obtain a clearer idea of the individual instructions, we used Halliday's lexical density measures to evaluate sentence complexity. We found that the lexical density for the second bulleted instruction for Activity 1.2 is reasonably high – there are 15 content words per clause (considering the entire sentence as one clause) – that is, the sentence has a lexical density of 15.



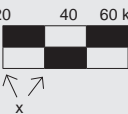




Scale	Distance measured on a map	Actual distance
1 mm is 1 km	50 mm	$\frac{50 \text{ mm}}{1 \text{ mm}} \times 1 \text{ km} = 50 \text{ km}$
1 mm is 10 km	28 mm	1.2.1.....
1:100	40 mm	$40 \text{ mm } 100 = 4\,000 \text{ mm}$
1:100 000	3 cm	1.2.2.....
20    40    60 km 	$y$ 	$\frac{y}{x} \times 20 \text{ km} =$
		1.2.3.....
(Try more than one scale) 1.2.4.....	20 mm	40 km

Figure 2: Activities 1.1 and 1.2

We now consider Activity 1.4 which appears in Figure 3.



1.4.1 (a) Use the map on page 4 to estimate the distance from your school to the marked entrance to the Kruger National Park. (5)

(b) Why can the answer in (a) only be an estimate? Give 2 reasons. (2)

1.4.2 When the CTA was copied at a certain school the map on page 4 was reduced to 80%. What influence can this have on your answer in 1.4.1? Explain your answer. (3)

1.4.3 Up to 1994 the KNP stretched 350 km along the Mozambican border and was on a verage 60 km wide. Use this information to determine the approximate area of the Park before 1994, in km<sup>2</sup>. (2)

1.4.4 According to one source, the actual area of the Kruger National Park before 1994 was 2 149 700 hectares. Convert your answer from 1.4.3 to hectares and explain why your answer differs from the actual area. (2)

1.4.5 According to an agreement with the governments of Mozambique and Zimbabwe, the Kruger National Park will become part of the great Limpopo Transfrontier Park. The eventual size of this Park will be 100 000 km<sup>2</sup>. Calculate the % increase if this Park compared to the size of the Kruger National Park before 1994 (2 149 700 ha). (3)

Figure 3: Activity 1.4

From Table 2, it is evident that Activity 1.4 in Figure 3 had a low readability score of 43% and a grade level index of 12. A reading suggests that instructions 1.4.3, 1.4.4 and 1.4.5 from this activity are reasonably difficult to follow. The Flesch-Kincaid tests indicate that learners below a grade level of 12 would find it difficult to understand the language. In addition, the lexical density scores for Question 1.4.5 revealed that the last sentence was the most dense, containing 12 content words. The first sentence, when viewed as containing two clauses, has 5 and 9 content words in the first and second clause, respectively. However, this sentence contains 14 content words altogether, making it a very dense statement. The instruction for Question 1.4.5 then requires the learner to try to make sense of this information comprising a heavy language load.

Across the selection of tasks that were shown, the readability measures confirm that the instructions are too wordy and not easily

understood. They do not conform to recommendations for maximum readability and comprehension suggested by Thompson *et al* (2002). Although questions 1.1, 1.2 and 1.3 are easily understandable, these tests indicate that other questions would not have been easily understood by the average English-speaking Grade 9 mathematics learner. For ESL learners, it would have been even more challenging to decode the instructions.

In the assessment setting for Grade 9 mathematics learners, these demands were experienced as overwhelming. Considering the readability demands and lexical density of texts in Activities 1.2 and 1.4, together with the demands of synthesising a large load of information for Task 2, it is not surprising that the learners did not know what to do. If learners do not understand the instructions, they cannot proceed further to demonstrate their mathematics knowledge and skills. We now hear from three learners about their evaluation of the language used in the tasks.

### 3.3 Interviews

We report on three interviews with a sample of three students whose performance in their class assessments varied.

#### 3.3.1 Sihle

Sihle is a 15-year-old African male learner whose home language is isiZulu, but he uses English as a medium of communication with his peers. Sihle obtained an average of 56% in his class mark (CA) and 37% in the CTA. In his interview, his closing comments were

For the CTA I would say now those who like the department of education shouldn't just send it ma'am [...] They should not be sending such hard things for pupils, which they know some pupils may lack from schoolwork just because not understanding what is going on. And for a teacher, ma'am, the teacher must not just come and leave the CTA on the table without explaining it to the pupil. First the teacher has to explain the CTA to the pupil so that the pupil can give him- or herself time to learn how to do the CTA.

This excerpt conveys his frustration at not being able to understand what he was supposed to do. His comment “which they know some pupil may lack from schoolwork because not understanding what is going on” is not very clear because of his grammar but when probed further it emerged that he meant that learners may perform badly in

school work (such as CTA), because they did not understand the tasks or what was expected from them. He was also disappointed that his teacher had not explained to them what they were supposed to do – he wished somebody could have explained what was expected from him.

### 3.3.2 Thabani

Thabani is a 15-year-old African male learner whose home language is isiXhosa. Thabani is an outstanding award-winning learner, participates in sport, is a class representative on the School Representative Council for Learners, is the captain of the English junior debating team, and participates regularly in interschool debates. He also participates in various mathematics challenges and has received various medals for mathematics Olympiad challenges. He obtained 86%, 92% and 100% for his first, second and third term marks, respectively, leading to an average of 93%. However, Thabani obtained 50% in the CTA.

Thabani's response in the interview conveyed a sense of his anger and disappointment. He described his perceptions of the CTA as follows:

The first time I saw it [CTA] it was very difficult for me to relate to the questions and it was something I have never done before. I did make an attempt but some of the questions I wasn't able to answer them.

The following excerpt details his responses when asked to explain further:

*Researcher:* What was the problem?

*Thabani:* ... I wasn't, I was not able to understand what were the questions.

*Researcher:* What made the questions so difficult that you could not understand? You are a very *good* maths student. So why did you have a problem?

*Thabani:* Just the way it was written.

*Researcher:* How was it written?

*Thabani:* The language.

Thabani identified the language as a barrier to his understanding of the instructions. His later comments reveal how powerless he felt:

I was under a lot of pressure. Because just being me and not being able to answer the questions. A lack of understanding just makes you feel

stupid about yourself. But with the CTA it seems like almost all the questions were like difficult. I wasn't quite certain what I was writing.

These comments convey his anger at not being able to understand the questions, which left him feeling stupid and angrier. His closing comment, "I wasn't quite certain what I was writing", suggests that he responded without understanding what he was supposed to do.

### 3.3.3 Cleo

Cleo is a 15-year-old African female learner who is an average student in mathematics. She obtained a class mark (CA) of 47% and 28% for the CTA.

In her interview, Cleo indicated that she was disappointed that the CTA was so different from the usual mathematics tasks they did in their classroom assessments. When asked how it was different, she replied "It's because, ma'am, we had to like, we had to read the passage and then we had to answer, and then you have to calculate". This suggests that Cleo saw the passages of language as a barrier between herself and what she needed to do [calculate]: "The lines. I mean not the lines, the passage" is what made it different from the usual maths tasks.

She reiterated that the CTA was "different from doing maths" or "classwork" She shared her frustration in trying to navigate through all the "passages" in order to get to the mathematics:

... in maths we like expecting things like you gonna get sums, and then they gonna say calculate the distance of this. And then here we get we had to calculate from Skukuza. You get to, what's this? [...] There's solve for  $x$ , OK. Its like calculate this to get this[ ...] Why don't they just say OK, here there is 20 km, OK, calculate here and here? Like don't *say*, don't give us the passage. The passage. I mean the words *ja*. And just give us the equations only.

We can hear her frustration when she cries out, "don't say, don't give us the passage", showing that she is tired of reading through the lines and trying to understand the instruction. She pleads: "... just give us the equations only".

## 4. Discussion and concluding remarks

The results of the internal assessment scores and the external CTA revealed that the differences were statistically significant. One of the issues uncovered in this study concerns the language and instructions



of the task. We used two easily available readability instruments to measure a sample of questions. Although readability statistics are not absolute measures, they can provide broad guidelines about whether a passage will be easily understood by Grade 9 learners. We then used a lexical density test to check individual sentence complexity. These tests together showed alarming trends in that many of the instructions and passages would be considered too dense to be understood by the average Grade 9 learner. The interviews with three learners of varying mathematical ability also identified the readability of the instructions of the CTA as a major hurdle for them and they expressed frustration in trying to understand what they were expected to do in the mathematics assessment. They did not attempt to demonstrate their competence in the required mathematics procedures or reasoning because of their failure to understand the instruction itself. This barrier implicates the validity of the CTA as a tool to measure their mathematics competence. Thabani, an award-winning learner in mathematics Olympiad competitions, achieved 100% in a class-based assessment, yet the results of the CTA tool suggest that he can successfully solve only 50% of Grade 9 level tasks. In addition, the performance of all the learners was, on average, 24% lower than their class assessments. Noting that the concept of validity applies to the ways in which we interpret and use the assessment, if the CTA information was used for benchmarking or diagnostic purposes, the information provided by the test is not valid.

Abedi & Lord (2001) comment that in their study higher performing learners also had strong language ability and did not find it difficult to understand original items. Students in low-level mathematics classes benefitted more from the language modification of test items. However, in this study, all three learners identified the language as a barrier to their identifying the 'calculation' that was needed. Even Thabani who was a high achiever was unable to cross the readability barrier.

It is important to note that, in addition to the readability factor, the validity of this assessment tool was also compromised by other factors such as teachers' mediation attempts, poor face validity, excessive use of unnecessary visual signifiers, messy numbers, context-specific language that the learners did not understand, implicit assumptions made by the task designers, and convoluted mark schemes. However,

on the basis of this study, we are not able to make pronouncements about which of these variables had a stronger effect on the results, or which could account for most of the variance in scores. Further research similar to that carried out by Shaftel *et al* (2006), that could attempt to quantify effect sizes associated with the different variables, would add greatly to knowledge about contextualised mathematics assessments.

We want to emphasise that our criticism of the CTA tool is about its suitability as a summative assessment, and not about its use in classroom activities. Tasks presented in the CTA would work well in a classroom setting with discussions and investigations facilitated by the teacher who could help learners understand the instructions and dense language.

What are some of the lessons that this CTA experiment in assessment has taught us? Strategies to aid ESL learners are always relevant and the use of a readability test could help task designers in the development of valid assessments. When mathematics assessments are designed using large passages of text, task designers should ensure that the information is easily understood. Shaftel *et al* (2006: 121) comment that test developers should “pay greater attention to the general language development of the students being tested and use wording that does not introduce additional comprehension hurdles over and above the required content”.

In addition, authorities should ensure that assessment programmes are not dominated by contextualised assessments. The use of an extended context as a basis for all the assessments is not recommended. Application of mathematics to real life is but one aspect of overall mathematics proficiency, and assessment programmes must cater for a range of different skills. We also recommend that assessment tasks set around extended contexts requiring the use of much information would be better placed as projects or investigations that could be carried out over a period of time and not in an examination-type setting. We want to reiterate that contextualised assessments are an important facet of mathematics learning, and that skilful design of contextualised mathematics items must ensure that the task is not obscured by the language.

## Bibliography

- ABEDI J & C LORD  
2001. The language factor in mathematics tests. *Applied Measurement in Education* 14(3): 219-34.
- ALLAN S, M MCGHEE & R VAN KRIEKEN  
2005. *Using readability formulae for examination questions*. London: Qualifications and Curriculum Authority.
- ALLEN B & S JOHNSTON-WILDER (eds)  
2004. *Mathematics education*. London: Routledge Falmer.
- BANSILAL S  
2008a. Assessing the validity of the Grade 9 mathematics common tasks for assessment (CTA). Presentation at the 5th conference of the Association of Commonwealth Examinations and Accreditation Bodies, 9-14 March 2008, Pretoria.
- 2008b. Does the CTA assess what it claims to? Presentation at the KZN regional meeting of the Association for Mathematics Educators in South Africa (AMESA), 22 February 2008, Durban.
2010. How much freedom does a teacher have in designing a learning event when adhering to assessment prescription? *Education as Change* 14(1): 77-90.
- BENJAMIN R J  
2012. Reconstructing readability: recent developments and recommendation in the analysis of text difficulty. *Educational Psychology Review* 24(1): 63-88.
- CHISHOLM L & B LEYENDECKER  
2008. Curriculum reform in post-1990s sub-Saharan Africa. *International Journal of Educational Development* 28: 195-205
- CHISHOLM L, J VOLMINK, T NDHLOVU, E POTENZA, H MAHOMED, J MULLER, C LUBISI, P VINJEVOLD, L NGOZI, B MALAN & L MPHAHLELE  
2000. *A South African curriculum for the twenty-first century*. Report of the Review Committee on Curriculum 2005. Pretoria: Department of Education.
- CLARK J & C LINDER  
2006. *Changing teaching, changing times. Lessons from a South African township classroom*. Rotterdam: Sense Publishers.
- COHEN L, L MANION & K MORRISON  
2000. *Research methods in education*. London: Routledge Falmer.
- COOPER B  
1998. Using Bernstein and Bourdieu to understand children's difficulties with 'realistic' mathematics testing: an exploratory study. *Qualitative Studies in Education* 11(4): 511-32.

- COOPER B & M DUNNE  
1998. Anyone for tennis? Social class differences in children's responses to national curriculum mathematics testing. *The Sociological Review* 46(1): 115-48.
2004. Constructing the 'legitimate' goal of a 'realistic' maths item. Allen & Johnston-Wilder (eds) 2004: 69-90.
- COOPER B & T HARRIES  
2003. Children's use of realistic considerations in problem solving: some English evidence. *Journal of Mathematical Behaviour* 22: 451-65.
- DEMPSTER E & V REDDY  
2007. Item readability and science achievement in TIMSS 2003 in South Africa. *Science Education* 91: 906-25.
- DEPARTMENT OF EDUCATION (DOE)  
1997. *Curriculum 2005: lifelong learning for the 21st century*. Pretoria: Department of Education.
2002. *Draft framework for the development of CTAs*. Pretoria: Department of Education.
2003. *National Curriculum Statement, Grades 10-12 (General): Mathematical Literacy*. Pretoria: Government Printers.
- HALLIDAY M A K  
1993. Some grammatical problems in scientific English. Halliday & Martin (eds), 1993: 69-85.
- HALLIDAY M A K & J R MARTIN (eds)  
1993. *Writing science: literacy and discursive power*. London: Falmer.
- HARRISON S & P BAKKER  
1998. Two new readability predictors for the professional writer; pilot trials. *Journal of Research in Reading* 21(2): 121-38.
- HUGO W J M, V WEDEKIND & D WILSON  
2010. *The state of education in KwaZulu-Natal: a report to the Provincial Treasury*. Pietermaritzburg: KZN Provincial Treasury.
- JANSEN J D  
1998 Curriculum reform in South Africa: a critical analysis of outcomes-based education. *Cambridge Journal of Education* 28: 321-31.
- KHAN M B  
2009. Grade 9 learners' experiences of the Common Tasks for Assessment in Mathematics. Unpubl M Ed thesis. University of KwaZulu-Natal.
- KAHN M  
2004. For whom the school bell tolls: disparities in performance in senior certificate mathematics and physical science. *Perspectives in Education* 22(1): 149-56.

- NITKO A J  
2001. *Educational assessment of students*. New Jersey: NJ: Merrill Prentice-Hall.
- PRINS E D & J M ULIJN  
1998. Linguistic and cultural factors in the readability of mathematics texts: the Whorfian hypothesis revisited with evidence from the South African context. *Journal of Research in Reading* 21(2): 139-59.
- RAKOW S J & T C GEE  
1987. Test science, not reading. *Science Teacher* 54(2): 28-31.
- REDDY V  
2006. *Mathematics and science achievement at South African schools in TIMSS 2003*. Cape Town: Human Sciences Research Council.
- SHAFTTEL J, E BELTON-KOCHER, D GLASNAPP & J POGGIO  
2006. The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment* 11(2): 105-26.
- SHORROCKS-TAYLOR D & M HARGREAVES  
2000. Measuring the language demands of mathematics tests: the case of the statutory tests for 11-year-olds in England and Wales. *Assessment in Education* 7(1): 39-60.
- THOMPSON S J, C J JOHNSTONE & M L THURLOW  
2002. Universal design applied to large scale assessments. NCEO Synthesis Report 44.
- VALE P  
2012. Mathematical literacy test items and student errors: investigating linguistic complexity. Presentation at the UMALUSI conference, 10 May 2012, Muldersdrift, South Africa.
- WEBER E (ed)  
2008. *Educational change in South Africa: reflections on local realities, practices, and reforms*. Rotterdam: Sense Publishers