Strategies for answering multiple choice questions among South African learners

What can we learn from TIMSS 2003?

Presented at the 4th Sub-Regional Conference on a Assessment in Education, hosted by Umalusi from the 26th to the 30th June 2006

Edith R. Dempster, School of Education & Development, University of KwaZulu-Natal

Abstract

This study investigated the strategies used by South African learners in answering text-only multiple choice items in TIMSS 2003. The investigation focused on 20 items in which learners showed a strong preference for an incorrect answer. The trend was particularly noticeable among learners who attended schools in which all the learners and teachers speak an African language as their home language. The study arose from a concem about the readability of TIMSS items as a valid instrument for assessing the scientific knowledge of learners who lack proficiency in English.

Results are presented for two groups of learners: those who attend 'Africans only' schools, and those who attend mixed-race schools where most of the teachers speak English as a home language ('non-African schools'). These schools previously catered for the Coloured, Indian and white communities.

Sentence complexity was significantly higher in questions where >40% of learners chose an incorrect answer than in questions where learners answered by random guessing, and where >40% of learners chose the correct answer. Only 12 out of 73 items were correctly answered by >40% of learners from 'African' schools, whereas 36 items were correctly answered by >40% of learners attending 'non-African' schools.

Five strategies were detected in items in which >40% of learners attending 'African' schools chose an incorrect answer. The most common strategy (9 items) consisted of favouring the

answer that contained the greatest number of familiar words. Learners rejected answers containing words that they did not recognize. Other strategies were selecting answers that contained words that were also in the question (2 items), choosing an answer that contained all the options (1 item), misunderstanding the question (2 items), and selecting answers that indicate misconceptions (4 items). Two items could not be classified.

The strategies indicated that learners did not understand the twenty questions clearly, and resorted to strategies based on 'tricks' that, with luck, might be the correct answer. Learners attending 'non-African' schools also resorted to 'tricks', although learners were more likely to select the correct answer than the incorrect answer. The results of this analysis raise further concerns about the level of content knowledge among South African Grade 8 learners, and therefore the content validity of TIMSS in South Africa. The lack of appropriate content knowledge is compounded by low levels of proficiency in English, coupled with unclear wording of some TIMSS items.

Introduction

TIMSS 2003 has once again confirmed the poor performance of South African Grade 8 learners in Mathematics and Science at Grade 8 level (Reddy 2006). As with the two previous studies, South African learners achieved the lowest score of all participating countries, including other African countries such as Ghana, Botswana, Tunisia and Morocco. The 2003 study was the first to be conducted with a group of learners who had experienced outcomes-based education throughout their schooling, yet their scores were the same as that achieved in 1998, and lower than the scores achieved in 1995.

In 2003, 8,952 learners participated in the TIMSS study, 11.6% of whom wrote the test in Afrikaans, while the remainder wrote the test in English. Most of those who wrote in English were not home-language speakers of English, with 84% of them being African learners attending schools that served the African population group under the apartheid system of government. The performance of this group was significantly lower than that of learners who attended mixed-race schools that were previously designated for the white, Indian and Coloured population groups (Reddy, 2006). In 2002, the enrolment in 'African' schools was about 96% African, while 'non-African' schools had about 40-50% African learners, and about 50% learners of the race group for which the school was originally established (Reddy 2006).

The universal validity of TIMSS questions has been questioned by a number of researchers, since the content tested may not match the content taught in all the participating countries. TIMSS items are validated by a panel comprising representatives from all participating countries. Items are included if they match the taught curriculum in 70% of the participating countries. In 2003, South African official curriculum was C2005, in which content was deliberately under-specified. Teachers were expected to choose content that would enable them to achieve the learning outcomes, within content areas that were specified only in very broad terms. Thus the content tested in TIMSS may have been invalid for South African schools, since it could vary between schools, and since the official intended curriculum was a deliberately vague document (Reddy 2006).

A scan of TIMSS released items raised our concerns about an additional source of invalidity of TIMSS 2003: the readability of certain multiple choice items, particularly for South African learners. English is the official medium of instruction in most South African schools, including those where all learners and teachers speak one or more of the indigenous African languages. This description applies to 84% of the learners who wrote the TIMSS tests in English. For these learners, English is a second language in the urban schools, and a foreign language in rural schools (Setati et al, 2002), and their scaled science score in TIMSS 2003 was 199, compared with a mean of 483 for learners attending the most resource-rich schools in South Africa, which were previously designated for the white community. The international average scaled score was 474 (Reddy 2006).

A previous study investigated the readability of 73 multiple choice items from the Science section of TIMSS 2003 (Dempster & Reddy in press). The selected items consisted only of text, thus eliminating confounding variables such as visual literacy, or ability to interpret

graphs or tables. We hypothesized that the percentage of learners selecting the correct answer on each item would be negatively correlated with readability factors such as sentence complexity, number of unfamiliar words and number of long words.

Readability formulae such as the Flesch-Kincaid Grade Level Formula and the Fog Formula are only reliable if they are applied to a sample of text containing at least 100 words (Allan et al., 2005), a condition that is generally not met by multiple choice items. Evaluating the readability of multiple choice items is therefore problematic.

Reading research has shown that vocabulary load and syntactic complexity are the most robust predictors of readability. Vocabulary load is measured by the number of syllables, or number of letters in a word, or location on a frequency list. Syntactic complexity is measured by the number of words per sentence, and/or the number of embedded clauses or phrases within the sentence, or the lexical density (number of content words per clause). After testing several models and variables over ten years or more, Homan et al (1994) arrived at the prediction equation that became the Homan-Hewitt Readability Formula. This formula was designed for short sections of text, such as is found in multiple choice questions.

Three factors proved to be good predictors of readability and were used in the Homan-Hewitt readability formula. These factors were:

- sentence complexity, measured as the average number of words per Hunt's T-Unit (WNUM). Homan et al. (1994) defined sentence complexity as the average number of words per clause, where a clause is the shortest grammatically correct sentence into which a piece of writing can be divided. Thus a complex sentence contains one T-unit, whereas a compound sentence may contain two or more T-units.
- number of unfamiliar words, defined as the number of words that were not familiar to 80% of 4th-Grade children (WUNF), as listed in The Living Word Vocabulary (Dale & O'Rourke 1981).
- Number of long words, measured by how many words have seven or more letters (WLON).

The Homan-Hewitt Formula is stated below.

Readability Level = 1.76 + (0.15 x WNUM) + (0.69 x WUNF) – (0.51 x WLON)

The outcome of the formula relates to Grade level for American learners. The formula was validated with a sample of 782 learners with known reading ages, selected from the 2nd to 5th-grade classes in five elementary schools. All the children included in the sample had mastered the subject matter of the test at a level of 75%, thus eliminating content validity as a potential confounding factor (Homan et al. 1994). The readability levels proved to be good predictors of performance, with learners able to correctly answer more questions where the readability level was at or below the child's grade level.

Allan et al. (2005) have criticized the usefulness of readability formulae for examination questions, stating that subject-specific terminology is likely to inflate the readability level,

while it is part of the knowledge required for the subject. They also point out that readability formulae such as Flesch-Kincaid, Fog, and Homan-Hewitt are calibrated for American Grade levels. If used in other countries, they need to be calibrated against the reading competence of children in those countries.

Given the concerns expressed about the applicability of readability formulae to examination questions, and the necessity to calibrate the grade levels to the country, we did not calculate a grade level for TIMSS items (Dempster & Reddy in press).

Since *The Living Word Vocabulary* was developed and standardized for American children, it was not an appropriate word list for South Africa. We were unable to locate a similar index of familiarity for South African learners, and substituted a primary school dictionary for South African learners in its place. The dictionary selected contains a list of 3 300 words that learners should know by the end of primary school, or Grade 7 (Blacquière et al. 1995). The assumption is that Grade 8 learners in South Africa should be familiar with most of the words in this junior dictionary. Sentence complexity and number of long words are not affected by the context of the country or the school, and were calculated in the same way as Homan et al. (1994).

We found that learners attending 'non-African' schools were more likely to randomly guess an answer if the sentence complexity was high, while African learners at African schools were more likely to select an incorrect answer if sentence complexity was high (Dempster & Reddy in press). The average number of unfamiliar words, and the average number of long words per item were not correlated with percentage of learners answering correctly.

An average of 28% of African learners attending African schools selected the correct answer in 73 selected items. This was significantly lower than the average of 39% for learners attending non-African schools. Moreover, learners attending African schools were more likely to select randomly among the answers, and more likely to favour an incorrect answer than learners from non-African schools. In fact, learners attending African schools showed a strong preference for an incorrect answer in 20 of the 73 items, or 27.4% of the items. In these items, 40% or more of the learners selected one of the distractors. By contrast, preference for an incorrect answer occurred in nine items among learners attending non-African schools.

This study seeks explanations for the pattern referred to above. What strategies were learners using in selecting a plausible answer? Can the strategy be explained in terms of the readability of the items? Pollitt and Ahmed (2001) have questioned whether TIMSS assesses science or readability. They analysed several TIMSS items, and showed that learners could apply different forms of logic that led them to an incorrect answer, depending on how they interpreted the question. This study attempts to do the same for 20 TIMSS items in which 40% or more of the South African learners favoured an incorrect answer. The phenomenon occurred more frequently among African learners attending African schools, where the level of proficiency in English is lowest among the range of South African schools. However, it is also apparent among learners attending non-African schools, whose proficiency in English is higher.

Methods

The method of sampling schools, learners, design of test items and administration of the test is described elsewhere (Mullis et al., 2001). A total of 8,952 learners from 192 schools participated in TIMSS 2003 in South Africa. While the total number of learners who wrote the test in English is 7,912, the number of learners per question is smaller because in terms of the TIMSS matrix design, not all learners answered every question. The results presented here represent 6,658 learners attending African schools and 1,254 learners attending non-African schools throughout South Africa.

The study estimated readability using three variables identified by Homan et al. (1994) as the most useful predictors of readability of multiple choice questions: sentence complexity, number of unfamiliar words, and number of long words. Words that were not listed in the Shuters Junior Dictionary for Southern Africa were classified as 'unfamiliar words. Where a word used in a TIMSS item differed from the listed word by more than two letters, it was classified as unfamiliar. Thus for example, 'burn' is listed in the dictionary, but not 'burning', therefore 'burning' is classified as an unfamiliar word. In addition, many words take on different meanings when used in science than in other spheres of life. If the scientific meaning of the word was not given in the dictionary, it was classified as an unfamiliar word. This technique probably under-estimates the number of unfamiliar words for African learners in African schools, but it was the closest available technique in the a bsence of a familiarity index for South African learners.

Twenty items in which 40% or more of the learners attending African schools selected an incorrect answer were studied to understand how learners may have reached their choice. We also relied on anecdotal evidence gained from students and learners on strategies that they use when confronted with multiple choice questions.

Results

Mean (±SD) sentence complexity for the 20 selected items was 12.7±4.6, mean number of unfamiliar words per item was 4.4±4.5, and mean number of long words was 5.2±3.8. Sentence complexity was significantly higher in questions where >40% of learners chose an incorrect answer than in questions where learners answered by random guessing, and where >40% of learners chose the correct answer. Mean numbers of unfamiliar words and long words did not differ significantly among the three types of answers.

The items were divided into six categories, based on the strategy that best explains the pattern of choice. Examples are given below for released items only.

1. Choosing an answer that contains a term that is also in the question. Two items of the 20 fell into this category.

Example 1:

The burning of fossil fuels has increased the carbon dioxide content of the atmosphere. What is a possible effect that the increased amount of carbon dioxide is likely to have on our planet?

A. A warmer climate
B. A cooler climate
C. Lower relative humidity

D. More ozone in the atmosphere

Sentence complexity of the question is 14.67, it contains 8 unfamiliar words and 12 long words. The correct answer is A.

Table 1: Percentage of learners selecting each answer in example 1

	A	B	C	D	n
African schools	19.6	17.3	12.3	50.8	1019
non-African schools	45.8	8	12.3	34	212

The word *atmosphere* is the only word that occurs in the question and in option D. It was a strong distractor for African learners. Learners attending non-African schools favoured the correct answer, but the second most popular choice was D. Anecdotal evidence from learners and university students indicates that matching terms in the question and in the answers is a widespread practice when students do not understand the question, or do not know the answer. It was also evident in questions where two or more possible answers contained terms that were in the stem, as in example 2.

Example 2:

A balloon filled with helium gas is set free and starts to move upward. Which of the following best explains why the helium balloon moves upward?

A. The density of helium is less than the density of air.
B. The air resistance lifts the balloon up.
C. There is no gravity acting on helium balloons.
D. The wind blows the balloon upward.

Sentence complexity is 8.1, there are 10 unfamiliar words, and 11 long words. The correct answer is A.

Table 2: Percentage of learners selecting each answer in example 2

	A	B	C	D	n
African schools	14.2	37.6	17.9	30.3	1021
non-African schools	40.9	26.1	19.7	13.3	203

The word *helium* appears in the question, and also in answers A and C, but it is an unfamiliar word and answers A and C are rejected by African learners. Option A also contains the unfamiliar word *density*, and is not favoured by the African learners. Options B and D contain the words *balloon* and *up* or *upwards,* which are also in the question. These were favoured by 67% of the African learners. Learners in non-African schools favoured the correct answer A, but the second-favourite answer was B, indicating that learners applied the same logic as learners in African schools.

2. Choosing the option that contains words that are familiar. Nine items of the 20 were placed in this category.

Example 3:

The sun is an example of which of the following?

A. Comet
B. Planet
C. Galaxy
D. Star

The sentence complexity is 11, and the item contains three unfamiliar words and two long words. The correct answer is D.

Table 3: Percentage of learners selecting each answer in example 3.

	A	B	C	D	n
African schools	5.3	42	12.7	40	1022
Non-African schools	7.6	25.3	9.1	58.1	198

Comet and *galaxy* are unfamiliar words, and few learners selected those answers. Answers B and D were equally attractive to learners in African schools, while over half of the learners in non-African schools selected the correct answer, D. Nevertheless, distractor B was the second most popular answer for this group.

Example 4

Which group of energy sources are ALL renewable?

A. Coal, oil and natural gas
B. Solar, oil and geothermal
C. Wind, solar, and tidal
D. Natural gas, solar and tidal.

Sentence complexity is 8, and there are eight unfamiliar words and 5 long words. The correct answer is C.

Table 4: Percentage of learners selecting each answer in example 4

	A	B	C	D	n
African schools	54.5	10.4	16.5	18.6	510
non-African schools	37.6	8.3	35.8	18.3	109

Renewable, solar, geothermal, and *tidal* are unfamiliar words in this item. Understanding the term *renewable* is key to understanding the question. Clearly, in this item, African learners favoured answer A, which consisted only of familiar words. Among learners attending non-African schools, A was a more popular choice than the correct answer, C.

3. Answers that indicate that learners misunderstood the question. Two items were place in this category.

Example 5:

A small, fast-moving river is in a V-shaped valley on the slope of a mountain. If you follow the river to where it passes through a plain, what will the river most likely look like compared with how it looked on the mountain?

A. Much the same.
B. Deeper and faster.
C. Slower and wider.
D. Straighter.

Sentence complexity is 14.7, and the item contains 4 unfamiliar words and 5 long words. The correct answer is C.

Table 5: Percentage of learners selecting each answer in example 5

	A	B	C	D	n
African schools	12.7	41.8	24.4	21.1	994
non-African schools	8.1	34.4	41.6	15.8	209

It appears that learners attending African schools understood the question as "What will the river look like on the mountain?" The item is very badly phrased, with many qualifiers obscuring the question. Learners attending non-African schools fared somewhat better, with 41.6% selecting the correct answer, but B still attracted 34.4% of the learners.

4. Answers that indicate misconceptions or misunderstanding of the concept. Four items fitted into this category.

Example 6:

Which of the following is NOT a mixture?

A. Smoke
B. Sugar
C. Milk
D. Paint

Sentence complexity is 8, there is one unfamiliar word, and two long words. The correct answer is B.

Table 6: Percentage of learners selecting each answer in example 6

	A	B	C	D	N
African schools	46.4	15.2	26.9	11.5	988
Non-African schools	41.4	22.7	27.1	8.9	203

Clearly option A was the most popular choice for both groups of learners. D was the least popular choice for both groups. It indicates a misunderstanding of the concept of mixtures, which can exist in gaseous form as well as in liquid form.

5. Selecting the answer that contains all the options. One item illustrated this strategy.

Example 7

The fossils that are found in the oldest layers of sedimentary rock were formed from which types of organisms?

A. Only organisms that lived in the sea.

B. Only organisms that lived on land.
C. Only organisms that lived in the air.
D. Organisms that lived on the land, in the sea and in the air.

Sentence complexity is 19 for the question, 7 for option A, 6 for B, 7 for C and 13 for D. The item contains 5 unfamiliar words, and 7 long words. The correct answer is A.

Table 7: Percentage of learners selecting each answer in example 7.

	A	B	C	D	n
African schools	11.3	30.2	12.4	46.1	1030
non-African schools	12.3	25.1	3	59.6	203

The key clue to answering this question is the word *oldest*, and the question assumes that learners know something about the sequential evolution of vertebrates in the history of life. In the absence of this knowledge, selecting D covers all the options, and was more popular among learners attending non-African schools than those attending African schools. The second-favourite selection was B, and can be explained in terms of learners' experience of rocks on land, rather than sedimentary rocks forming in the sea.

6. Unknown explanation.

Two items could not be categorized in any of the above categories. One is an unreleased item, and the other is shown here.

Example 8

Eating leafy vegetables is important for human health. This is because leafy vegetables are a good source of which of the following?

A. Protein
B. Carbohydrates
C. Minerals
D. Fat

Sentence complexity is 11, the item contains nine unfamiliar words and eight long words. The correct answer is C.

Table 8: Percentage of learners selecting each answer in example 8

	A	**B**	**C**	**D**	**N**
African schools	43	28.4	18.4	10.3	1057
non-African schools	47.3	22.4	27.3	2.9	205

Minerals is defined in the Shuters Junior Dictionary in terms of substances extracted from the earth during mining operations. Thus, minerals as part of a healthy diet may be an unfamiliar use of the term for South African learners. However, it is unclear why protein should be such a popular choice for both groups of learners.

Discussion

Content validity is questionable for TIMSS in the South African context, since the science covered in South African schools differs substantially from other countries that participated in TIMSS (Reddy 2006). This clearly affected the way South African learners experienced the TIMSS tests, with many items testing knowledge that South African learners had not acquired through the Natural Sciences curriculum. Lack of content knowledge meant that few questions were analysed and answered from learners' knowledge base, but from an alternative set of strategies. Compounding the lack of subject knowledge is the fact that most of the South African learners were participating in TIMSS with insufficient proficiency in English to enable them to understand all the questions.

The items analysed in this study share a common factor: all of them resulted in learners favouring an incorrect answer because of the strategies they used to select an answer to a multiple choice question. Where learners could not apply their strategies, they resorted to random guessing, which was more prevalent among the learners attending African schools than those attending non-African schools. Only 12 of the 73 items included in this study were correctly answered by more than 40% of the learners attending African schools, while 36 of the 73 items were correctly answered by more than 40% of the learners attending non-African schools.

Pollitt and Ahmed (2001) have pointed out that questions have construct validity only if the learners' minds do what we intend them to do when answering the question. They present evidence that by focusing only on the content words in an item, learners could (and do) select an incorrect answer to an item. Linguists agree that the lexically heavy content words (nouns, verbs, adverbs and adjectives) receive more attention from readers than the smaller grammatical words (Pollitt & Ahmed 2001). Through analyzing the content words and the associations these words may evoke in learners' minds, Pollitt & Ahmed (2001) made predictions about which choices would be favoured by learners. They illustrate their predictions with a number of TIMSS items where the pattern of choice across the four or five alternatives indicates rejection of certain answers, and favouring of the correct answer plus one incorrect answer. The incorrect choices can be explained in terms of Pollitt and Ahmed's predictions.

In the case of South African learners, whose scientific knowledge has a weak match with the knowledge tested in TIMSS, and who have the added disadvantage of writing the test in their second language, the strategies they use often have little to do with making sense of the question. Many African learners attending African schools are multiply disadvantaged by poor resources in their schools and at home, teachers who are not confident or well-qualified in science, and they lack proficiency in English (Reddy 2006). South African learners attending non-African schools generally have better-resourced schools and homes, better-qualified teachers, and better proficiency in English. Despite the unequal provision of resources, the best-performing South African learners (those who attend the ex-white schools) score only slightly above the international average.

Hewitt and Homan (2004) have highlighted the importance of readability in large-scale tests, referring to readability as the forgotten validity variable in standardized test items.. Our

analyses have demonstrated a link between sentence complexity and the ability of learners to select the correct answer. The complexity of the grammatical structure of some TIMSS items is not fully captured in a single variable, which counts only the number of words in each Hunt's T-unit. It does not capture the obscuring of the question within complex sentences, the use of the passive voice, questions phrased in the negative, large numbers of qualifying phrases and clauses, and reliance on prepositions and logical connectives.

Overall, the number of unfamiliar words was not significantly associated with the number of learners answering correctly, but analysis of individual items and learners' preferences demonstrates that unfamiliar words affect learners' choice, because they tend to favour answers that contain familiar words. Where the item gave no linguistic cues to work with, the learners' responses indicated random or near-random choice. This was the case for African schools with 17 of the 73 items, and for non-African schools, 7 items indicated random choice. Occasionally, the strategies result in a correct answer being selected, which then obscures the undeniable fact that most South African learners know very little of the science that is tested in TIMSS. If learners answered questions where they do not know the science by guessing randomly, the chance of choosing the correct answer is 25% (or 20% for questions with five alternatives). However, if learners apply strategies that lead them to the incorrect answer, the percentage selecting the correct answer drops to below 25%. The effect of so many questions where the percentage choosing the correct answer is well below 25% is to reduce the average percentage correct to a level just above guessing—the average percentage correct per item was just 28% for African learners attending African schools. The overall score on TIMSS may have been substantially better if learners had guessed the answers or left out questions they did not understand rather than applying linguistic strategies that led them to an incorrect answer.

This study highlights the importance of careful scrutiny of the wording of items to minimize the kinds of strategies used by South African learners in this study. Familiarity indices for all grades would be extremely helpful in writing textbooks and in setting examination papers that are accessible to all learners. Glossaries and dictionaries containing the terminology required for success at all levels of schooling would help teachers and learners to acquire the vocabulary necessary for acquisition of concepts and skills in terminology-rich subjects such as science. Learners should be directed away from resorting to the kinds of linguistic strategies evidenced in this study, since these strategies obscure the purpose of the test, which is to find out how much science learners know.

References

- Allan, S., McGhee, M. and van Krieken, R. (2005) *Using readability formulae for examination questions*. London: Qualifications and Curriculum Authority.
- Blacquière, A., Hoyle, P. and Thompson, C.I. (1995) *Shuters Junior Dictionary*. Pietermaritzburg: Shuter & Shooter.
- Dale, E. & O'Rourke, J. (1981) *The Living Word Book Vocabulary*. Chicago: World Book International.
- Dempster, E.R. and Reddy, V. In press. 'Item readability and science achievement in the TIMSS (2003) study in South Africa'. In *Journal of Science Education*.

- Hewitt, M.A. and Homan, S.P. (2004) 'Readability level of standardized test items and student performance: the forgotten validity variable'. In *Reading Research and Instruction* 43: pp. 1-16.
- Homan, S.P., Hewitt, M. and Linder, J. (1994) The development and validation of a formula for measuring single-sentence test item readability. In *Journal for Educational Measurement* 31 pp. 349-358.
- Mullis, I.V.S., Martin, M.O., Smith, T.A., Garden, R.A., Gregory, K.D., Gonzalez, E.J., Chrostowski, S.J. & O'Connor, K.M. (2001) TIMSS Assessment Frameworks and Specifications 2003. Boston College: International Study Center.
- Pollitt, A. and Ahmed, A. (2001) 'Science or Reading? How students think when answering TIMSS questions'. International Association for Educational Assessment Conference paper.
- Reddy, V. (2006) Mathematics and Science Achievement at South African schools in TIMSS 2003. Cape Town: HSRC Press.
- Setati, M., Adler, J., Reed, Y. & Bapoo, A. (2002) Code-switching and other language practices in mathematics, science and English language classrooms in South Africa. In J. Adler & Y. Reed (Eds.), *Challenges of Teacher Development: An Investigation of Take-up in South Africa* (pp. 72-93). Pretoria: Van Schaik.