Per Linguam 2014 30(2):1-19
http://dx.doi.org/10.5785/30-2-602

Journal Home Page:
http://perlinguam.journals.ac.za

Per Linguam

A Journal for Language Learning
Tydskrif vir Taalaanleer

ISSUES OF VALIDITY AND GENERALISABILITY IN THE GRADE 12 ENGLISH HOME LANGUAGE EXAMINATION

Colleen du Plessis
University of the Free State and Inter-Institutional Centre for Language Development and Assessment (ICELDA)

Very little research has been devoted to evaluating the national English Home Language (HL) curriculum and assessment system. Not only is there a lack of clarity on whether the language subject is being offered at an adequately high level to meet the declared objectives of the curriculum, but the reliability of the results obtained by Grade 12 learners in the exit-level examination has been placed under suspicion. To shed some light on the issue, this study takes a close look at the language component of the school-leaving examination covering the period 2008-2012, to see whether evidence of high language ability can be generated through the current selection of task types and whether the inferred ability can be generalised to non-examination contexts. Of primary interest here are the validity of the construct on which the examination is built and the sub-abilities that are being measured, as well as the validity of the scoring. One of the key findings of the study is that the language papers cannot be considered indicators of advanced and differential language ability, only of basic and general proficiency. The lack of specifications in the design of the examination items and construction of the marking memoranda undermine the validity and reliability of the assessment. As a consequence hereof, the inferences made on the basis of the scores obtained by examinees are highly subjective and cannot be generalised to other domains of language use. The study hopes to draw attention to the importance of the format and design of the examination papers in maintaining educational standards.

INTRODUCTION

The Home Language component of the National Senior Certificate (NSC) is a much neglected area of research. Very little attention has been devoted to evaluating the curriculum and its accompanying assessment protocol, and as a consequence hereof the results obtained by learners in the Grade 12 examination have been placed under suspicion amidst perceptions of a lowering of standards. Without evidence of a decline in standards, the chances are that no steps will be taken to introduce any corrective measures on the Home Language front.

An overview of research conducted by the Council for Quality Assurance in General and Further Education and Training (Umalusi), the statutorily mandated overseer of the National Senior Certificate (NSC), shows that only peripheral attention has been devoted to determining the standards of the Home Language subjects and that these have also not been subjected to benchmarking, unlike in the case of English First Additional Language (cf. Umalusi, 2009a, 2009b, 2009c, 2010, 2011). There is little clarity as to whether the Home Language curriculum is being offered at an adequately high level as patently required by the

old and new curricula, and uncertainty about the reliability of the results obtained by learners in the exit-level examination. The low pass mark of 40% required for a Home Language subject (Department of Basic Education, 2013b: 36) and the high average pass rate of 94% for English HL over the period 2008-2012 (Department of Basic Education, 2012a: 59) serve to reinforce perceptions of low standards.

This article takes a close look at the language component of the Grade 12 English Home Language examination papers covering the period 2008-2012, in an attempt to shed some light on the issue of standards and the generalisability of language ability inferred on the basis of examination scores to non-examination domains. Of primary interest in the content analysis of the selection of examination papers are the validity of the construct on which the examination is built, the definition and description of the sub-abilities that are being assessed, the choice of language tasks and the validity of the scoring. The latter aspect is equally important, as unreliable scoring will obstruct the inference of language ability and will therefore, with regard to current, orthodox notions of validity, undermine any validity argument (Read, 2010; Van der Walt, 2012). As a secondary objective, this study hopes to draw attention to the importance of the format and design of the examination papers in maintaining standards. Evidence that language learning is taking place needs to be generated through the system of assessment being used. Owing to the fact that severe shortcomings have been identified with the school-based continuous assessment (Umalusi, 2012c: 30), which contributes up to 37.5%[i] of the final overall percentage obtained by each learner, the examination papers serve as the most concrete form of evidence of the extent of teaching and learning in the current dispensation, underlining their importance.

BUILDING A VALIDITY ARGUMENT

Much of the emphasis in the validation of any language test or examination centres around the issue of construct validity – a notion which 'integrates considerations of content, criteria and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility' (Messick, 1995: 742). In line with current thinking, it is not the test or language examination instrument on its own that is considered to be valid, but the kind of evidence or data produced through the scores obtained by candidates on the basis of which inferences of ability can be made (cf. Kane, 2004: 136; Weir, 2005: 12; Chapelle, 2012: 21). This evidence pertains to the selection of examination content and tasks, the contexts within which the latter are to be performed and scored, and the effect and justifiability of inferences based on the scores (cf. Weideman, 2012).

Defining and articulating the construct is considered absolutely essential from the start of the design process (Weir, 2005: 18) if appropriate evidence of ability is to be produced. The confidence that may be placed in any language examination is considered to be directly proportional to the evidence collected in the process to support the validity of the evaluation instrument (Davies *et al.*, 1999: 220; Van der Walt, 2012: 145). Accordingly, construct validity can be supported in language testing by ensuring that the abilities to be assessed are founded on accepted theories of language, cognition and communicative competence, and by aligning these with suitably designed task types (Blanton, 1994; Bachman & Palmer, 1996; Van Dyk & Weideman, 2004a, 2004b; Weideman, 2011; Young, 2012; Patterson & Weideman, 2013).

The superordinate construct of the Home Language curriculum that derives from a content analysis of the newly introduced Curriculum and Assessment Policy Statement (CAPS) and its predecessor on which it is premised, the National Curriculum Statements (NCS), has been conceptualised in a report for Umalusi as follows:

> 'The assessment of a differentiated language ability in a number of discourse types involving typically different texts, and a generic ability incorporating task-based functional and formal aspects of language' (Du Plessis, Steyn & Weideman, 2013: 20).

This construct can be elucidated further on the basis of the distinction made in CAPS between two levels of mastery identified by Cummins (Cummins & Davison, 2007: 353): 'interpersonal communication skills required in social situations and the cognitive academic skills essential for learning across the curriculum' (Department of Basic Education, 2011: 8). Whereas the first type of skills, *basic interpersonal communicative skills (BICS)*, pertains to conversational language, the latter kind, *cognitive academic language proficiency (CALP),* requires of learners a much higher order of thinking and level of language ability. It is this more advanced proficiency that is to be reflected in the differentiated and generic abilities examined in the Home Language papers.

CAPS underwrites the principle of '*high* knowledge and *high* skills' and the minimum standards to be attained are to be '*high*, achievable standards in all subjects' (Department of Basic Education, 2011: 4). In an attempt to clarify the potential confusion surrounding the term Home Language, CAPS makes it clear that the reference to Home Language shall 'be understood to refer to the *level* and not the language itself' (emphasis added; Department of Basic Education, 2011: 8). It is thus the standard of the curriculum that distinguishes Home Language from First Additional Language, and in this context the accompanying level of examination should be that traditionally associated with the assessment of a first language. Since determining the desired standard of the examination paper can be seen as an integral part of the articulation of its construct, any undermining of what is supposed to reflect a high ability will weaken the validity argument. Advanced ability is implied inherently in the conceptualised construct of CAPS in the differing ability required of examinees to respond to a variety of discourse types that demand the mastery of distinct language features.

The attention devoted to ensuring construct validity in language testing relates, furthermore, to the importance of being able to generalise the results obtained by candidates to non-testing domains (Young, 2012). In the absence of generalisability, little value can be attached to the results obtained in an examination beyond the assessment context. In this sense a test or examination should correlate well with 'indices of behaviour that one might theoretically expect it to correlate with' (Weir, 2005: 18). Weir (2005) refers to such correlation as theory-based validity, an intrinsic part of construct validity. Stated differently, the underlying language and cognitive processing that takes place when performing language-related tasks in authentic contexts needs to be replicated during the assessment process. If the selection of examination content and criteria is to have validity, then the choice of tasks will be 'representative of the larger universe of tasks of which the test is assumed to be a sample', with due consideration to the 'linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed' (Weir, 2005: 19). Bachman and Palmer (1996) consider this alignment to form part of the notions of authenticity and interactiveness, both of which, together with construct validity, they include under the utility or 'usefulness' that is for them the prime consideration in language assessment. Accordingly, examination tasks should be representative of language tasks executed outside the examination room (i.e.

target language use) in which 'language ability (language knowledge plus metacognitive strategies), topical knowledge, and affective schemata' are drawn on to formulate a response (Bachman & Palmer, 1996: 39). This interactional understanding of a construct of language competence implies an ability to 'infer from test performance something about both practice-specific behaviour and a practice-independent, person-specific trait' (Young, 2012: 181).

Apart from the importance of conceptual clarity on the construct, a further essential condition for the validation process relates to the issue of scoring validity, more commonly referred to as reliability or the consistency of the measurement (Bachman & Palmer, 1996; Jones, 2012). Scores may be deemed to be reliable if they remain consistent from one set of tests and tasks to another. Reliability is thus a function of score consistency between different administrations of tests and tasks, as well as between the different subtests and tasks of the same assessment instrument. Although today reliability is generally considered as a necessary quality of a language test, it is not an aspect that on its own can provide sufficient evidence of the validity of the measuring instrument (Weir, 2005: 24; Weideman, 2009). There is also a distinction to be made between reliability resulting from valid marking procedures, and reliability resulting from the performance of test items and their ability to discriminate between candidates of differing proficiency. The latter form is usually determined statistically. Thus, in order to establish reliability in the case of the Grade 12 English Home Language examination, markers would need to be consistent in awarding the same marks for the same performance (intra-rater reliability). At the same time there also needs to be consistency between the marking of different markers (inter-rater reliability). Although rubrics are used to reduce subjective marking, these do not guarantee marking consistency since raters may not apply the rubrics uniformly. Statistical techniques such as Multifaceted Rasch (MFR) analysis may be useful in identifying inconsistent marking and whether statistical adjustment of marks is justified.

To date, no use has been made of statistical methods to determine the reliability of the examination items in the Grade 12 papers, and no subtest correlation data are available from the examining authorities either. This means that the current study has to rely on a theoretical framework to establish the potential reliability of the scoring in the absence of empirical data.

In summary, although the standard of the curriculum has been made explicit in CAPS, there is little certainty about the standard of the accompanying examination papers, hence the need for a validation study. In one of the few studies commissioned by Umalusi so far, 'The standards of the National Senior Certificate Home Language Examinations: A comparison of South African official languages' (Umalusi, 2012a), an attempt was made to categorise each examination item according to its level of cognitive demand on the basis of an adaptation of Barrett's taxonomy (see Umalusi, 2012a: 43-48). When the taxonomy proved to be problematic for this purpose, a follow-up study commissioned by Umalusi (2012b) was devoted to developing a framework for assessing cognitive challenge using a combination of revised formats of the taxonomies of Bloom and Barrett. However, this process merely served to confirm the difficulty and subjectivity of attempting to decide which questions were more cognitively challenging to learners without any empirical evidence in support thereof based on learners' responses. The same anomalies present in the first study simply resurfaced in the second, as the following example from the newly developed evaluative framework illustrates:

Level of cognitive demand	Type of cognitive demand	Explanation of categorisation. Questions which require students:	Examples
Higher order processes	**5. Synthesise or create**	To ***integrate*** ideas and information and relate parts of material, ideas, or information to one another and to an overall structure or purpose in a way that is ***relational***. To engage in ***original creative thought*** and design and put elements together to form a coherent whole and make a new or ***unique*** product ***showing emotional, aesthetic or literary sensitivity***	*You are selling a second-hand item (e.g. a Walkman, a CD player, an item of clothing). Create an advertisement which will be placed on the notice board at school.* *Write an essay of between 250 and 300 words titled 'As I looked at that photograph…'*

Table 1: Extract from the Umalusi typological framework for determining cognitive demand of examination items (Umalusi, 2012b: 86)

It is highly disputable that the mentioned example tasks would require cognitive challenge of the highest order. Rather than being asked to carry out such basic communicative tasks, examinees could be challenged to display their creativity and originality in argumentative types of tasks that would require reasoning and problem solving ability.

A more systematic and sophisticated approach is needed to determine the level of ability required, and preferably one that is premised on internationally accepted principles for the validation of high-stakes language examinations. In view of the fact that it is not within the ambit of the current study to undertake a full validation of the language papers, two key aspects of validity will be considered as the first of a number of necessary steps in building a validity argument, namely construct validity and scoring validity, also referred to as reliability.

METHODOLOGY

Since the Grade 12 language examination is not only an assessment of language proficiency, but the measurement of mastery of the content of the curriculum that defines the high-level proficiency required by the syllabus, conceptual clarity was in the first instance gained on the underlying construct of the examination as encapsulated in the aims and principles of the curriculum and learning programme. Secondly, the construct was articulated in a number of specifications of ability deriving from the content analysis of the curriculum. Thereafter, a detailed content analysis of each examination paper and its accompanying marking memorandum was made to ascertain the degree of alignment with the curriculum, as well as the desirability of using the existing task types as a valid and reliable measure of high language ability.[ii]

It should be mentioned that the examination papers scrutinised in this paper were set using the National Curriculum Statements. A cursory comparison of the NCS and newly introduced CAPS, however, revealed that the new curriculum was an abridged and more user-friendly version of its predecessor, and that the objectives were essentially the same. A decision was then taken to examine the extent to which the English Home Language papers were aligned

with the new CAPS document, and whether they could be retained in their current format once the new curriculum had been rolled out in full in 2014.

To execute the above objectives, the language component of the Grade 12 November English Home Language examination (Paper 1)[iii] was analysed in detail, covering the period 2008-2012. The five-year period was chosen as this goes back to the year when a common national examination was set for all learners as part of a new educational dispensation. Each question in each respective section was scrutinised from the perspective of a potential examinee and possible responses compared to those elucidated in the marking memoranda. The purpose here was to determine the clarity of the questions, completeness of the memoranda, any discrepancies between the question papers and memoranda, and the desirability of the set tasks.

In order to undertake the analysis, each of the language-related abilities specified in CAPS was allocated a code. A limitation of the above method is that more than one classification per examination item was possible. An attempt was thus made to identify the main purpose of each, using the suggested answers contained in the memoranda as a guide. The purpose of the code was to determine the spread of curriculum content and sub-abilities covered.

Marking was designated as subjective where a personal viewpoint requiring subjective evaluation was to be expressed. In those cases where there was a definite correct answer to be provided and no interpretation of responses required, marking was considered objective. What should be kept in mind, however, is that even where questions are not to be marked globally, the memoranda stipulate that the marker is given the right to consider other responses and that the memoranda are intended to be used as a guide and not prescriptively. This naturally opens the door for further subjective marking.

Each examination item was also categorised according to the type of response required, i.e. closed-ended or open-ended. The latter type requires learners to construct their own responses, which may lead to subjective marking. Closed-ended questions generally have higher reliability since there is a definite correct answer, but these are limited in their ability to test productive ability (cf. Alte, 2005: 111-112).

MAIN FINDINGS

Section A of Paper 1

The text comprehension part of the paper (Section A) generally covers the reading of two texts, one of which may be a visual text. Of particular concern in this section is the nature of the texts selected and their difficulty. Passages in the 2008 and 2009 papers were ridiculously easy with low Flesch-Kincaid grade levels of 6.8 and 7.[iv] It is encouraging to note the tendency since 2010 to select more advanced texts, although some of the content remains troublesome (see Addendum A for the selection and grade levels). The following excerpt from the November 2012 paper (Department of Basic Education, 2013a) serves as an example:

Text passage: This April, South Africans were able to reflect on the past 18 years since we took that giant step towards becoming a country that can boast one of the most democratic constitutions in the world. Theatre in South Africa has always been a dynamic forum that has given us the

courage to grapple with the state of the nation. Our writers, stand-up comedians, satirists and community-based artists have used their remarkable talents to create and nurture a climate that has allowed us all to become active participants in our democracy.

Question 1.1: Why is theatre considered 'a dynamic forum'? (2 marks)

Memorandum: Theatre is considered a 'dynamic forum' as it has nurtured a climate of democracy. Those involved in the theatre have encouraged us to become participants in this democracy.
[if a candidate explains the concept of 'dynamic forum', award 2 marks.]
[if a candidate lifts directly from the passage, do not award more than 1 mark.]

It is obvious that the question is formulated in too general terms. Some learners may give their own opinions, while others may simply quote a phrase from the text. There may be a number of unanticipated responses different to those contained in the memorandum and the possibility exists that an acceptable answer may be scored as incorrect by markers who adhere strictly to the memorandum. The text passage is poorly written, with a particular lack of coherence and cohesion (compare the first two sentences), and contains more than one sweeping statement. Moreover, the example question anticipates a connection between the new South Africa and theatre, without the text itself providing any coherent link. It is further noticeable that the text expresses a number of opinions, yet no questions are included to assess whether learners can distinguish between fact and opinion. It is also problematic that full marks can be allocated where candidates explain what the words 'dynamic forum' mean (e.g. a lively platform), without placing this in the context of drama and theatre. The selection of a text with a strong political theme is another contentious point and something that should best be avoided, especially in the light of South Africa's history where education was used as a tool to further the interests of the apartheid regime.

The directive issued by the Department of Basic Education in 2012 (Circular E 2, Department of Basic Education, 2012b) to include longer reading passages in future papers and a related visual text is to be welcomed. More cognitive processing is involved with longer texts and the greater number of examination items that can be set facilitates generalisation of ability to other domains requiring reading. The use of short and undemanding texts compromises theory-based validity (Weir, 2005: 74).

In addition, scoring validity in Section A of the examination is troublesome. Most of the items in the comprehension section count more than one mark (see table 2), but hardly any indication is given to the candidates of how marks will be earned.

Number of questions counting 1 mark	1
Number of questions counting 2 marks	31
Number of questions counting 3 marks	25
Number of questions counting 4 marks	3
% of questions counting more than 1 mark	**98%**

Table 2: Mark distribution in Section A of Paper 1 (2008-2012)

The lack of specification on how marks will be allocated per item is unfair to examinees. Moreover, the analysis reveals that on average up to 70% of the items could potentially be scored subjectively. Many of these items require the expression of an opinion (on average 56%), making the performance of this ability a characteristic feature of section A. The preponderance of questions such as 'Do you agree …', 'In your opinion ...', and 'Suggest

why ...', explains the high percentage of subjective marking involved, which would impact negatively on the reliability of scores.

Another striking feature is the lack of closed-ended questions. Only one of the 60 questions over the five-year period required a single word for a response. When examinees have to respond to an item by writing a few sentences to earn between one and four marks, and no clear indication is given of how marks will be allocated, the language testing principles of validity and reliability are jeopardised.

Section B of Paper 1

In this part of the examination, examinees are required to summarise a text passage. Although a summary type task assesses reading skills, it is usually considered a writing task (Hughes, 2003; Weigle, 2002; Weir, 2005; Yu, 2013) in which learners have to show their skill in understanding the content, distinguishing between essential and non-essential or supportive information, and their ability to manipulate language by condensing the essence of the message in a coherently written paragraph. In the HL papers, the summary section was originally designed as a combined reading and writing task, but from 2010 onwards examinees have no longer been obliged to construct a paragraph. They may simply list a number of points. Although there is a stipulation that 'sentences and/or sentence fragments must be coherent' (cf. English Home Language Memorandum Paper 1, 2012, p. 5; Department of Basic Education, 2013a), this is likely to undermine the validity of the task, considering that coherence may be difficult to establish on the basis of fragmented points. Another point of criticism is that only perfunctory language ability is needed to produce sentence fragments, as compared to the skill it takes to produce a well-organised paragraph. Further to this, allowing more than one response format makes it impossible to compare responses and award marks equitably. Phrases lifted from the text passages cannot be marked in the same way as a coherently constructed paragraph: they simply do not provide evidence of the same sub-skill.

The table that follows shows the selection of texts used for summary writing over the five-year period 2008-2012.

Year	Topic	Word count	Readability		Sub-ability[v]	Marking
			Flesch reading ease	Flesch-Kincaid grade level		
2008	Books and reading	359	54.2	10.1	44, 69, 88	Subjective
2009	2010 Soccer World Cup	330	58.4	9.4	44, 69, 88	Subjective
2010	Children's rights and freedom of action	349	62.7	8.2	44, 69	Subjective
2011	Power of positive thinking	347	57.5	8.7	44, 69	Subjective
2012	The meaning of face	370	49.1	11.4	44, 69	Subjective

Table 3: Overview of summary task in Section B

The summary tasks are mostly of a general nature with little indication being given to the examinees as to how marks will be allocated. As pointed out earlier, the memoranda specify that marking is global, and thus may potentially be subjective. What is particularly worrying, however, is the absence of penalties for lifting phrases from the text or exceeding the required length. If high language ability is to be measured, this type of summary writing cannot be used. To compound matters, the text passages are already short, and the shorter the passage, the more difficult it potentially becomes to provide a meaningful summary.

More careful consideration should, therefore, be given to the nature and lengths of the texts selected for summarising, and the response format. In its current format Section B of the paper serves very little purpose and cannot be considered reliable. It lacks theory-based and content validity, serving neither as a test of reading comprehension nor of writing ability. The ability to 'comprehend reading matter and organise their thoughts in writing' (Hill, 1991) is nonetheless one of the academic skills that can benefit students in all fields of study. When designed well, a summary task should display high construct validity and authenticity, and the inferred ability should be generalisable to other domains.

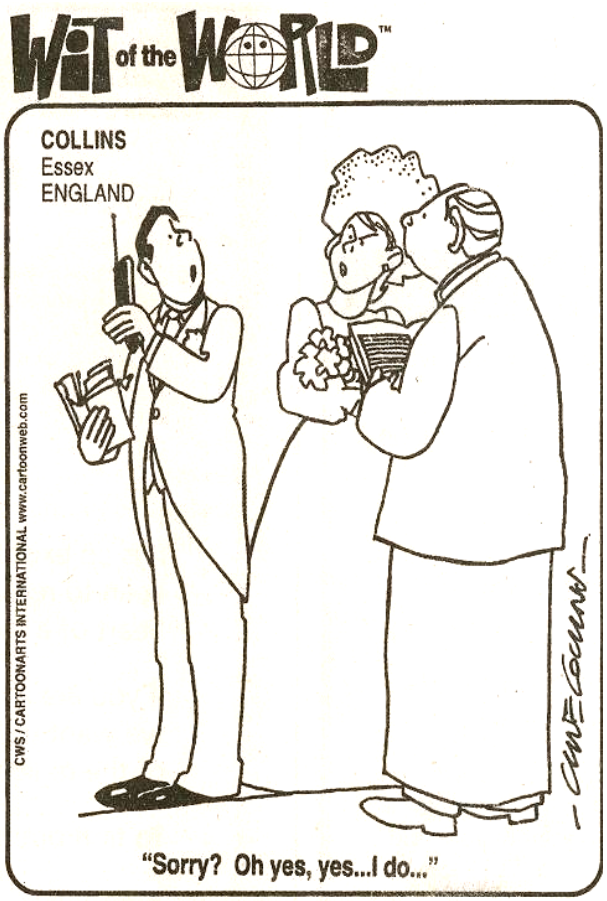
Part of the problem with the formulation of the summary task in the Grade 12 papers may relate to divergent conceptualisations of summary writing. Yu (2013) finds incongruities in the way summary writing is operationalised and assessed. In view of the fact that student responses to and interpretations of summary writing are influenced by their previous experiences and assumptions, he considers summary writing not to be a uniform ability or unitary process, but a multidimensional and unique kind of writing that requires integrated language ability (reading, analysing, condensing and restructuring in writing), making it a genre of its own. What is more, summary writing in one kind of discourse, and for one kind of audience, may be different from that required of the genre in others: the typical differences among various discourse types may occasion specific variations of how the concept of summarising is understood in each. These concerns illustrate the need for a common understanding of the task on the part of the examiners, markers and examinees – an aspect that appears to be missing at the moment.

Section C of Paper 1

In this third section of the paper, the focus falls to a large extent on the analysis of advertisements and cartoons, in addition to language use and text editing. The inclusion of so many visual texts in this part of the paper is based on an erroneous assumption in the curriculum statements that 'for many learners, the screen rather than the printed page is the source of most of their information' (Department of Basic Education, 2011: 23). Such a supposition fails to recognise the difficulty of many learners to access online information, and that, irrespective of the mode of communication, the written verbal text still provides more information to learners than any non-verbal slanting frames, fonts or photographs, some of the typographical and other components of texts favoured by examiners in this respect.

The study of the meaning of visual signs, which forms part of the curriculum, derives largely from the work of Ferdinand de Saussure (1974) who noted the distinction to be made between the visual image as the signifier and the concept it represented as the signified (cf. Culler, 1986: 8). The problem, as Berger (1999: 71) points out, is that the 'relationship between a signifier and signified is arbitrary, and therefore always open to question'. Any examination items pertaining to visual elements such as graphics, fonts, frame sizes and body language, will be open to any number of different interpretations, an aspect that can

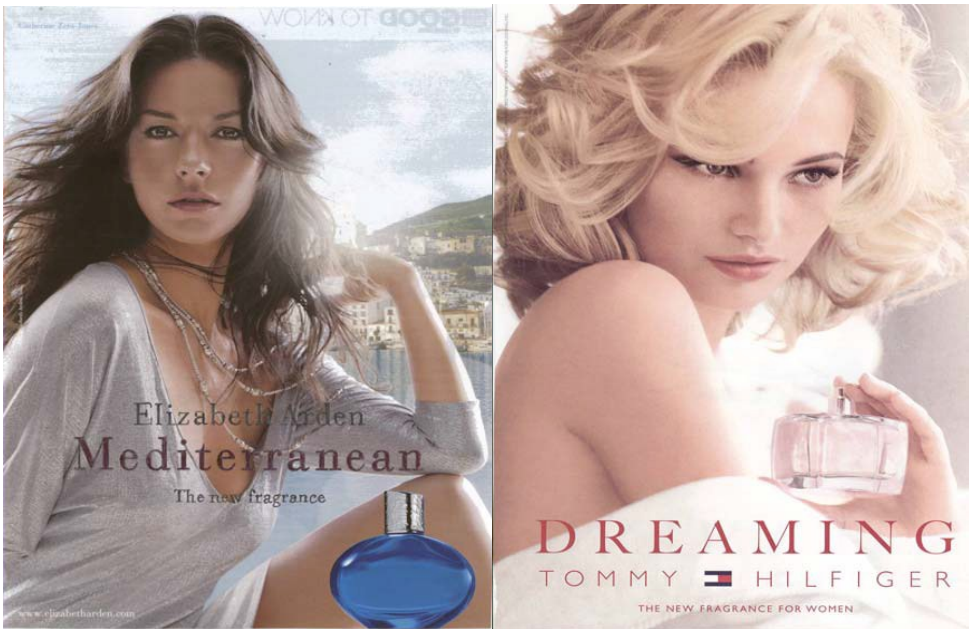
complicate scoring validity. There is a further complexity too. The reading of visuals such as cartoons may depend on cultural and extraneous knowledge, making the interpretation of images a potentially unfair construct in an examination of English language ability. The following example of a culturally biased item is the cartoon that forms part of question 4 of the November 2010 English HL paper (Department of Basic Education, 2013a):



Question: 4.1 Discuss what the cartoonist is satirising (2 marks).

Memorandum: The cartoonist satirises people's obsession with modern gadgets such as cell phones. This obsession interferes with traditional/normal considerations. People cannot be separated from their cell phones, even for something as important as their own wedding.

Cultural bias is evident in the setting depicted. Examinees would need to be familiar with Western religious wedding customs to understand the cartoon. There are also items with gender bias. The following two perfume advertisements in the November 2008 paper (Department of Basic Education, 2013a) serve as illustration.



Learners are expected to comment on the choice of models and words in the visuals. To make the necessary associations, extraneous knowledge of the Mediterranean is required in the case of the first advertisement.

Question: Discuss how the advertisers of the perfumes in Texts E and F appeal to their respective target markets, with reference to the following:
3.2.1 Words in the advertisements (2 marks)
3.2.2 Choice of models (2 marks)

Memorandum:

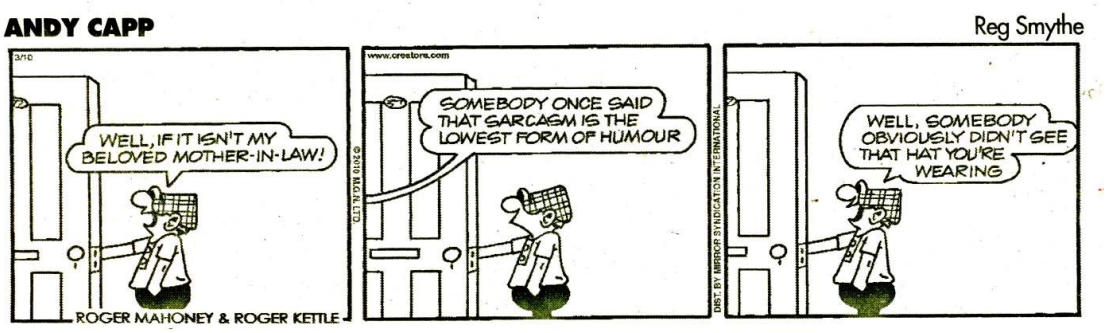
3.2.1 *Dreaming* – gives consumer idea of world of fantasy / bedroom.
New fragrance – something new on the market.
Name of famous designer – lends status / implies wealth.
Mediterranean – outdoor / exotic / historical.
Candidate may mention any of the above – one for each advertisement.
(1 mark per advertisement; credit discussion of *New Fragrance* if repeated for each.)

3.2.2 *Dreaming* – model appears romantic / coy / alluring / feminine / soft and gentle. *Mediterranean* – model appears assertive / confident / forceful / forthright / challenging. (Any description of beauty/sex appeal, e.g. both models are beautiful/sexy, award full marks.)

The suggested answers in the memorandum are all of a subjective nature and open to debate, but fortunately examinees can mention that the words 'new fragrance' in each advertisement refer to a 'new' product and earn their marks. Such examination items do not provide any evidence of language ability or critical language awareness, apart from being biased towards a certain conception of femininity.

Perhaps the strongest reason for excluding visuals such as photographs and advertisements such as the above in a language examination relates to the irrelevance of analysing such images in real-life contexts. Many of the tasks in Section C of Paper 1 lack content validity and violate the principles of authenticity relating to target language use.

The nature of the marking is once again potentially unreliable in this section of the examination, with the number of items requiring subjective marking surpassing those that can be marked objectively. There is also an over-representation of the assessment of the ability to express an opinion (24 of the 88 questions analysed in this section, in other words more than a quarter), and an abundance of open-ended item types. The problem seems to be related to both the selection of visual texts and kinds of items set, as evident in the following example extracted from the November 2011 paper (Department of Basic Education, 2013a).



Question 4.2.2 The cartoonist does not show the mother-in-law in any of the frames. Do you think that this is an effective technique? Motivate your response. (2 marks)

Memorandum: Yes. The reader can supply his/her own idea of a hideous hat: this is more effective than drawing one./ The big gap in the relationship between Andy and the woman is suggested by her being out of the frames.
OR
No. I think it would have been very effective if the cartoonist had shown the mother-in-law wearing a hideous hat.
[Consider and credit other valid responses.]

Markers have the prerogative to accept or reject the responses of the learners, or to give all learners full marks, considering that there are no definite answers: From the memorandum we can see that the same marks are allocated for responses that reflect inferential and higher order thinking (e.g. the nature of the relationship between Andy and his mother-in-law), and those that merely require an opinion such as 'No, the hat and mother-in-law should have been shown'. The effect hereof is that the item does not discriminate between learners of differing ability. The above example is typical of the kinds of questions and items included in this section of the examination papers over the past five years.

On the positive side, the last task type in Section C, which covers language use and text editing (question 5), is both representative of the curriculum content and contains very few discrepancies between items and the marking memoranda. The task requires examinees to identify incorrect language use, explain the use of punctuation and to display knowledge of grammatical structure, *inter alia*. The absence of subjective marking also makes this section of the paper potentially more reliable than the other sections.

To summarise, the main factors impacting negatively on the validity of the examination papers, and the generalisability of the scores obtained to non-examination domains, relate to the following: deficiencies in the memoranda and an extremely high percentage of marking subjectivity, the over-inclusion of open-ended and constructed response items, the over-representation of questions related to the expression of an opinion, and the lack of indication

to examinees as to how marks will be earned. Table 4 reflects the core findings relating to sections A and C over the five-year period of review.

		2008	2009	2010	2011	2012	Average
% marks potentially subjective	Section A	60%	73%	53%	73%	93%	70%
	Section C	60%	47%	40%	33%	37%	43%
% of open-ended items included	Section A	100%	92%	100%	100%	100%	98%
	Section C	72%	61%	67%	50%	50%	60%
% of items requiring an opinion	Section A	50%	42%	42%	62%	82%	56%
	Section C	33%	33%	27%	22%	17%	26%

Table 4: Leading factors contributing to unreliability of scoring in sections A and C

If we add Section B, the summary writing task, which also requires global and subjective scoring, we are left with a set of highly unreliable examination papers that fall far short of providing generalisable evidence of language ability.

CONCLUSION

The English Home Language examination plays an important role in the South African education dispensation. Access to university is granted on the basis of the matriculation results, and school-leavers stand a better chance of gaining employment in the economic sector if their language skills are good (Solidarity Research Institute, 2012). It is thus desirable that the ability inferred on the basis of examination scores be generalisable to post-matriculation settings. In order to achieve the latter, the principles of validity and reliability need to be applied during the respective design phases and administration of the examination papers. This could substantially enhance the credibility of the examination results as indicators of high educational standards.

On the basis of the findings of the content analysis of both the curriculum and examination papers, no conclusion can be reached that the selection of English Home Language papers evaluated can be considered a valid or reliable assessment of high language ability, and the credibility of the examination results remains questionable. The main reasons for this appear to be related to the underrepresentation of the construct on which the examination is premised, and a problematic system of scoring.

Although the papers as a whole may, if one employs a very lenient set of criteria, be pronounced representative of the language curriculum, they fail to meet the objectives of the National Curriculum Statements and CAPS to provide evidence of *high language ability*, the supposedly distinguishing feature of the HL papers. Rather than being indicators of advanced and critical language ability, the examination papers instead provide (mostly subjective, but certainly unreliable) evidence of basic language ability. Examinees are afforded little opportunity to display a command of language across different discourse types and the ability to handle typical features of discourse, as required by the curricula, even in the summary

writing section. No distinction is apparent between the language situation and the conditions for using language in that situation.

In as much as the curriculum advocates a communicative approach to language and the integration of skills, the examination papers still compartmentalise these. Employing a skills-neutral view of language could create a more authentic context and allow examinees to navigate between different discourse types and tasks while displaying their integrated ability to read, analyse, evaluate and write about what they are reading and thinking. Consideration should, therefore, be given to revising the entire format of the examination papers to reflect current views that advocate a more holistic and integrated approach to language testing through the employment of multi-skill constructs and multi-dimensional tasks. This could allow greater variation of task types and the performance of differential ability.

Further to the above, the construction of the memoranda and basis on which marks are allocated are in need of urgent revision. It is equally essential that examinees understand how marks will be earned, and that markers award marks consistently and equitably. The analysis of question papers and memoranda reveals that even challenging questions can be undermined when answers that show little insight are accepted. In this respect, the memoranda need to stipulate very clearly what constitutes an acceptable answer and how marks will be awarded. Another contentious issue relates to the disparities that exist in the qualifications and capabilities of the markers (Umalusi, 2012c: 318), and the complexities of overcoming rater bias. There is a pertinent need to counter the number of open-ended items with item types that can be scored objectively, such as multiple-choice questions. The inclusion of closed-ended items to supplement the constructed-response items could also create an opportunity for computerised marking in some sections. This would enable the generation of reliability statistics and facility values that could clarify the difficulty of items, and at the same time ensure some consistency in the marking.

In conclusion, the importance of the format and design of the examination papers in maintaining educational standards is clear from the findings of the study. On the matter of whether the English Home Language subject is being offered at a suitably high level, there is less certainty. Here there is a need for benchmarking with other multilingual countries to compare curricula and assessment systems at first language level.

REFERENCES

ALTE (Association of Language Testing in Europe). 2005. Materials for the guidance of test item writers. Language policy division, Council of Europe. Available at http://www.alte.org/attachments/files/item_writer_guidelines.pdf

BACHMAN, LF & AS PALMER. 1996. *Language testing in practice*. Oxford: Oxford University Press.

BERGER, AA. 1999. *Signs in contemporary culture: An introduction to semiotics.* 2nd edition. Salem, Wisconsin: Sheffield Publishing Company.

BLANTON, LL. 1994. Discourse, artefacts and the Ozarks: Understanding academic literacy. *Journal of second language writing*, 3(1): 1-16.

CHAPELLE, CA. 2012. Conceptions of validity. In Fulcher, G & F Davidson (Eds), *The Routledge handbook of language testing.* New York: Routledge. 21-33.

CULLER, J. 1986. *Ferdinand de Saussure.* Ithaca, New York: Cornell University press.

CUMMINS, J. & C DAVISON (Eds). 2007. *International handbook of English language teaching Part 1*. New York: Springer Science+Business Media, LLC.

DAVIES, A, A BROWN, C ELDER, K HILL, T LUMLEY & T MCNAMARA. 1999. *Studies in language testing: Dictionary of language testing.* Cambridge: Cambridge University Press.

DEPARTMENT OF BASIC EDUCATION. 2011. Curriculum and assessment policy statement: Grades 10-12 English Home Language. Pretoria: Department of Basic Education.

DEPARTMENT OF BASIC EDUCATION. 2012a. National Senior Certificate Examination Technical Report 2012. Pretoria: Department of Basic Education.

DEPARTMENT OF BASIC EDUCATION. 2012b. Circular E 2 of 2012. Amendments to the examination guidelines for official Home Languages (HL); First Additional Languages (FAL) and History: National Senior Certificate (NSC). Pretoria: Department of Basic Education.

DEPARTMENT OF BASIC EDUCATION. 2013a. Past exam papers. [online](Available from: http://www.education.gov.za/Examinations/PastExamPapers/tabid/351/Default.aspx). Accessed 2013-12-15.

DEPARTMENT OF BASIC EDUCATION. 2013b. *National policy pertaining to the programme and promotion requirements of the National Curriculum Statement Grades R-12*. Available at : http://www.education.gov.za/LinkClick.aspx?fileticket=Rcf0UfEfk5s%3D&...]

DE SAUSSURE, F. 1974. *Course in general linguistics.* Glasgow: Fontana.

DU PLESSIS, C, S STEYN & A WEIDEMAN. 2013. Towards a construct for assessing high level language ability in Grade 12. Unpublished report submitted to the Umalusi Research Forum, Pretoria, 12 March 2013.

HILL, M. 1991. Writing summaries promotes thinking and learning across the curriculum: But why are they so difficult to write? *Journal of Reading*, 34(7):536-539.

HUGHES, A. 2003. *Testing for language teachers*. 2nd edition. Cambridge: Cambridge University Press.

JONES, N. 2012. Reliability and dependability. In Fulcher, G & F Davidson (Eds), *The Routledge handbook of language testing.* New York: Routledge. 350-362.

KANE, M. 2004. Certification testing as an illustration of argument-based validation. *Measurement*, 2(3):135-170.

MESSICK, S. 1995. Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50:741-749.

PATTERSON, R. & A WEIDEMAN. 2013. The typicality of academic discourse and its relevance for constructs of academic literacy. *SAALT Journal for Language Teaching*, 47 (1): 107-123.

READ, J. 2010. Researching language testing and assessment. In Paltridge, B & A Phakiti (Eds), *Continuum companion to research methods in applied linguistics.* London: Continuum International Publishing Group. 286-300.

SOLIDARITY RESEARCH INSTITUTE. 2012. *The South African labour market and matriculants' prospects*. Available at: http://www.moneyweb.co.za/mw/action/media/download File?media_fileid=17832

UMALUSI (Council for Quality Assurance in General and Further Education and Training). [S.a.] Available at: http://umalusi.org.za.html

UMALUSI (Council for Quality Assurance in General and Further Education and Training). 2009a. *2008 Maintaining standards report part 1: Overview.* Pretoria: Umalusi.

UMALUSI (Council for Quality Assurance in General and Further Education and Training). 2009b. *2008 Maintaining standards report part 2: Curriculum evaluation.* Pretoria: Umalusi.

UMALUSI (Council for Quality Assurance in General and Further Education and Training). 2009c. *2008 Maintaining standards report part 3: Exam paper analysis.* Pretoria: Umalusi.

UMALUSI (Council for Quality Assurance in General and Further Education and Training). 2010. *Evaluating the South African National Senior Certificate in relation to selected international qualifications: A self-referencing exercise to determine the standing of the NSC.* Joint research project undertaken by Umalusi and Higher Education South Africa (HESA). Pretoria: Umalusi.

UMALUSI (Council for Quality Assurance in General and Further Education and Training). 2011. *All the cattle in the kraal: An overview of Umalusi's research 2003-2011.* Research report. Pretoria: Umalusi.

UMALUSI (Council for Quality Assurance in General and Further Education and Training). 2012a. *The standards of the National Senior Certificate Home Language examinations: A comparison of South African official languages.* Pretoria: Umalusi.

UMALUSI (Council for Quality Assurance in General and Further Education and Training). 2012b. *Developing a Framework for assessing and comparing the cognitive challenge of Home Language examinations*. Pretoria: Umalusi.

UMALUSI (Council for Quality Assurance in General and Further Education and Training). 2012c. *Technical report on the quality assurance of the examinations and assessment of the National Senior Certificate (NSC) 2012.* Pretoria: Umalusi.

VAN DER WALT, JL. 2012. The meaning and uses of language test scores: An argument-based approach to validation. SAALT *Journal for language teaching*, 46(2):141-155.

VAN DYK, T & A WEIDEMAN. 2004a. Switching constructs: On the selection of an appropriate blueprint for academic literacy assessment. SAALT *Journal for language teaching*, 38(1):1-13.

VAN DYK, T & A WEIDEMAN. 2004b. Finding the right measure: From blueprint to specification to item type. SAALT *Journal for language teaching*, 38(1): 15-24.

WEIDEMAN, A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African linguistics and applied language studies*, 27(3):235-251.

WEIDEMAN, A. 2011. Academic literacy tests: Design, development, piloting and refinement. SAALT *Journal for Language* Teaching, 45(2):100-113.

WEIDEMAN, A. 2012. Validation and validity beyond Messick. *Per linguam* 28(2):1-14.

WEIGLE, SC. 2002. *Assessing writing.* Cambridge: Cambridge University Press.

WEIR, CJ. 2005. *Language testing and validation: An evidence-based approach.* New York: Palgrave Macmillan.

YOUNG, RF. 2012. Social dimensions of language testing. In Fulcher, G & F Davidson (Eds), *The Routledge handbook of language testing.* New York: Routledge. 178-193.

YU, G. 2013. The use of summarization tasks: Some lexical and conceptual analyses. *Language Assessment Quarterly*, 10:96-109.

BIOGRAPHICAL NOTE

Colleen du Plessis is a lecturer in the Department of English at the University of the Free State. Her fields of interest include language testing, academic literacy development and applied language studies in education.
Email address: duplessiscl@ufs.ac.za

ADDENDUM A

2008	**Text A: History in the making**	
	Discourse field: Politics (1994 elections)	
	Number of words	320
	Flesch Reading Ease	68.5
	Grade level	6.8
	Text B: Untitled	
	Discourse field: Politics (new South African identity)	
	Number of words	358
	Flesch Reading Ease	47.4
	Grade level	12.5
2009	**Text A: The games that bring us together**	
	Discourse field: Social (sport and games)	
	Number of words	432
	Flesch Reading Ease	75.1
	Grade level	7.0
	Text B: Youth sport for a healthy nation	
	Discourse field: Social (sport and health)	
	Number of words	131
	Flesch Reading Ease	49.9
	Grade level	10.7
2010	**Text A: Comic strips and cartoons**	
	Discourse field: Academic (education)	
	Number of words	344
	Flesch Reading Ease	47.9
	Grade level	11.3
	Text B: Nelson Mandela comic book launched	
	Discourse field: Politics (establishing a democracy)	
	Number of words	419
	Flesch Reading Ease	41.1
	Grade level	12.4
	Text C: Untitled cartoon	
	Discourse field: Social (birthday wishes)	
	No readability statistics available (too little text)	
2011	**Text A: Untitled**	
	Discourse field: Politics (unity through sport)	
	Number of words	715
	Flesch Reading Ease	43.5
	Grade level	12.7
	Text B: Invictus (film poster)	
	Discourse field: Politics (unity through sport)	
	No readability statistics available (too little text)	
2012	**Text A: The arts celebrate and inspire our democracy**	
	Discourse field: Politics (using arts to establish a democracy)	
	Number of words	833
	Flesch Reading Ease	50.8
	Grade level	12.0
	Text B: R150m Soweto Theatre packs entertainment punch (advertisement)	
	Discourse field: Social (entertainment)	
	No readability statistics available (too little text)	
Average Flesch Reading Ease		**53.03**
Average grade level		**10.68**

Readability statistics and themes of comprehension texts in Section A of Paper 1 (2008-2012)

[i] Assessment tasks completed during the course of the year as part of continuous assessment contribute 25% towards the final mark. However, oral assessment tasks also carried out during the normal school programme contribute a further 12.5%, making the total contribution 37.5% (CAPS: 75).

[ii] Past examination papers and memoranda are easily obtainable from www.education.gov.za.

[iii] Paper 2, which covers Literature, was not included in the analysis, as the construct of this examination differs from that of Paper 1. Similarly, Paper 3, which assesses writing, will be reported on in a separate study.

[iv] An objective indication of the readability of texts can be calculated through programmes such as Flesch Reading Ease and Flesch-Kincaid Grade Level (available in Microsoft Word) that calculate the average length of sentences and number of syllables in each word.

[v] Sub-ability 88 of the coding system used requires of learners the ability to write texts that are coherent using conjunctions and transitional words and phrases. Sub-abilities 44 and 69 relate to the ability to make notes and summarise the main and supporting ideas respectively.