Hyphenated validity: Comparability of assessment protocols in English language between two examination boards

Muchativugwa Liberty Hove
South Africa, North-West University, Mafikeng Campus
muchativugwahv@gmail.com

Abstract

This paper reports the discrepancies in assessment protocols between Cambridge International Examinations and the South African Umalusi in the 2010 exit English First Language examinations. The longitudinal study commenced in 2007 when twenty purposively selected learners were sponsored by a South African benefactor and enrolled at a private school that offered them a CIE-oriented curriculum in place of the OBE one that they would have pursued in government schools. Based on their entry competencies in English, the learners participated in the design and implementation of a task-based syllabus out of which suitable teaching and learning materials were developed. Through recursive formative, summative and other internal assessment strategies, in particular the quantitative measure called the hypotaxis index, the learners' performance in the final CIE examinations was relatively reliably predicted. The predicted grades, CIE final assessment papers, grades and marks are presented in this study and compared to the final grading scale(s) used by Umalusi in order to demonstrate the underlying ideological patents and paradigmatic shifts from one examining board to the other. Other validity and reliability issues are also analysed to highlight the (in) comparability of assessment protocols between examination boards.

Keywords

Validity, reliability, validity-as-language; validity-as-culture; hypotaxis index; task-based syllabus; standards comparability

Summary

Validity and reliability issues are context-specific and largely remain contentious in the light of (in) comparable benchmarks across examining boards.

Introduction

Exit examinations are generally understood to be reliable measurement instruments whose principal objectives are to screen for purposes of entry into higher education studies and provide selection criteria for purposes of employment. Based on these constructs of validity and reliability, exit examination grades and certificates must have, *ipso facto*, recognized currency and, by extension, the grades should be comparable across examining boards (Griesel, 2003). This paper argues that the predictive validity of examination grades, especially the comparability of these grades across examination boards, remains a subjective element and a promise, and is not an empirical fact, especially in English as a First Language proficiency assessment. Taking cue from the various reports from higher educational institutions in South Africa in particular (Hendricks, 2006; Jansen, 2011), it is apparent that many learners are leaving schools with certification that is problematic and difficult to benchmark. Again, the premise in this study is that test items that constitute the various English as First Language examination papers should be of real and sufficient complexity at the appropriate grade-level to function as useful empirical indicators of the cognitive abilities and proficiency levels of the candidates who pass such examinations.

Background

The Telkom Foundation (TF) sponsored twenty learners from previously disadvantaged primary schools in Limpopo and the Western Cape provinces in South Africa to enrol at a private school, the International School of South Africa (ISSA) located in the North-West Province, in 2007. There was a subsequent group of 32 learners from the same poor state schools who were also sponsored in 2008 to enrol at ISSA. These learners were purposively selected for a longitudinal case study on cognitive academic language proficiency development and, in tandem, the design and implementation of a relevant and efficient syllabus that aimed to respond to their linguistic challenges and educational needs. Amongst others, one of the greatest needs was "to succeed against all odds" in their CIE exit examinations. As part of recursive formative and summative school-based assessment procedures, the learners' potential to succeed in the Cambridge International Examinations (CIE) was systematically gauged and profiled as "predicted grades" in March 2010. Subsequently, these learners wrote the CIE exit examinations in November 2010 and their grades in these final examinations were compared to those that had been predicted in order to establish degrees of divergence and convergence at the research site. This comparative analysis was also perceived as a post-test evaluation of the efficacy of the task-based syllabus that had been developed *in situ* to facilitate their acquisition and development of competency skills that would enable them to succeed in the CIE exit examinations. For the first time in the history of Cambridge International Examinations (CIE), the examinations board published both the grade symbols and the actual examination marks for each candidate per subject. Meanwhile, in South Africa, there was a public outcry against the Department of Education (DoE) that marks for subjects such as geography, accounts, mathematics and English had been "massaged" and adjusted upward by Umalusi, the quality assurance body, for the National School Certificate (NSC)- (Jansen, 2011). This study especially sought to establish the comparability of the grade threshold marks for English Language in particular since the cohort that participated in the longitudinal study would have written the NSC examinations in South Africa had they not been sponsored by the Telkom Foundation.

Literature review

Downing and Halaydna (2006) present a generic examination cycle as the process by which a "standard score" or "grade threshold standard" is established.

Table 1 shows this generic process.

Examination process	**Process attributes and description**
Overall plan and examination level	Systematic guidance for test development at the appropriate grade/exit level
Content definition	Cognitive domains and essential sources of content-related validity Specific delineation of test constructs
Test specifications	Operational definitions of content, framework for validity, and defensible sampling of content and cognitive domains
Test item development	Development of effective questions, defensible question formats, validity evidence related to evidence-based principles, training of examiners/item

	writers and effective item editing
Test design and test assembly	Test forms (essay, multiple choice, structured) and pre-testing/post-testing analyses
Test production	Validity issues concerned with quality control
Test administration	Validity issues concerned with standardization and timing
Passing scores	Establishing defensible passing scores, validity issues concerning cut/threshold scores, comparability of year-on-year standards, constancy of score-scales/deviation
Reporting test results	Quality control, meaningfulness and timeliness, challenges to test results
Test technical reports	Systematic and thorough examiners' reports, documentation of validity evidence, recommendations.

For both CIE and NSC examinations, assessing writing in English First Language is based on band scales. Band scales are holistic and depend largely on what Gannon (1985:61) calls impression marking. Hyland (2002), Larsen-Freeman (1978) and Wolf-Quintero (1998) all argue that analytical measures are more appropriate measurement scales to establish linguistic accuracy, syntactic and grammatical complexity in learners' writing. It must be conceded that the studies cited here are principally concerned with establishing the developmental index in learner language. The studies argue that as learners become more proficient users of the English language, they write more clearly, more accurately and that the texts they produce are more grammatically and lexically complex (Naves, 2006:4).

Notwithstanding the research tangent which these researchers take, the systematized and operational protocols in the two examination boards surveyed in this study remain holistic. What connects the practices of the examining boards to the research discussed above is the focus on accuracy, complexity and language range. Skehan and Foster (1997:22) suggest that accuracy is concerned with how well language is produced in relation to the rule system of the target language. For complexity and range, Skehan and Foster (1997: 97) submit that this competency construct entails the capacity to use and control more advanced language and that this capacity involves a greater willingness on the part of the writer-candidate to take risks through ambitious sentence structures and diction usages.

In light of the arguments here, it is logical to perceive English First Language examinations administered by CIE and DoE as assessment instruments that seek to empirically measure the candidates' linguistic proficiency as evidenced in their accuracy, complexity of structures and vocabulary range. In the Cognition Hypothesis, Schmidt (2001) argues that gradually increasing the cognitive demands of language tasks pushes learner-writers to greater accuracy and complexity. Skehan and Foster(1999), on the other hand, contend that the more cognitively demanding the language task is, the more likely it would be that learner-writers will attend to conveying meaning first and to linguistic complexity and accuracy last. Whereas Schmidt is reporting on the integration of information-processing and interactionist explanations of language task effects, his hypothesis is applicable to an exit examination design where the summative assessment tasks set are generally perceived as eliciting a range from low to higher cognitive language proficiency skills. Until more empirical evidence becomes available to bolster the claims of either Schmidt or those of Skehan and Foster, text structure and linguistic texture, it would appear, override text content in the English language examinations offered by both CIE and NSC.

In this study, the constructs of accuracy, sentence complexity and vocabulary range were measure by using a quantitative measure, the hypotaxis index. The hypotaxis index quantifies a writer's ability to coordinate and subordinate ideas in order to express more sophisticated ideas. In a stretch of written discourse, the index tallies the coordinated and subordinated sentence structures and calculates their percentage as a total of the rest of the sentences that makes up the text. The higher the percentage of error-free, accurate and complex, subordinated and well-coordinated sentences, the more successful and linguistically competent the writer of the text. Implicit in this hypotaxis index is the fact that the language task set for formative, summative and exit assessment should provide the candidate with a meaningful task that allows the candidate to express ideas to the pitch of their abilities.

Facets of test validity

According to Messnick (1989), there are two principal facets of validity, that is, the evidential basis and the consequential basis, and these two could be discussed under two categories, that is, test interpretation and test use. For the evidential basis of validity, the important segments are construct validity, relevance and utility. Consequential validity is hardly an issue with testing boards and agencies yet this is the most crucial implications that any high stakes examination has to live with. Consequential validity relates to the value implications of the test, and most significantly, the social consequences of these tests. These facets are summed up in Table (i).

Table (i): Facets of test validity

	Test interpretation	Test use
Evidential basis	Construct validity	Construct validity, relevance and utility
Consequential basis	Value implications	Social consequences

Research design and methodology

This study adopts a multi-method approach. Its qualitative dimension seeks to explore the quality and depth of the assessment instruments used by two examining boards, Cambridge International Examinations and the Department of Education’s matriculation examination papers in English as a First language. This approach follows an understanding of qualitative research as “a process of systematic enquiry into the meanings of the test constructs and the skills that the candidates are expected to exhibit” (Grafanki, 1996:329). The quantitative dimension provides the statistical and graphic evidence emerging from the investigation in order to highlight the discrepancies between the examining boards.

This study sought to monitor “comparability between different tests of different forms” of the same subject (Newton, 2007). Both English First Language examinations by CIE and DoE were of current interest to the researcher, especially the fact that the research participants were transitioning from one curriculum form to another. The premise in this study was that the CIE and DoE English First Language examination papers would be of “comparable difficulty since they were set to a curriculum framework and test specifications that were explicit in their respective syllabi” (Greaney & Kellaghan, 1996).

A needs-analysis survey was conducted to establish entry level competencies since the research participants were being weaned from one curriculum orientation (OBE) to a new one (CIE). A task-based English language syllabus was then collaboratively designed and implemented at ISSA, having taken into cognizance the language needs of the research participants (Hove, 2010). The implementation of the syllabus reflected two major considerations:

1. Specific threshold levels were defined for language learning objectives and also for teaching purposes. These were pitched at grade-appropriate levels and were meant to reflect the levels of the participants’ language competence.

2. A breakthrough threshold level was defined, in the sense that this competence level reflected closely the exit examination level towards which the research participants were studying.

This framework for language pedagogy sought to relate the language course to assessment benchmarks. It provided a meta-language to interrogate both language-learning objectives and language-assessment levels. A content analysis of test items from previous CIE examinations provided space for the creation of language teaching guidelines that were perceived as relevant for meeting the competency levels expected in the CIE English First Language examination papers. In tandem, formative and summative classroom assessment practices were guided by the comparability of learner-performance indicators to those anticipated in the exit examinations. A new challenge emerged pertaining to the OBE curriculum at this stage: C 2005 was undergoing change and therefore the matriculants were going to sit an entirely new examination in 2009. This examination did not have any comparative precedent, except the one that it was replacing in South Africa. There were, therefore, three elements that were considered critical for this comparability study:

1. The specification of the content and purpose of the examination, i.e. test specifications, item-writer guidelines, examiner training, and previous examination reports on standards;
2. Standardization and interpretation of the performance and competency levels, i.e. suitable standardized materials such as exemplar scripts to benchmark current performance against performance standards in comparably recent years; and
3. Empirical validation concerns, i.e. routine item-by-item comparability and test calibration. The 2010 question papers for both examining boards were analysed in terms of the constructs that each question tested and how these competency constructs compared between the test papers in terms of difficulty and the cognitive demands that the questions made on the candidates.

Results and discussion

In terms of item development and validity evidence related to adherence to evidence-based principles, the CIE English First Language paper has a higher face-validity compared to the DoE paper. The CIE paper has had the same format since 2005 while the DoE one has no historical precedent to compare with. Of course this observation does not overlook the curriculum and political imperatives in the South African educational ecology and the need to revise both curricula and examination protocols.

The 2010 examination paper for English Home Language from DoE displays a somewhat indefensible sampling of content domains when gauged against the prevailing syllabus specifications. The summary question, for instance, asks the candidates to *list* the points only, without asking and insisting on the linkages between these points:

> Your teacher has asked you to deliver a short talk to your classmates during the English oral period on how to take care of your takkies. Read the article below and summarise the main points for inclusion in your article.
>
> Instructions
>
> 1. List seven points in full sentences using approximately 70 words.
> 2. Number your sentences from 1 to 7.
> 3. Write only one point per line.
> 4. Use your own words as far as possible. [10 marks]

In listing, the cognitive demands are apparently lower than the cognitive demands of writing in continuous form and adhering to stylistic principles such as concision and cohesion, a test construct that was evident in the CIE question paper of the same year. The CIE summary question was set out as follows:

> Summarise (a) the evidence that the orchestra described in passage B is "really terrible" and (b) what Signor Allesandro thinks are the qualities of a great conductor, as described in Passage A. Use your own words as far as possible. You should write about one side in total…Up to fifteen marks will be available for the content of your answer, and up to five marks for the quality of your writing. [20marks]

One glaring difference is in the length of the reading passages: CIE asks the candidate to read two passages concurrently, each one of them approximately 90 lines, and extends this to test the candidate's ability to make the selection of summary points and link them in continuous writing, while the DoE task, in contrast, is set on a very trite passage that is only 17 lines long.

The CIE marking scheme for the summary question explicitly states what it seeks to test. For 15 marks, the question tests candidates' reading to be demonstrated in how they (as spelt out in the Reading Curriculum component):

a. Understand and collate explicit meanings
b. Understand, explain and collate implicit meanings and attitudes
c. Select, analyse and evaluate what is relevant to specific purposes.

As a higher-order skills question, the summary task seeks to screen candidates on their abilities to perform at the appropriate grade level and their ability to demonstrate the relevant skills as outlined in both the marking scheme and the curriculum objectives. This summary task recognizes the links between the reading and the writing skills of the candidates, hence for 5 marks, it rewards the candidate's ability to:

a. articulate experiences and express what is thought, felt and imagined
b. order and present facts, ideas and opinions
c. understand and use a range of appropriate vocabulary
d. use language and register appropriate to audience and context
e. make accurate and effective use of paragraphs, grammatical structures, sentences, punctuation and spelling.

Compared to the DoE summary task in the examination where only "ordering and presenting facts" in list form is assessed, the CIE summary task in the examination is understood as a more credible construct of the skills embedded in "summarizing." In addition, the CIE summary task asks the candidates to "use own words as far as possible" and the final response ought to be "one page" in length. This is a clearly more valid assessment task when compared to the mere "list(ing) of at least seven points" that was set as the DoE summary task.

For the comprehension test items, there is also evidence of lower order skills being tested in the DoE paper when compared to the CIE one. The first reading passage in the DoE paper, "There's a Hippo on My Stoep!" is very simple in terms of the reading levels associated with a school leaving examination such as this matriculation one. Notwithstanding the simplicity of the text, the questions set on it could be classified as lying on a continuum between simple recall, application and basic analysis. There is no evidence of questions at the higher analysis, evaluation and synthesis domains:

> 1.1. How does Jessica come to live with the Jouberts? (2)
> 1.2. Why does Jessica sleep on the stoep? (2)

> 1.7. State two points from the passage which show that the Jouberts now regard Jessica as their "child." (2)

In contrast, the CIE comprehension question for the same year asked candidates the following two questions, based on two significantly challenging reading passages:

> Question 1
> Immediately after the sequences that you have just read, Signor Allesandro gives a TV interview. The interviewer asks three questions: Some people say you are an eccentric man whose behaviour is odd at times. Are they right? Can you explain the unexpected happenings that took place at the beginning of your Beethoven concert? Do you think that the time has come for you to retire from conducting? Write the words of the interview.
> Base your answer on what you have read in Passage A. Write between one and a half to two pages. Up to fifteen marks will be available for the content of your answer and up to five marks for the quality of your writing. [20 marks]
> Question 2
> Re-read the descriptions of (a) Signor Allesandro's enjoyment of the curry in paragraph 1 and (b) the traffic jam in paragraph 3. Select words and phrases from these descriptions, and explain how the writer has created effects by using this language. [10 marks]

In order to fully respond to Question 1 in the CIE paper, the candidate has to focus on the three parts: the eccentric behaviour of Allesandro, the unexpected happenings and whether or not Allesandro should retire. The first part insists that candidates read and understand the character of the "great conductor" and in particular his arrogance. In the latter parts of the question, the candidates also need to make judicious interpretations of both character and behaviour, based on what they have read. Candidates are tested on their ability to go beyond a mechanical reproduction of parts of the text. The format of the interview, even though the interviewer's questions are provided, is another test construct that seeks to measure the ability to articulate experience, express what is thought and felt, present ideas in an acceptable format and use language and register appropriate to the task set.

Question 2 in the CIE examination, for instance, is marked for the candidate's ability to select effective or unusual words and demonstrate an understanding of ways in which language is purposely made effective by the writer's conscious choices. The test construct seeks to establish the candidate's ability to select words that carry specific meanings, including implications. Commenting on a writer's language is in itself already a meta-linguistic task and the candidates are cognitively stretched to make sensible comments on the language of the writer and the consequent effects that are created through this usage.

Questions 1.1. to 1.8. based on the first passage are similar in their taxonomy to Qustions 2.1. to 2.8. based on Passage B in the DoE examination paper. Question 2.4. for instance asks the candidates:

> "State whether the following statement is true or false and give a reason for your answer. Buyiswa has two biological daughters. [2 marks]

True-False questions are, in general, hackneyed test constructs and even though they could be defensible, they do not sufficiently pose cognitive challenges on the learner.

The last question on Passage B in the DoE paper, "Give a suitable title for the passage, using no more than six words" is worth two marks and is very predictable to any candidate who has read through the magazine article on self-actualisation and personal fulfilment.

Passing scores for the DoE English Home Language paper are a cause for concern. Pegged at 40%, this pass mark is comparably lower than the 60% cut-off point for grade C in the CIE paper. One important feature of any examination process is “establishing defensible passing scores.” The scales that validate performance descriptors in the DoE paper in this instance are skewed to promote “mediocrity” (Jansen, 2010). Table (ii) shows the performance distribution of the candidates and the grades awarded by CIE in 2010:

Table (ii): CIE Performance distribution of candidates by grade and percentage

Passing grades	A*	A	B	C	D	E	F	U	G
Raw score	91-100	81-90	71-80	61-70	51-60	41-50	31-40	21-30	00-20
Number of candidates	1982	19820	75316	172434	295318	392436	408292	329012	289372
% of total candidates	0.1	1.0	3.8	8.7	14.9	19.8	20.6	16.6	14.6
Total number of candidates	1982	21802	97118	269552	564870	957306	1365598	1694610	1983982

N= 1 983 982 candidates for First language English in 2010, CIE

For the DoE examinations in 2009, Table (iii) shows the performance distribution of the candidates and the grades awarded:

Table (iii): DoE Performance distribution of candidates by grade and percentage

Passing grades	1	2	3	4	5	6	7	8	9
Raw score	91-100	81-90	71-80	61-70	51-60	41-50	31-40	21-30	00-20
Number of candidates	8592	9129	15036	23091	33294	94512	121362	114381	97197
% of total candidates	1.6	1.7	2.8	4.3	6.2	17.6	22.6	21.3	18.1
Total number of candidates	8592	17721	32757	55848	89142	183654	305016	419397	516594

Reporting test scores, especially the test scores of a high- stakes examination such as the CIE and DoE matriculation examination, entails a high degree of quality control and appropriate timing. DoE results for 2010 were released in February, almost a full month after CIE results had been released in January. This compares unfavourably on the timeliness of the release of public examination results, especially considering that CIE processed 1 983 982 English First Language candidates’ results while DoE processed 516 594 candidates. Another major hurdle, in addition to the time lapse, relates to the controversy surrounding the publication of these results. Grading and grade review procedures were inadequate and lacking in uniformity in the case of the DoE examinations. Thirdly, whereas standardization is viewed as a statistical necessity as well as a procedural one, the integrity of the South African English Home Language examination was severely challenged as the performance standards of the candidates were “adjusted upwards”(Howie, 2009).

Comparable curriculum outcomes versus comparable proficiency and performance outcomes

This paper has indicated that there have been two significant changes in the South African curriculum: the transition from apartheid to a democratic dispensation necessitated the first change, while a human resources and curriculum implementation challenge necessitated the second one. The second change, which has brought the more problematic hiatus, needs to be examined more deeply. Whereas the political agenda has pushed for these paradigm shifts in the spirit of "redressing the imbalances of the past," it is also important to observe that this shift could have disadvantaged learners through extraneous factors such as teacher under-preparedness, the novelty of new materials and the introduction of unfamiliar assessment techniques, including, amongst others, continuous assessment. The comparable outcomes perspective contends that the first cohort of students on RNCS 2007 should have grades equivalent to the last cohort on the old curriculum. Considering the test questions in the English First Language from CIE and DoE, it is possible to conclude that through "social moderation," item difficulty and item discrimination analyses, the CIE questioning and response calibration offered higher cognitive challenges when compared to the DoE papers. On test design and test assembly, i.e. the test forms, such as essay, multiple choice and structured questions, the CIE test papers offered more robust test constructs than DoE. Whereas the curriculum blueprints of OBE (South Africa) and CIE might compare favourably, specifically with regard to operational definitions of content and frameworks of validity, the DoE test papers offered indefensible samples of content and cognitive demands.

Passing scores

In terms of "defensible marks" for each grade awarded, CIE used the following distribution to award the respective grades:

Table (iv): CIE defended grade cut-off points

Component	Maximum mark available	A: Minimum mark required for grade	C: Minimum mark required for grade	E: Minimum mark required for grade	F: Minimum mark required for grade
Paper 2	50	31	23	17	N/A
Paper 3	50	30	23	15	11

The threshold for grade B is set halfway between those for grade A and C. The threshold for grade D is set halfway between those for grade C and E. The threshold for G is set as many marks below the F threshold as the E threshold is above it. Grade A* does not exist at the level of an individual paper component but is, at the grade review meetings, awarded to those outstanding candidates who performed at comparably high levels relative to the two preceding examination years. Such internal comparability checks are set as checks and balances for the "standards" of the subject and the paper components. This breakdown was not available at the time of researching for this study from Umalusi and DoE in South Africa, but it would have been revealing to establish the "arbitrariness" of these grade boundaries.

Conclusion

Negative wash back, especially in the form of test corruption and test score pollution, becomes a significant challenge for any curriculum. Since examinations are such "a primary disciplinary site" (Shohamy, 2001: xxii), "what will be taught in schools is what will be examined, and what is not examined will not be taught" (Vinjevold, 2005: 16). From the previous discussion on the two boards' protocols on test development and assembly, test production and passing scores (defensibility, comparability), it would appear that the DoE is threatened by a "hyphenated validity" where its credibility at certification is threatened and undermined when a majority of candidates who have not reached a defensible level of achievement are certified as competent. Bishop (1998) argues, convincingly, that examinations across boards should be content equivalent where this equivalence is evident in the type of tasks, types of questions, knowledge-cum-cognitive domains and the test distribution matrix.

References

DOWNING, S.M. & HALAYDNA, T. M. 2006. *Handbook on test development*. London: Routeledge.

GANNON, P**.** 1985. *Assessing writing: Principles and practice of marking written English.* London: Arnold.

GEE, J. 1996. *Social linguistics and literacies: Ideology in discourses*. London: Falmer Press.

GREANEY, V. & KELLAGHAN, T. 1996. *Monitoring the learning outcomes of education systems: Directions in development.* Washington. D.C.: World Bank.

GRIESEL, A.B. 2003. Triangulation. Unpublished paper, Department of Social Research: Loughborough University, UK.

HOWIE, S.J. 2009. Standard setting: Some issues and considerations. Presented at Umalusi workshop for setting standards, Pretoria, 23 April 2009.

LARSEN-FREEMAN, D. 1978. An ESL index of development. *TESOL Quarterly, 12*(4), 493-448.

LILLIS, T. 2001. *Student writing: Access, regulation and desire*. London: Routeledge.

NAVES, T**.** 2006. The long-term effects of an early start on EFL writing. Unpublished PhD dissertation. Universidad de Barcelona. English Department, September 2006.

REA, M. 2004. Academic literacies: A pedagogy for course design. *Studies in Higher Education*, vol.29, No.6, pp.739-756

SKEHAN, P. & FOSTER, D. 1997. Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*(3), 1-27.

SKEHAN, P., & FOSTER, D. 1999. The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*(1), 93 – 120.

VINJEVOLD, P. 2005. Matriculation: What is to be done? pp.58-63.

WOLF-QUINTERO, K., INAGAKI, S. & KIM, H.-Y. 1998. *Second Language development in writing: Measures of fluency, accuracy and complexity. Technical Report 17*. Manoa, Hawaii US: University of Hawaii Press