

PUT LISTENING TO THE TEST: AN AID TO DECISION MAKING IN LANGUAGE PLACEMENT

Fiona Marais & Tobie van Dyk

Unit for Afrikaans and English, Language Centre, Stellenbosch University

Universities and other higher education establishments throughout the world, including South Africa, have become concerned about the academic literacy levels of the students they enrol. The problem at most South African tertiary education institutions, which is in line with global trends, is certainly considerable, with almost a third of the students identified as being at risk. A lack of ability in academic discourse is seen as a major cause of students' failure to complete their courses within the given period. In 2006, as part of a nationwide attempt to remedy the academic literacy crisis, Stellenbosch University, along with other academic establishments, officially decided to implement a test of academic literacy in both English and Afrikaans. At Stellenbosch University, the English version of this test is known as TALL (Test of Academic Literacy Levels). It was developed to assess reading and writing abilities in an academic context. The results are used to 'stream' students into programmes which assist them in acquiring the various skills deemed necessary for their academic success. Students are sorted according to their TALL results into 'high risk' and 'low to no risk' categories, however, a need has been identified for further screening of the borderline students whose performance in the test falls between these two groups. Administrative and logistical limitations have, thus far, prevented listening skills from being included in the construct of TALL, but there is general consensus that listening is an important skill, particularly at university level. The focus of the research project, reported on in this article, was to design, and put into practice, an academic listening test (ALT) to assist in decision making regarding the placement of first-year students in approved language courses at Stellenbosch University. The qualitative and quantitative results obtained from the various administrations of ALT were analysed to determine the reliability and validity of the test. The final phase of the study involved the correlation of these results with those of TALL to establish whether ALT could assist the TALL administrators in making more informed decisions.

1. INTRODUCTION

There is increasing concern among universities and other higher education establishments throughout the world about the through-put rates of students in general, and more specifically of the first year intake. South Africa is no exception to the decrease in global through-put rates; however, the South African context should also be borne in mind. As a result of several changes in the educational landscape over the last number of years (among others the introduction of a new school curriculum, the replacement of the former final matriculation examination with a new end of school exam and certificate, and the massification of higher education), it is commonly believed that many students arrive under prepared at tertiary institutions (Van Dyk, Zybrands, Cillie & Coetzee, 2009:333). A comprehensive study by

Scott, Yeld and Hendry (2007) found that only 30% of students in the South African higher education system graduated after five years, while 14% are still registered and 56% eventually left without graduating at all. The estimated national completion rate therefore is currently calculated at 44%.

There are, of course, many variables influencing study success, however, most would agree (Van Schalkwyk, 2008:2) that an under-preparedness for the ‘intellectual demands of higher education programmes’ has often been cited as a contributory factor to the current problem. More specifically, a lack of proficiency in the language(s) of teaching and learning and an inability to deal with the language demands of higher education have a detrimental effect on student success. Weideman (2003:56) rightfully points out that a lack of ability in academic discourse (especially where reading and writing is concerned – often referred to as academic literacy) is considered to be one of the major causes of academic failure. The same author also suggests that mechanisms be put in place to address these issues.

2. PROBLEM STATEMENT AND RATIONALE

In 2006, as part of an attempt to support students in terms of their academic literacy development, Stellenbosch University, in collaboration with academics from other universities, officially decided to implement a test of academic literacy in both English and Afrikaans. The English test, known by the acronym TALL (Test of Academic Literacy Levels) is a paper-based, diagnostic and placement test that focuses on reading and writing tasks, similar to those which would be required of students at tertiary education level. The results of TALL are currently used to place students in specific programmes that assist them to acquire some of the skills needed for academic success. Students are sorted according to their TALL results into the broad categories of *high risk* and *low to no risk*. However, TALL is a first and robust screening of students’ reading and writing abilities and the need for a more refined method of screening borderline or Code 3 students, whose performance in the test falls between the above-mentioned two categories, has been identified.

Although administrative and logistical limitations have, thus far, prevented listening and speaking skills from being included in the construct of TALL, there is general consensus that both of these are also important skills for survival at university. In fact, listening skills are considered particularly important at university level for obvious reasons and this realisation impacts directly on the research problem of whether ALT could assist TALL in the screening of students. This forms the rationale and constitutes the relevance of the research presented in this article, whereby an academic listening test was designed, operationalised and examined for usefulness as an added dimension to TALL. To avoid confusion, the academic listening test will henceforth be referred to as ALT (Academic Listening Test).

3. RESEARCH DESIGN

As mentioned above, various constraints have prevented listening skills from being tested at Stellenbosch University to date. The initial focus of our research, therefore, was to design and implement an appropriate listening test (ALT) as an additional screening measure for students. The second phase involved a correlation of the results of ALT with those of TALL

to determine whether a listening component would provide a basis for making more informed decisions about placement when assessing the literacy levels of incoming students.

ALT was designed as a computerised test which could be used to qualitatively and quantitatively assess the academic listening skills of a selection of first-year university students. The various texts, tasks and questions included in the test were intended to measure a wide range of listening skills. Moreover, ALT attempts to measure listening competency in understanding both explicit and implicit information (Bejar, Douglas, Jamieson, Nissan & Turner, 2000:5). In accordance with the ethical requirements of Stellenbosch University, ALT was required to include a disclaimer which provided prospective test takers with information on the rationale behind the test and stipulated that no student would be disadvantaged by taking the test. In addition, candidates were made aware of the fact that data would be referred to collectively and not individually, therefore names, student numbers or other personal information were not included in the study. An opportunity was also provided for every candidate to give their consent to the results being used for research purposes.

To address issues of content, face and construct validity, a pilot test was carried out, in which participants responded to a questionnaire. The pilot testing was also an important check for any technological problems which could threaten the construct validity of ALT.

In order to determine the reliability of ALT, a retest was conducted a month after the initial testing and the results obtained from both administrations of ALT were analysed to present an argument for validation. A further validation study involved the use of internal consistency coefficients to assess the degree of reliability for the four individual tasks, as well as for each item included in the tasks. Alderson, Clapham and Wall (1995:178) state that 'a classic concurrent validation would involve comparing scores on the test in question with scores on some other test known to be valid and reliable'. Since TALL has proven reliability and validity (Van der Slik & Weideman, 2005; Van der Walt & Steyn, 2007), the correlation coefficient of 0.72 ($p = 0.00$) measured on the first ALT administration and 0.67 on the second, provided convincing evidence of concurrent validity.

4. THEORETICAL FRAMEWORK

4.1 ACADEMIC PREPAREDNESS

Lecturers at Stellenbosch University have noticed an increasing lack of critical and analytical thinking skills expressed through language (verbal reasoning) amongst their students over the last number of years. The concept of building an argument by providing the necessary evidence, as well as distinguishing between fact and opinion, seems to pose a problem for some students. This appears to indicate a lack of the necessary academic literacy skills required for success at a tertiary level (Van Schalkwyk, 2008:2). Tests of academic literacy, therefore, are designed to assess the degree to which students possess the necessary cognitive and linguistic capabilities to cope with university courses. In a country such as South Africa, with its widely diverse population, range of schooling standards and socio-economic situations, this seems to be of particular importance (Cliff, 2003:2; Van Dyk & Weideman 2004:2). Since school-leaving results may be inadequate in reflecting the potential of entry-level students to succeed in higher education (Cliff, 2003:2; Van Dyk & Weideman, 2004:9), test designers have to identify the kinds of tasks students will be called upon to perform in real-life situations and then attempt to replicate them as closely as possible in the test.

Furthermore, it appears to be widely acknowledged in the literature, that language tests are a means of measuring general or specific language abilities through the execution of tasks (Bachman, 1990; Bachman & Palmer, 1996; Weir, 1993).

Since a test of academic literacy would need to predict the future ability of entry-level students to meet the higher-order linguistic requirements of academic learning, it would be necessary to assess both their language ability and their thinking skills. This could then be used to gauge their preparedness for, and odds of success at, university, as well as the type of support that may be required to facilitate this.

4.2 ACADEMIC LISTENING ASSESSMENT

Before embarking on the design of a test of academic listening competency, as was the case in this study, it was necessary to do thorough research into the listening process as it is perceived by various scholars. Literature on listening strategies and factors that affect listening comprehension also provided essential information on listening as a construct. However, researchers have yet to agree on a widely-accepted definition of listening comprehension. This could be due to the numerous different processes and variables which are involved, making it almost impossible to provide a single comprehensive definition (Wagner, 2002:1). Nevertheless, accord seems to exist among researchers regarding the characteristics which make up the listening process (Brindley, 1998:172; Dunkel, Henning & Chaudron, 1993:180; Lynch, 1998:3).

In the development of ALT, tasks were selected that are representative of those usually required of first-year students at a university. Some examples of the skills which are included in the construct of both TALL and ALT are:

- understanding academic vocabulary in context;
- making a distinction between important and less important information;
- being able to infer meaning from implicit rather than explicit information (Van Dyk & Weideman, 2004:10).

4.3 ACADEMIC LISTENING TEST CONSTRUCT

Many researchers refer to listening as a two-stage process (Buck, 2001:51; Chaudron & Richards, 1986:113; Rost, 1990:33; Shohamy & Inbar, 1991:29; Weir, 1993:98). These two stages consist of bottom-up processing involving the more *local* skills such as the identification of details and extraction of facts, and top-down processing which requires interpreting the more implicit information such as inferencing or listening for gist. However, the two processes do not seem to take place in any particular sequence and they often occur simultaneously in a so-called parallel process (Rubin, 1994:211). This makes it very difficult to attribute task responses to any one particular skill or construct (Brindley, 1998:173; Buck, 2001:106).

The target language use (TLU) domain of ALT is set within the context of a university with its accompanying features such as lectures and tutorials. As is repeatedly mentioned in the literature, the language skills cannot be separated and a good example of this is specifically in a lecture situation where listening, reading and writing are fully integrated. Students listen to a lecture, take notes and then use the notes for study or assignment purposes. According to Ferris and Tagg (1996:299) and Kuehn (1996:29), academic listening tasks have proved challenging for all students, regardless of their mother tongue. In addition, Flowerdew

(1994:11) has found that the processing required for effective academic listening is far more complex than, for example, listening to a conversation. Rost (2002:162) maintains that this is because academic listening is mostly a non-collaborative or one-way listening process, of which a lecture is the most typical example. Furthermore, since lectures play such an important role in any academic programme, effective listening in lectures is fundamental for all students who wish to succeed at university (Flowerdew, 1994:7).

In general, the literature seems to emphasise the highly complex nature of the listening process, as well as the inherent difficulties in attempting to measure listening abilities. Furthermore, according to Brindley (1998:181), a lot of research still needs to be done towards a better empirical basis on which to design listening test specifications in the future and it is to this end that this paper wishes to make a contribution.

5. THE DESIGN AND OPERATIONALISATION OF ALT

The design process began with test specifications which determined both the method and the content of ALT. This test *recipe* stipulated the type and length of texts, details of the instructions, as well as how the test would be scored (McNamara, 2000:31). The framework of ALT was based on the theories and approaches of several researchers in the field, such as Buck (2001), Weir (1993), Wagner (2002) and Jordan (1997), as well as the compilers of TOEFL (Bejar, Douglas, Jamieson, Nissan & Turner, 2000). Since listening comprehension is an internal process which cannot be observed directly, researchers, up to now, have had to resort to assessing the more easily measured skills associated with the listening process (Brindley, 1998:172; Weir, 1993:98; Rost, 1990:33). For the purposes of this investigation, it was decided to use the *two-stage* listening model of bottom-up and top-down processing skills, mentioned above, as a theoretical framework on which to base the abilities to be tested by ALT.

ALT was adapted to fit into the assessment format of Blackboard, the learning management system (LMS) used at Stellenbosch University. The reasoning behind this decision was the ease and accuracy of scoring, as well as the computer's ability to instantly calculate statistical data. A design consideration which had to be reckoned with specifically was the issue of bias. Since the delivery mode of ALT is through the computer, careful consideration had to be given to ensuring that test performance would not be significantly affected by the level of a candidate's computer skills. In computerised testing, test designers need to be mindful of the test method effects, such as the quality of the recordings and the layout of the test (Douglas, 2000:277). Since reading on screen is known to be more difficult than on paper, font size and spacing are also important considerations. According to Buck (2001:255), the overriding issue is whether the computer can deliver tests that are more true to life and have a more realistic listening construct than conventional tests.

In order to address at least some of the issues raised above, ALT was divided into four sections or tasks which were placed in an *easier-to-more-difficult* order and test takers were advised of the listening purpose for each task. Clear instructions were given at the beginning of each task and, where necessary, additional information was given for some of the questions. All the tasks were designed to assess certain abilities, as well as represent the academic TLU domain on which they were based. The explanation below gives a more detailed description of the construct of the different tasks.

Task 1: This section required test candidates to *listen for certain instructions* and measured three main abilities, which were to:

1. recognise and remember specific instructions which include warnings, suggestions, recommendations and advice;
2. recognise the function of non-verbal cues such as stress and intonation, as indications of emphasis; and
3. deduce the meaning of words from the context.

6. Task 1: Question 2 (Points: 8.0)

The lecturer recommends certain steps which must be taken in order to access course material. Put the following steps in the correct order by selecting the step you think comes first next to **a)**, the second next to **b)** and so on.

Matching pairs

a)	- Select choice -
b)	Go to the Departmental webpage Scroll down to 'Language Change' Click on 'Ling 2'
c)	Click on 'Undergraduates'
d)	- Select choice -

Save and View Next Next Question

Finish

7. Task 1: Question 3 (Points: 2.0)

Where would a student listening to these instructions expect to find the 'copies' that the lecturer refers to?

a. In Lisa's office.

b. In his office.

c. In the thesis office.

d. In Delise's office.

Save and View Next Next Question

Finish

FIGURE 1: Screenshot of Task 1

Task 2: This comprised an extract from a first-year Psychology lecture and measured the ability to:

1. identify the main theme of a lecture;
2. recognise and recall important details and specific information presented in the text;
3. recognise and recall the stated opinion of the lecturer;
4. distinguish between the most important information and the supporting detail; and
5. concentrate on, and process, a long piece of text.

13. Task 2: Video (Points: 0.0)



When the video clip is finished, select 'True' and click **Save and View Next** to continue to the questions.

True False

14. Task 2: Question 1 (Points: 2.0)

What is the main theme of the lecture? Is it to determine whether:

- a. there are separate brain functions for object recognition and facial recognition?
- b. there is a part of the brain which is specially adapted to recognizing objects?
- c. there is a special module of the brain which is dedicated to recognizing faces?
- d. object recognition and facial recognition occur in the same part of the brain?

FIGURE 2: Screenshot of Task 2

Task 3: In this section, candidates listen to two students having an informal discussion on the legal aspects of euthanasia. Test takers are required to:

1. deduce information based on background knowledge of the topic and to fill in any gaps through awareness of the context;
2. make inferences about the speakers' attitudes or opinions;
3. distinguish between the most important information and the supporting detail;
4. identify the main theme, as well as the supporting arguments presented in the discussion;
5. identify reformulation as a means of agreement in a dialogue; and
6. listen for attitudes and opinions expressed both explicitly and implicitly.

30. Task 3: Question 3 (Points: 8.0)

Decide whether the following opinions are expressed by only one of the speakers or whether they agree. Type **T** for Tessa, **D** for Danie or **B** for both, in the boxes provided.

The doctor who overdosed his patient with morphine was justified in his actions.

An individual's rights are infringed by not allowing him to seek assistance in his suicide.

The 'right to life' does not mean the same as the 'right to death'.

A person should not be able to kill someone who is suffering.

35. Task 3: Question 8 (Points: 2.0)

Danie makes a statement that strong medical evidence is necessary before decisions can be made on the legalization of euthanasia. Tess shows her agreement by:

a. adding an extra piece of information to support his statement.

b. reinforcing Danie's point by rephrasing what he has just said.

c. quoting from the South African Law Commission.

d. stating that the doctor and the patient's family should be in agreement.

FIGURE 3: Screenshot of Task 3

Task 4: A senior lecturer in Business Science is speaking about foreign investment in this task; candidates need to be able to:

1. use an understanding of the text to fill in the content words omitted from the summary. According to Buck (2001:71), if test takers are asked ‘to fill in blanks on a summary of the passage ... [it] forces them to process the meaning’;
2. concentrate on, and process, a long piece of text;
3. listen for details and specific information; and
4. identify reformulation or paraphrasing.

40. Task 4: Question 1-16 (Points: 32.0)

The audio clip will play for a second time and you can type in the missing words. The words that come before and after the gap, will indicate the type of word that has been omitted, so make sure your responses fit into the grammatical context of the sentence. Words must be spelled correctly so that the computer is able to recognize them. A list of words is provided alongside the text which can be used to check the spelling of your answers.

Foreign direct investment is often considered to be merely a source of capital for a country but there are more complex issues which need to be considered. One of the primary considerations is the way in which **1** [] are structured. It is possible for a change of **2** [] to take place without money coming into the country. Therefore, what is of real significance is **3** [] control. Foreign investors, with their accompanying new technologies, practices and training strategies can result in improved performance, independent of capital funds.

South Africa is regarded as a 'chronic **4** []' when it comes to FDI. An explanation for this could be that there is a very strong **5** [] sector in South Africa and foreign firms find it difficult to get a **6** []. Generally, however, the inflow of technologies and managerial **7** [] is considered to be beneficial to South African industry.

An aspect of multinational investment is that there is a **8** [] effect. If a firm is below this, then it is below the basic level of **9** []. This can result in the firm being forced to close due to the arrival of foreign firms. This clearly has a very negative effect on a country's local **10** []. However, firms that are above the threshold, **11** [] from the challenges presented by the arrival of foreign firms. It provides an incentive to improve their goods and services as well as an opportunity to learn from the **12** []. Therefore, when considering the advantages and disadvantages of foreign direct investment, the **13** [] factor is the situation at home.

In conclusion, it is clear that governments and multinationals are motivated by very different goals. Multinationals, aim to make money and provide their **14** [] with good returns. Governments, on the other hand, have a duty to their citizens. However, both **15** [] and **16** [] play important roles in ensuring that South African firms remain above the threshold level where foreign direct investment has strong beneficial effects.

Wordlist

- benchmark
- benefit
- budget
- capability
- capacity
- challenges
- competence
- control
- corporations
- deals
- deciding
- determining
- domestic
- dominant
- economy
- efficiency
- entr ee
- experience
- expertise
- foothold
- governance
- government
- important
- industry
- internal
- infrastructure
- investors
- labour
- local
- management
- managerial
- multinationals
- overachiever
- profit
- revenue
- shareholders
- talent
- threshold
- underperformer
- ventures

FIGURE 4: Screenshot of Task 4

After completion of the design phase of ALT, a pilot project was conducted before the inception of the main study, so as to receive qualitative feedback on ALT and make sure that there were no technical hitches or possible bias. For this, a group of nine first-year students studying Health Science and three lecturers from the Unit for Afrikaans and English at Stellenbosch University’s Language Centre volunteered to complete ALT and respond to a questionnaire. The 16 questions included in the questionnaire, amongst others, concerned the representativeness and relevancy of the tasks, level of difficulty, listening purpose, as well as the listening skills assessed in each task. This feedback was important for gauging both content and face validity (the general opinion of peers). In addition, issues pertaining to the construct validity of ALT were also included in the questionnaire. These comprised aspects

such as: the appropriateness of texts and tasks; clarity of instructions; ease of navigation; and the sound and visual quality of the media files included in ALT. Apart from the above-mentioned questions, the questionnaire supplied to the lecturers contained an additional page of questions which related to their opinion of ALT as an effective indicator of academic literacy. At the end of the main questionnaire (for students and staff), respondents were given an opportunity to make any additional comments pertaining to aspects not covered in the questionnaire.

For the collection of quantitative data, volunteers from a group of six hundred and twenty-seven first-year Bachelor of Science students were asked to complete ALT. These students had all attended a semester of the module, *Scientific Communication Skills 172*, either in English or in Afrikaans, which provides assistance in developing academic literacy skills. For reasons of reliability, the administration of ALT was conducted in two parts, comprising an initial testing and a retest involving the same test, a month later. Ninety-seven students completed both administrations of ALT, which took place in a multimedia lab, situated in one of the University's areas for computer use, since headphones were a prerequisite for the test. The administration of all the tests was monitored by a supervisor to ensure that there were no technological problems or distractions which might affect the performance of the test takers.

6. DISCUSSION OF RESULTS

6.1 QUALITATIVE FINDINGS

According to McNamara (2005:25), test design begins by making decisions concerning the content of a test; in other words, the test construct. The purpose of the test should be stated as fully as possible since it is this *purpose* that will influence the content of the test (Weir, 1993:19). Bachman and Palmer (1996:178-9) maintain that it is not enough merely to select tasks *relevant* to the construct domain; they must also *represent* the target setting. Therefore, the content aspect of construct validity has to be addressed by considering both the relevance of the content, as well as how representative the tasks are.

In addition to the abilities that the test designer wishes to measure, the methods used to test these abilities also have a significant impact on test performance (Bachman, 1990:156; Alderson *et al.*, 1995:44). This is known as the *method effect* and developers of tests generally strive to reduce its influence. Bachman (1990:9) suggests that the biggest challenge facing testers of language is to make sure that the test methods will reflect performance that is characteristic of ability in a non-test situation. The purpose of the questionnaire used in this study, thus, was to ascertain the perceptions of ALT with regard to various types of validity and the degree to which the respondents thought the method effects impacted on the results. It was important, therefore, that the answers to these qualitative questions were based on the opinions of both experts and students.

The findings, based on the questionnaire used in the pilot project, thus concerned the assessment of content, face and construct validity of ALT through the opinions of both colleagues and students. For face validity, it is important for *non-testers*, such as students, to 'comment on the value of the test' and for experts in the field to 'judge the test', in the case of content validity (Alderson *et al.*, 1995:172). As has been mentioned, matters of construct

validity involve issues such as layout (particularly in computerised testing); clarity of instruction; and level of difficulty.

All the lecturers included in the pilot project felt that the tasks were completely relevant and representative and the majority of the students was satisfied with the relevancy of Tasks 1 and 2, with some reservations about Tasks 3 and 4. Since content validity is determined by professional opinion, the response from the lecturers seemed to indicate that ALT can be deemed content valid with regard to the representativeness of its content.

Overall, both experts and students were satisfied with the level of difficulty of the tasks. There was also general consensus on the computerised format of the test being easy to navigate and clearly laid out. The sound and visual quality of the audio and visual clips were considered good, rather than excellent, which is sometimes a price that has to be paid for authentic, non-staged recordings. The majority of the participants in the pilot project thought the instructions clear and unambiguous. This feedback seems to indicate that method effects were kept to a minimum and did not have a significant negative impact on the construct validity of ALT.

6.2 QUANTITATIVE FINDINGS

A recurring theme present in the literature surrounding language testing is the necessity for developers of tests to ensure, as far as possible, that tests are both reliable and focused on relevant validity. These two concepts are so enmeshed that a test cannot be reviewed for one without the other being taken into consideration. According to Alderson *et al.* (1995:187), a test cannot be valid if it is not reliable. In other words, a test has to be consistent in its measurement or it cannot be considered accurate. On the other hand, a test can be reliable without being valid if consistent test results are recorded but the test does not measure what it was designed to assess. However, Alderson *et al.* (1995:188) maintain that, when conducting validation studies, the most important consideration is ‘whether the test yields a score which can be shown to be a fair and accurate reflection of the candidate's ability’.

Various statistical analyses were carried out as part of the validation of ALT itself, as well as of the two sets of scores it provided in the assessment of academic listening. Included in the analyses, was an examination of how the various components related to one another, as well as an investigation into the validity and reliability of the test as a whole.

6.2.1 RELIABILITY OF ALT

6.2.1.1 INTERNAL CONSISTENCY

The first validation of the test involved computing the internal consistency for each of the four tasks by calculating the Cronbach alpha coefficients as indicated in Table 1 below. Cronbach's alpha is a reliability measure that determines the ‘average correlation of items in a survey instrument to gauge its reliability’ (Reynaldo & Santos, 1999). The item-total correlation for each item was also measured, in order to determine how each item related to the other items in the scale.

TABLE 1: Cronbach alpha coefficients

	Alpha (Test 1)	Alpha (Test 2)	Number of items
TASK 1	0.77	0.72	7
TASK 2	0.72	0.82	12
TASK 3	0.31	0.51	9
TASK 4	0.92	0.94	16

The alpha reading of 0.31 on Task 3 in the first administration of the test was the lowest of the eight readings. This was largely to be expected, since Task 3 was based on implicit rather than explicit information and test takers were required to make use of inferencing skills in order to respond to the questions. This reading improved to 0.51 in the second sitting of ALT, possibly because the candidates knew exactly what to listen for in the retest. The overall internal consistency for Task 2, as shown by the Cronbach alpha coefficients of 0.72 and 0.82 in the first and second administrations respectively, display an acceptable level of reliability. The reliability coefficients of 0.94 and 0.92 in Task 4, on both the test administrations, are substantially higher than the norm of 0.8 (Weir, 2005:29), which indicates strong reliability. The readings of 0.77 and 0.72 in Task 1 also show adequate levels of reliability. The test, as a whole, only displayed an alpha measurement of 0.63, which was to be expected, since the test intentionally contained a variety of heterogeneous items.

6.2.1.2 CONSISTENCY OF MEASUREMENT OVER TIME

When parallel methods of checking for reliability were used in the test-retest, ALT showed that, as a complete test, it can be considered reliable. The two administrations of ALT were carried out a month apart to determine reliability over a period of time and the Spearman correlation method was used to gauge the consistency of measurement of ALT over time. Spearman’s Rho, a variant of Pearson’s *r*, is used as a correlation coefficient with rank data and expresses the degree to which two variables relate to one another (Plonsky, 2009). The results of this analysis are given in Figure 5, shown below.

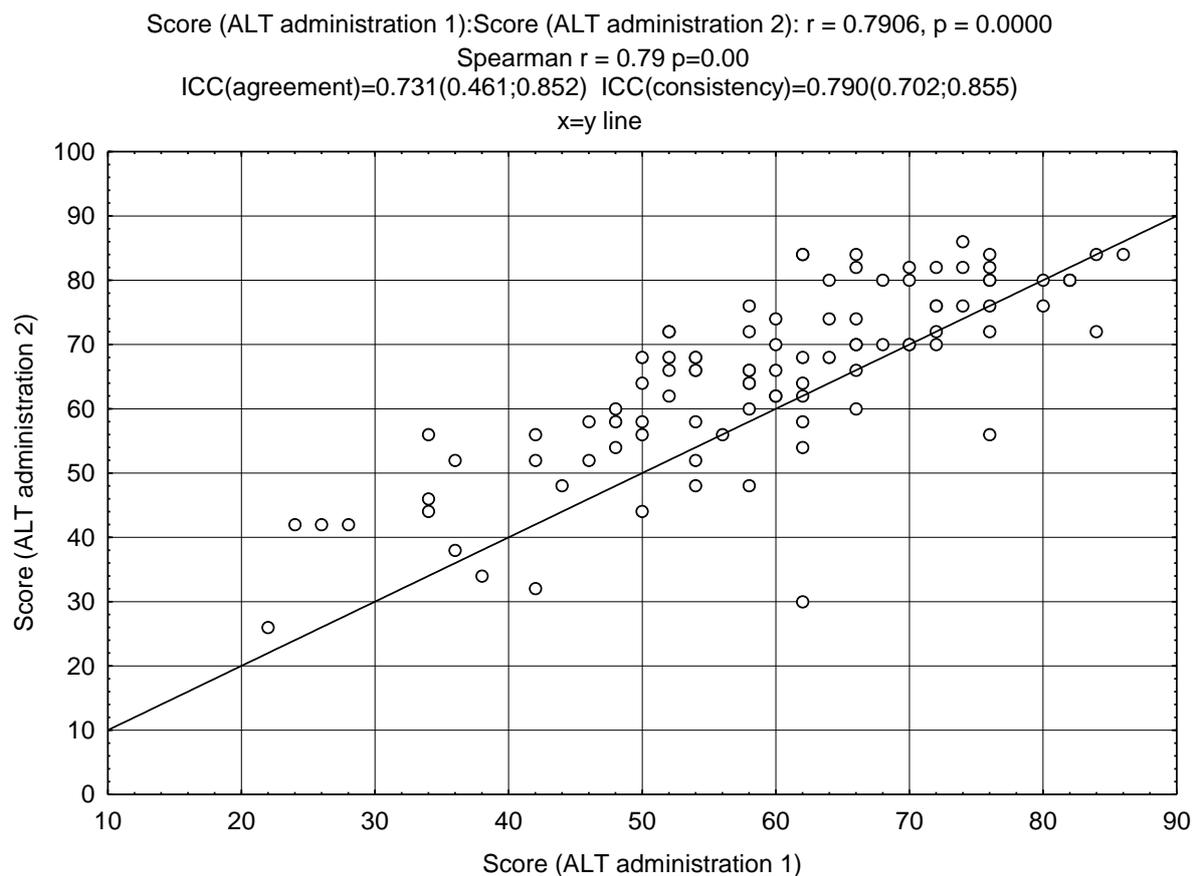


FIGURE 5: Comparison of scores on ALT administration 1 with ALT administration 2

The resultant correlation value of 0.79 in Figure 5 presents acceptable evidence that ALT has good consistency of measurement and can thus be considered reliable. The p value of 0.00 is also evidence that the correlation is significant. The intra-class correlation or ICC (agreement) has a reading of 0.73, which, once again, indicates good correspondence between the first and second administration.

Because of an assumption that candidates would find the second sitting of ALT easier than the first, the *agreement* reading of the ICC has a built-in penalty to control for bias, which the ICC (consistency) does not include. This explains the higher reading of 0.79 for the 'consistency' measurement and confirms that the second sitting of ALT yielded better results than the first. Visual proof of this assumption is provided by the pattern that emerges from the graph. This is illustrated by the clustering of circles above the line which indicate that scores on the second ALT administration were indeed higher than on the first. In spite of this improvement on the second sitting of ALT, the correlation between the results of the first and second administrations of the test is significant at 0.79 ($p = 0.00$), which indicates a strong accord between the two sets of scores. This, in turn, is also indicative of good test reliability.

Reliability, according to Davies (1990:52) is a 'statistical reassurance of consistency of result'. In other words, the results obtained are dependable. The quantitative results mentioned above are evidence of the reliability of ALT, which is of utmost importance, since no test can be valid if it is not reliable.

6.2.2 VALIDITY OF ALT

6.2.2.1 CONSTRUCT VALIDITY

For reasons of construct validation, the correlations of the different test sections/tasks were calculated by using the Spearman correlation coefficients to determine the degree of difference or similarity in the attributes being tested. Alderson *et al.* (1995:183-4) recommend assessing the construct validity of a test by comparing these different components or sub-tests with each other. The purpose of having different test sections is to test different abilities; this is usually indicated by reasonably low correlation coefficients. The correlations were thus expected to be in the region of 0.3 to 0.5 (Alderson *et al.*, 1995:184). Since a correlation is only considered significant if its p value is smaller than 0.05, the p value will be considered in the following discussion.

As can be seen in Table 2, shown below, the correlation coefficients between Tasks 1 and 2, Tasks 2 and 3 and Tasks 2 and 4, respectively, are within or close to, Alderson's recommended parameters and all show a p value of 0.00, which is an indication of significance. However, the correlation coefficients between Tasks 1 and 4 ($p = 0.00$) and Tasks 3 and 4 ($p = 0.01$), although significant, are slightly below the minimum of 0.3. Tasks 1 and 3 ($p = 0.64$) show an insignificant correlation, which is likely to be an indication of quite different traits being measured in the two tasks.

TABLE 2: ALT Sub-test correlations

Task	1	2	3	4
1				
2	0.37 ($p = 0.00$)			
3	0.04 ($p = 0.64$)	0.35 ($p = 0.00$)		
4	0.28 ($p = 0.00$)	0.53 ($p = 0.00$)	0.24 ($p = 0.01$)	

In the correlation between Task 1 and Task 2 in ALT, the reading of 0.37 ($p = 0.00$) stands to reason, since the items included in both these tasks require the test taker to listen for specific, mostly explicit information, but they are not so similar as to be testing exactly the same skills. However, when Task 1 was correlated with Task 3, the correlation was insignificant, since Task 3 focused on entirely different listening skills from those required in Task 1. Task 3 was based on implied information, where candidates had to infer meaning from what they had heard in the audio clip, while Task 1 involved recognising and remembering specific instructions. The correlation between Task 2 and Task 3 ($p = 0.00$) was within the optimum parameters of 0.3 and 0.5, which was surprising, since the traits that Task 2 was supposed to measure had more in common with Task 1, which showed an insignificant correlation with Task 3. The correlation of 0.28 between Task 1 and Task 4, which proved to be significant with a p value of 0.00, is closer to the 0.3 mark. This was to be expected, since Task 4 also involved listening for facts and deducing the meaning of words from the context; indeed, one might have expected the correlation to be even higher. The high correlation of 0.53 between Task 2 and Task 4 ($p = 0.00$) was also somewhat unexpected, as, although there were shared traits, the two tasks essentially were designed to test mostly different abilities. The low but

still significant correlation of 0.24 ($p = 0.01$) between Task 3 and Task 4, again, was to be expected, since the traits measured by the two tasks had little in common. These correlations all indicate that the individual tasks each assesses different abilities to a greater or lesser degree, which seems to demonstrate some measure of construct validity.

6.2.2.2 CONCURRENT VALIDITY

It was decided, for purposes of concurrent validity, to use the June 2008 TALL results as the criterion against which to measure the two administrations of ALT, since all the tests were administered within approximately the same time frame.

According to Alderson *et al.* (1995:178), the correlation coefficient in a concurrent validity study should range from 0.5 to 0.7. Both the first and second ALT administration showed a significant correlation coefficient with TALL. A correlation coefficient of 0.72 ($p = 0.00$) on the first ALT administration and of 0.67 ($p = 0.00$) on the second, signifies that both values fell well within the optimal boundaries. Alderson *et al.* (1995:178) state that ‘a classic concurrent validation would involve comparing scores on the test in question, with scores on some other test known to be valid and reliable’. Since TALL has proven reliability and validity (Van der Slik & Weideman, 2005; Van der Walt & Steyn, 2007), the correlation coefficient of 0.72 ($p = 0.00$) measured on the first ALT administration and 0.67 on the second, provide convincing evidence of concurrent validity. In light of the fact that both TALL and ALT are language tests assessing academic communication skills, this result was not unexpected since some aspects are common to both test constructs, for example, understanding vocabulary from the context and making inferences from implied information. However, the correlation is not so high as to be an indication that the two tests are measuring exactly the same skills. The results, therefore, seem to show that the reading and writing skills assessed in TALL and the listening skills measured in ALT, although related, are still differing aspects of language ability.

As has been mentioned above, there are various types of validity which are confirmed either by qualitative or quantitative results, or both. These findings are all based on the relationship between the test instrument and the domain to be measured (Davies, 1990:6). In the case of ALT, there appears to be adequate qualitative evidence of content and face validity, as well as convincing quantitative proof of concurrent validity. In terms of construct validity, data from both qualitative and quantitative studies were computed to produce significant proof of this all-important type of validity.

6.3 COMPARISON OF ‘BORDERLINE’ OR CODE 3 SCORES ON TALL AND ALT

As was reported in the section explaining the problem statement, one of the main aims of this research was to determine whether ALT would be able to assist in more accurate screening of students placed in the borderline category on the basis of their performance on TALL. Since the preliminary sample group consisted of only four students, a reliable correlation between their TALL scores and their performance on ALT was not possible. It was then decided to do a replication study of the correlation between the scores of candidates in the borderline category of TALL and their performance on ALT. In this replication study, students whose scores on TALL, January 2009, placed them in the borderline category, were targeted and asked to volunteer to complete ALT. Twenty-five sat for the test and their results were correlated with those of TALL. A correlation value of 0.03 yielded a p value of 0.89, which signifies no correlation. Figure 6, below, shows the candidates’ results on TALL and ALT

respectively, and one can clearly see that there is no statistical correlation between the two. Nonetheless, it could also signify that, in this category, TALL and ALT are measuring very diverse skills. Moreover, since it is here that refined screening is needed; ALT could indeed be contributing different information regarding the linguistic skills of the test takers.

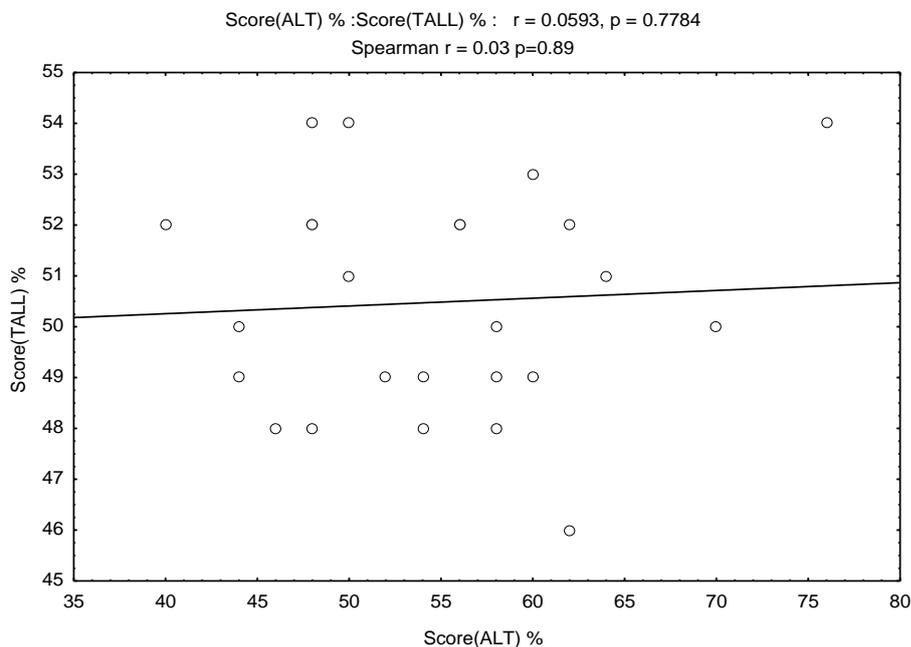


FIGURE 6: Comparison of code 3 scores on TALL and ALT – the replication study

The Code-3 candidates' scores on TALL, all fell between 46 and 54%, but their results on ALT were between 46 and 76%, with 14 students showing improved results except for one, who remained the same. However, no statistical pattern emerged and there was no significant correlation with TALL. The small sample size resulted in a restricted or truncated range, which could lead to an inaccurate or false reading of the data. It was therefore unfortunate that, given the limited number of students in both the initial and replication sample groups, a correlation between the borderline candidates on TALL and their results on the listening test could not be reliably calculated.

7. CONCLUSION

In a quest to confirm or refute the hypothesis of this investigation, namely that an academic listening test would be a useful added dimension to TALL, it was necessary to collect and analyse extensive qualitative and quantitative data.

The qualitative data to emerge from the validation study of ALT indicate that the method effects of the design, layout and mode of delivery that could adversely affect a candidate's performance, were kept to a minimum. The quantitative data demonstrate that ALT manifests good construct and concurrent validity, as well as evidence of reliability.

In this study, listening skills were assessed individually, as well as in combination with other language skills, so ALT should not be seen as a 'pure' assessment of listening ability (Bejar *et al.*, 2000:5). The integration of speaking, listening, reading and writing make up an

individual's proficiency in a language, but the mastery of these skills is seldom even (Lado, 1961:25). There has been some criticism of the skills-based approach to testing and Bachman and Palmer (1996:78) argue that, rather than trying to differentiate between the four language skills, it would be more useful to pinpoint specific tasks involving language use. These tasks could then be described in terms of their various characteristics, as well as the kinds of language ability they display. In this project, both the identification of TLU tasks that will be required of students in the future, and the simulation of those tasks in an academic listening test, formed the core of this study. However, owing to the small sample size, the original research was unable to provide conclusive evidence that ALT would be able to assist TALL in a more accurate screening of students. Further research to investigate the shortcomings of ALT, as well as the influence of the size of the sample group on the data, is necessary. This will form the basis for a doctoral study to research the relatively unknown field of listening testing. The Master's study, on which this article is based, will therefore be used as a springboard for future enquiry.

Nevertheless, the significance of listening as a facet of academic literacy was thoroughly investigated in this research and it can be concluded from the results that listening competency, although only a component of an integrated set of language abilities (Rost, 2002:172), plays an important role in the assessment of academic competency. Furthermore, since most validation studies are an ongoing process, there is always scope for revision and improvement in any language test. This process is reliant on research and there appears to be much research that still needs to be done in the field of listening testing. According to Wagner (2002:26), it is a field that has attracted an increasing amount of attention over the last ten years, but still presents many unanswered questions.

8. REFERENCES

- ALDERSON, JC, C CLAPHAM & D WALL. 1995. *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- ANDERSON, A & T LYNCH. 1988. *Listening*. Oxford: Oxford University Press.
- BACHMAN, L. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- BACHMAN, LF & AS PALMER. 1996. *Language testing in practice*. Oxford: Oxford University Press.
- BEJAR, I, D DOUGLAS, J JAMIESON, S NISSAN & J TURNER. 2000. *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series Report No. 19). Princeton, NJ: Educational Testing Service.
- BRINDLEY, G. 1998. Assessing listening abilities. *Annual Review of Applied Linguistics*, 18:171-191.
- BUCK, G. 2001. *Assessing listening*. Cambridge: Cambridge University Press.
- CHAPELLE, C. 2001. *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge: Cambridge University Press.
- CHAPELLE, C & D Douglas. 2006. *Assessing language through computer technology*. Cambridge: Cambridge University Press.

- CHAUDRON, C & JC RICHARDS. 1986. The effect of discourse markers on the comprehension of lectures. *Applied Linguistics*, 7(2):113-127.
- CLIFF, A. 2003. *Assessing the academic literacy skills of entry-level students using the Placement Test in English for Educational Purposes (PTEEP)*. [Online]. Available: <http://www.ched.uct.ac.za/seminars/archive2003/cliff2.pdf>. [20/01/2008].
- DAVIES, A. 1990. *Principles of language testing*. Oxford: Basil Blackwell.
- DOUGLAS, D. 2000. *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.
- DUNKEL, P, G HENNING & C CHAUDRON. 1993. The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal*, 77:180-191.
- FERRIS, D. & TAGG, T. 1996. Academic listening/speaking tasks for ESL students. *TESOL Quarterly*, 30(2):297-320.
- FLOWERDEW, J. 1994. Research of relevance to second language comprehension: An overview. In J Flowerdew (ed.). *Academic listening: Research perspectives*. Cambridge: Cambridge University Press. 7-29.
- JORDAN, RR. 1997. *English for academic purposes. A guide and resource book for teachers*. Cambridge: Cambridge University Press.
- KUEHN, P. 1996. *Assessment of Academic literacy skills: Preparing minority and limited English proficiency (LEP) students for postsecondary education*. California State University (UCLA). Report for the Improvement of Postsecondary Education, Washington D.C.
- LADO, R. 1961. *Language testing*. London: Longman.
- LYNCH, T. 1998. Theoretical perspectives on listening. *Annual Review of Applied Linguistics*, 18:3-19.
- MCNAMARA, T. 2000. *Language testing*. Oxford: Oxford University Press.
- PLONSKY, M. 2009. *Psychological statistics* [Online]. Available: <http://www.uwsp.edu/PSYCH/stat/7/correlat.htm#Vlb> [2009, 2 November].
- ROST, M. 1990. *Listening in language learning*. London: Longman.
- ROST, M. 2002. *Teaching and researching listening*. London: Pearson Education.
- REYNALDO, J & A SANTOS. 1999. Cronbach's Alpha: A tool for assessing the reliability of scales. *Journal of Extension*, 37(2) [Online]. Available: <http://www.joe.org/joe/1999april/tt3.php> [2010, 10 June].
- RUBIN, J. 1994. A review of second language listening comprehension research. *The Modern Language Journal*, 78(ii):199-221.
- SCOTT, I, N YELD & J HENDRY. *A case for improving teaching and learning in South African higher education*. Higher Education Monitor No. 6. Pretoria: Council on Higher Education [Online]. Available: <http://www.che.ac.za/documents/d000155/index.php> [2009, 17 September].

- SHOHAMY, E & O INBAR. 1991. Validation of listening comprehension tests: The effect of text and question-type. *Language Testing*, 8:23-40.
- VAN DER SLIK, F & AJ WEIDEMAN. 2005. The refinement of a test of academic literacy. *Per Linguam*, 21(1):23-35.
- VAN DER WALT, JL & HS STEYN (Jnr). 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort*, 11(2):138-153.
- VAN DYK, TJ & AJ WEIDEMAN. 2004. Switching constructs: On the selection of an appropriate blueprint for academic literacy assessment. *Journal for Language Teaching*, 38(1):1-13.
- VAN DYK, T, H ZYBRANDS, K CILLIE & M COETZEE. 2009. On being reflective practitioners: The evaluation of a writing module for first-year students in the Health Sciences. *Southern African Linguistics and Applied Language Studies*, 27(3):333-344.
- VAN SCHALKWYK, S. 2008. Acquiring academic literacy: A case of first-year extended degree programme students at Stellenbosch University. Doctoral dissertation. Stellenbosch: University of Stellenbosch.
- WAGNER, E. 2002. Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, Teachers College, Columbia University*, 2(1):1-39.
- WEIDEMAN A. 2003. Assessing and developing academic literacy. *Per Linguam* 19(1 & 2):55-65.
- WEIR, CJ. 1993. *Understanding and developing language tests*. Hertfordshire: Prentice Hall Europe.
- WEIR, CJ. 2005. *Language testing and validation*. Hampshire: Palgrave MacMillan.

BIOGRAPHICAL NOTES

Fiona Marais is with the Unit for Afrikaans and English at Stellenbosch University's Language Centre. She is currently involved with course design and lecturing in the field of academic literacy. She is particularly interested in language assessment and testing; she developed a computerised academic listening test as part of her Master's study. Email address: fcm@sun.ac.za

Tobie van Dyk, of Stellenbosch University's Language Centre, has been working intensively on different kinds of language tests designed for different purposes since 2004. His expertise, however, lies in the field of academic literacy testing and development. He currently heads up the Unit for Afrikaans and English in the Language Centre which is, among others, responsible for all language testing, as well as academic literacy development and the acquisition of Afrikaans and English at this university. In addition, he administers the collaborative effort of four local universities, known as the Inter-institutional Centre for Language Development and Assessment (ICELDA). Email address: tjvd@sun.ac.za