Criterion-related Validity of Strong and Weak Second Language Performance Assessments in a University Pathway Programme

MA in Language Testing (by Distance)

Nikolay Gochev

December 2013

Word Count: 16,322

Abstract:

The study investigates the predictive validity of three second language performance assessments in a foundation programme provided by a private college in partnership with a UK university. Adopting McNamara's (1996) distinction between weak and strong second language performance assessments, the study examines the relationship between the Final Academic Score awarded at the end of the foundation programme and language proficiency as measured by (a) two weak performance assessments: IELTS and Language for Study 1 (an in-house language test), and (b) a strong performance assessment: Skills for Study 1 (an end-of-course assessment). Spearman rank order correlation coefficients (r_s) were calculated for the three sets of scores. Significant positive correlations were found between the Final Academic Score and both the strong and weak second language performance assessment scores and sub-scores. However, the findings suggest that the strong performance assessment is a better predictor of students' academic achievement at the College than either of the weak performance assessments. Possible reasons for the findings are discussed in the light of the theoretical insights from the literature on second language performance assessment. The paper argues that the main factors which contribute to the better predictive power of the strong performance assessment are the task/context approach to construct definition, the use of indigenous marking criteria, and the untimed, process-oriented nature of the assessment. Limitations and implications for future research are also discussed.

Contents

1. INTRODUCTION

The UK is a popular study destination for international students. In 2011-2012, 435,230 international students were enrolled in UK universities (UK Council for International Student Affairs [UKCISA], 2013). Several routes into UK higher education are available to overseas students. Some students qualify for direct entry on a par with British students, for example EU nationals whose qualifications are equivalent to A levels, or students who hold the International Baccalaureate (IB), European Baccalaureate (EB) and Irish Leaving Certificate. Others can gain access to higher education via a range of pathway programmes such as Foundation Years or Foundation courses ("*Routes into university and higher education*", 2013). In addition to having appropriate academic qualifications, international students must satisfy English language entry requirements. The UK Border Agency [UKBA] (2013) specifies a minimum level of English language proficiency for university study of CEFR B2 and provides a list of approved English language tests. Furthermore, UK universities set their own English language entry requirements, which range from IELTS 5.5 to 7, or equivalent, for undergraduates (Hayatt & Brooks, 2009, p.14). Students who do not meet either the academic or language requirements have the option of enrolling in Foundation Programmes provided either by universities or private colleges, which typically include academic English, study skills, subject-specific and IT courses (Reinders, Moore, & Lewis, 2008, p. 11) aimed at improving students' academic qualifications and English language ability.

Undoubtedly, language assessments play a significant part in such programmes. Although a number of language proficiency tests such as IELTS and TOEFL are used for screening purposes (Howell et al, 2012), some authors have argued that their use as a measure of achievement in English for Academic Purposes (EAP) courses may be invalid (Alderson, 2000; Banerjee & Wall, 2006). Schmitt (2012) points out that the aim of EAP courses in pre-sessional and foundation programmes is broader than simply language teaching. Such courses are specifically tailored to the requirements of the academic target language use situation and therefore the assessments need to focus on measuring achievement rather than proficiency. Thus it could be argued that to justify the use of a particular assessment instrument, data must be gathered to provide evidence for the validity of the inferences drawn from that assessment. Both proficiency and achievement EAP tests claim to assess students' readiness for

university study. Therefore, one possible source of evidence is correlational research into the predictive validity of language assessments.

This dissertation begins with a Background section, which presents an overview of the literature on second language performance assessment. It outlines the distinction between strong and weak performance assessments based on their approach to construct definition and the criteria used to evaluate test-takers' performance. The construct in English for Academic Purposes testing is discussed, followed by two important aspects that impact on second language performance assessment, namely time and process-orientedness. After that, the importance of criterion-related validity within Messick's (1980) concept of validity as a unitary concept is discussed, and the results of previous predictive validity studies evaluated. Finally, a rationale for the current study is presented.

The second part of the Background section describes the research context and formulates three research questions which the study aims to answer. Section 3 focuses on the methodology and provides information about the participants of the study, the assessment instruments used, the data collection methods and data analysis procedures. Section 4 introduces the rationale for selecting the appropriate correlation coefficients for this study and reports the results of the correlational analysis. Section 5 attempts to answer the three research questions by analysing the extent to which the results from the study support the theory behind weak and strong performance assessments outlined in the literature review. The dissertation ends with a conclusion and suggestions for future research.

2. BACKGROUND

2.1 Literature review

2.1.1 Performance assessment

A performance test is ‘a test in which the ability of candidates to perform particular tasks, usually associated with job or study requirements, is assessed’ (Davies et al, 1999, p.144). The need for performance assessment arose in the 1960s for two main reasons. The first one, as the definition above implies, was to evaluate the language proficiency of overseas professionals, such as doctors and teachers, who sought employment in English-speaking countries, and of foreign students who wished to study in English-medium universities. The second reason was that language testing had to keep up with the developments in the classroom, which began to emphasise the teaching of communicative language ability (McNamara, 1996, p. 25). As Wigglesworth (2008) points out, compared with the traditional discrete item tests, performance based assessments aim to evaluate examinees on a wide range of language abilities. However, this assessment procedure introduced two factors missing from the traditional tests, namely *performance* on the part of the test-taker and a *judging process* used to evaluate that performance (McNamara, 1996, p. 10). Figure 1 (adapted from McNamara, 1996, p.9 and Skehan, 1998, p.170) sums up the characteristics of performance assessment.

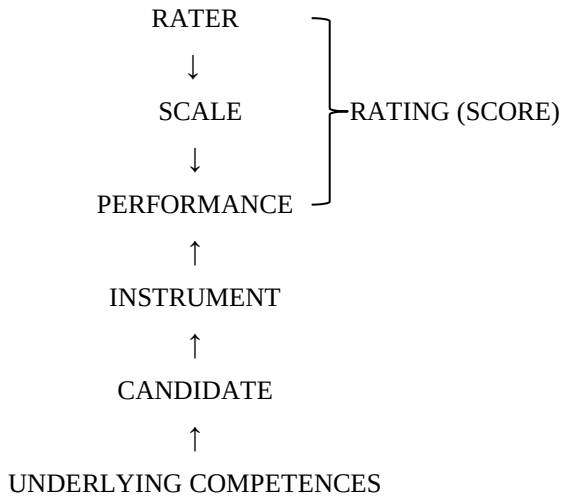

Figure 1. The characteristics of performance assessment

This model illustrates that a number of factors and the interactions between them may affect the rating of a candidate and consequently the decision taken on the basis of the test (Skehan, 1998, p.169). As the aim of assessment is to provide 'an adequate and accurate basis for making trustworthy interpretations' (Norris, 2002), I will next consider the role of the assessment instruments, the rating criteria and the context in which they are used in achieving that objective.

2.1.2 Strong and weak second language performance assessments

Performance assessments are often described as direct, authentic and realistic, designed to elicit and test real-world skills (McNamara, 1996; Bachman, 2007; Brown, 2004). However, McNamara (2006, p.43) argues that such features are often a matter of degree and that actual tests will most likely be positioned somewhere along the continuum from 'strong' to 'weak' second language performance assessments.

In the 'strong' sense, real-world tasks and criteria are used and the target of assessment is the execution of the task itself. Any assessment of language ability is incidental in as much as it contributes to the successful completion of the task (McNamara, 1996, p. 43). As McNamara (1996, p.43) puts it, 'adequate second language proficiency is a necessary but not a sufficient condition for success on the performance task'. Brown, Hudson, Norris and Bonk (2002) use the term 'task-based language assessment' to describe strong performance assessments, defining the construct of interest as the ability to 'accomplish particular tasks or task types'. In his discussion of construct definition in language testing, Bachman (2007, p.42) also characterises the construct of strong performance tests as task/context focused. Moreover, Skehan (1998, p.156) observes that such tests represent 'direct methods of gathering data' with little emphasis on underlying abilities, but with higher success rate in predicting performance in the real world. This success, according to Skehan, is due to a large extent to the fact that the 'real life approach' restricts itself to making predictions in specific, clearly defined situations.

Task-based performance is assessed by employing rating scales. Douglas (2000, p.68) refers to the criteria used by specialists in the real world as 'indigenous assessment criteria' and argues that such criteria should be used as the basis for constructing the

rating scales of specific purpose language tests. Moreover, since in the 'strong' sense successful completion of the task is the criterion according to which performance is evaluated, non-linguistic factors are seen as contributing to rather than diminishing the authenticity of the test. Thus, it is conceivable that lower general language proficiency may not result in failure for a candidate who has other relevant skills, while a highly proficient candidate may actually fail a strong performance test if they lack those relevant skills (Jones, 1985, cited in McNamara, 1996, p.39). Furthermore, Brindley (2009) points out that outside applied linguistics successful communication is often evaluated by applying performance criteria such as 'empathy, behavioural flexibility and interaction management' and that language testers may need to consider such criteria in performance assessment. Overall, strong performance assessments place less emphasis on language skills and focus more on task completion whose success is judged on the basis of indigenous criteria derived from the target language use situation.

In the 'weak' sense, the tasks of second language performance tests only resemble or simulate real-life tasks, but the actual focus is on 'language performance' (McNamara, 1996, p. 44). In other words, the tasks are the means through which an adequate sample of language is elicited for the purposes of determining the candidates' level of proficiency. This chimes in with Long and Norris's (2009, p.138) observations that in some tests tasks are used simply 'as a means for eliciting particular components of the language system which are then measured and evaluated'. Bachman (2007, p.42) defines the construct of weak performance tests as 'trait/ability focused', i.e. as 'an ability or capacity that resides in the individual' and therefore performance on the test allows for inferences to be made about a test-taker's 'underlying ability for language use'. McNamara (1996, p.44) points out that while weak performance tests may appear to mirror strong performance tests, such resemblance is largely superficial. They are mainly used because of their assumed face validity, positive washback and as a way of offsetting two major disadvantages of strong performance tests, namely time and expense. McNamara (1996, p.44) notes that most performance based general proficiency tests are of the weak type and even EAP tests such as IELTS are not strong performance tests. This view is supported by Brown (2004) who states that composition writing and oral interviews which are evaluated on the basis of linguistic criteria like accuracy, range and fluency are weak performance tests. A possible

consequence is that although the tasks that comprise weak performance tests like IELTS may resemble the tasks that students undertake at university, some students, who fail these simulated tasks because of low proficiency in English, are likely to adopt a range of strategies in the real world that will allow them to complete academic assignments successfully (McNamara, 1996, p. 42).

On the other hand, according to Bachman (2007, p.56), what distinguishes these two approaches is not so much the types of assessment tasks used but the inferences which are drawn from the test scores. The 'trait/ability' approach is used to support inferences about the abilities of the examinees, while the 'task/context' approach is mainly concerned with 'predictions about future performance on real-world tasks'. Figure 2 illustrates these two different views of language proficiency together with the possible interpretations of test results.

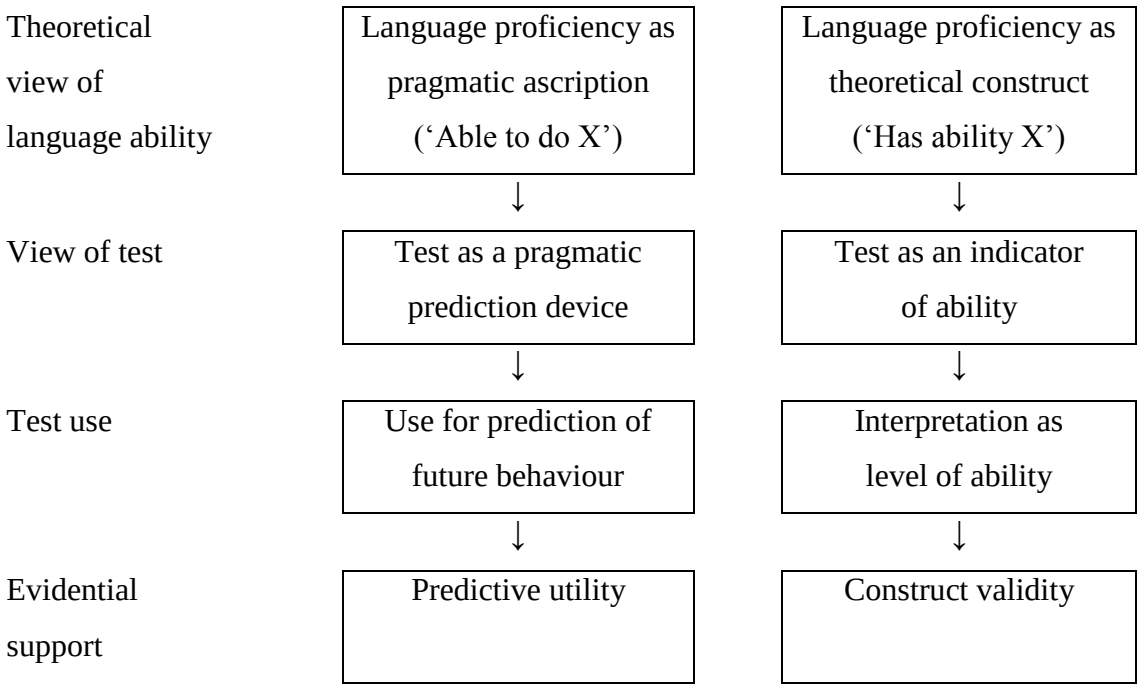

Figure 2. Contrasting views of language proficiency, uses/interpretations of test results, and validity (Bachman, 1990, p.254)

According to Bachman (1990, pp.253-254) evidence of 'predictability does not constitute evidence for making inferences about ability.' In other words, the inference

‘A is able to do X’ therefore ‘A has ability Y’ is not valid. What is not clear, however, is whether the inference ‘A has ability Y’ therefore ‘A is able to do X’ is a reasonable one to make. But this is exactly what weak performance assessments seem to be used for. Alderson, Clapham, and Wall (1995, pp.180-181) point out that proficiency tests like IELTS and TOEFL are used to make predictions about the future performance of students in English-medium universities. A strong argument against such an inference is the fact that native speakers of English do not necessarily perform better at university than international students, who are presumably less proficient (Dooey & Oliver, 2002; Cho & Bridgeman, 2012). This raises questions about the fairness of the gate-keeping function of weak performance assessments like IELTS, namely the extent to which cut-off scores on weak performance tests can be justified without recourse to scores from strong performance tests. If these two approaches support different inferences about the test-takers, it is arguable that no important decision should be taken on the basis of only one of them. Such a view is expounded by a number of researchers who emphasise that inferences about test-takers should be drawn from a range of sources to ensure fairness and accountability in language testing (Dooey & Oliver, 2002; O’Loughlin, 2011; Rees, 1999; Shohamy, 2001).

2.1.3 The construct in EAP

Evidently, the way the construct of interest is defined affects the inferences that can be drawn from test results. As Skehan (1998, p.153) observes the development of weak and strong performance assessments came from a conflict between three underlying issues in language testing: inferring ability, predicting performance, and generalising from context to context. Performance assessment in EAP contexts inevitably has to address these issues. Arguably, the most important question is: what construct should EAP tests measure? One common approach to defining the construct is analysis of the target language use situation (McNamara, 1997; Douglas, 2000), which invariably raises the issue of academic skills. Jordan (1997, p.7) provides a comprehensive list of skills that are essential for both native and non-native speaker students. In addition to the four skills of listening, speaking, reading and writing for academic purposes, emphasis is also placed on research and reference skills, i.e. effective library use, finding and analysing evidence and data, summarising, paraphrasing and using quotations and in-text citations. Although these skills are considered transferrable if

the students already possess them, such an assumption cannot be made automatically as students arrive in English-speaking countries with a range of previous educational experiences (Jordan, 1997, p.5). This means that the objectives of foundation courses for international students typically include the development of both language proficiency and adequate study skills. In terms of assessment, then, achievement tests whose marking criteria reflect the real-life skills necessary for the academic environment in the UK are more likely to provide an accurate measure of readiness to study in an English-medium university than a more general academic English proficiency test. As Banerjee and Wall (2006) point out using external EAP tests such as IELTS 'would risk construct under-representativeness' and therefore reduce the usefulness of the scores for the stakeholders, namely the various university departments and the students themselves.

Indeed, some authors (Green, 2007; Moore & Morton, 2007) have argued that the IELTS writing construct does not reflect academic writing at university. For example, Moore and Morton (2007) identify three main areas of discrepancy between IELTS writing task 2 and writing tasks in higher education. First, in IELTS essay tasks the test-takers are expected to rely on their personal experience and background knowledge. However, in higher education settings students are expected to make use of published sources, which they need to be able to paraphrase, summarise and cite in order to support the arguments put forward in their course assignments. Second, IELTS tasks do not include a variety of text types and genres, nor do they include the same range of rhetorical functions as academic tasks. Third, IELTS tasks are primarily *phenomenal* in nature as they focus on 'real-world […] situations, actions, practices', while university tasks can be *phenomenal* and *metaphenomenal*, i.e. emphasising 'abstract […] theories, ideas, methods' (Moore & Morton, 2007).

In an attempt to address such construct under-representativeness in large-scale EAP examinations, many institutions have introduced integrated reading/writing, listening/speaking, and reading/listening/writing tasks. Lewkowicz (1997) summarises the advantages of integrated tasks: they are direct, add realism to assessments, and are appropriate for homogeneous populations whose language use needs are clearly identifiable as in university settings.

Reading-to-write tasks are described as authentic because in academic contexts writing is always preceded by reading a wide variety of sources (Hamp-Lyons & Kroll, 1997; Weigle, 2004; Plakans, 2008; Gebril & Plakans, 2013). Furthermore, an impromptu unseen writing task is inauthentic because even if university students have to be able to write mini-essays in response to exam questions, they will have prepared and revised for that exam by reading the relevant coursebook and other assigned texts. A second argument for the use of reading-to-write tasks is that they may reduce the background knowledge effect as all candidates will have access to the same information on which to base their arguments (Weigle, 2004). Studies of reading-to-write tasks have shown that lower level students make less use of the source texts, perhaps as a result of weaker reading comprehension skills (Gebril & Plakans, 2013); that better inferences can be made about academic writing ability in terms of planning and discourse synthesis than on the basis of writing-only tasks (Plakans, 2008); and that preparing for a reading-to-write test has a positive washback as it enhances the skills that students require for academic study (Weigle, 2004). Moreover, Weigle (2002, p.178) argues that the construct of writing in academic contexts should include the ability 'to make use of all available resources' to communicate effectively with the intended audience rather than the ability to create texts by simply relying on one's own linguistic competence. She emphasises that although this 'complicates the definition and measurement of the construct enormously', it is a better reflection of what writers do in the real world.

Integrated Listening-and-Speaking or Listening-and-Writing tasks require test-takers to listen to a recording and then incorporate information from the listening text into speech, for example by giving an oral summary of the content of the talk, or into an essay, for example as support for their arguments. Such tasks are viewed as integral part of academic study where students are frequently required to make use of multiple sources, including lectures and podcasts (Taylor & Geranpayeh, 2011; Slaght & Howell, 2007). As with reading-to-write tasks, this complicates the construct definition as the task involves more than one skill. However, despite the drawback of 'muddied measurement' (Weir, 1990, cited in Lewkowicz, 1997, p.127), several authors emphasise the increased authenticity of integrated tasks (Douglas, 2010; Slaght & Howell, 2007). Indeed, based on analysis of the target language use domain in higher education, several language testing bodies have introduced integrated tasks:

The Test of English for Education Purposes (TEEP) contains an integrated reading-listening-writing task, where candidates use the reading and listening parts of the test as sources of ideas for an academic essay (International Study and Language Institute, 2013); the Canadian Academic English Language (CAEL) Assessment, in which students base their essays on three reading passages and one lecture on the same topic (CAEL, 2013); and the speaking and writing sections of TOEFL have been designed to integrate information from the reading and listening tasks (The TOEFL iBT Test, 2013).

In addition, since the assessment criteria which comprise rating scales 'act as the de facto test construct' (Knoch, 2011), in the EAP context they should simulate the criteria applied by subject specialists in university settings. There is some evidence that for subject tutors content takes priority over language. For example, a study by Errey (2000) used introspective verbal reports to investigate the extent to which five university lecturers from five different departments were influenced by language errors in student essays. The analysis of the verbal reports showed that *content* and *use of sources* were ranked as the most significant factors that lecturers took into account when assessing students' work, followed by *coherence* and *organisation*. Out of 24 factors, *grammatical accuracy* and *accurate vocabulary* were ranked 10th and 18th respectively. However, the study also revealed inconsistencies in the ways lecturers actually dealt with language errors, with some lecturers ignoring or tolerating them, while others deducted marks for linguistic inaccuracies. This conclusion seems to support an observation by Hartill (2000) that the linguistic quality of students' essays is 'left very much to the individual tutor's discretion' and that in many university departments there is a tendency for less strict emphasis on grammatical accuracy.

2.1.4 Untimed out-of-class and timed in-class assessments

Another important aspect of performance assessments is the time allotted and the place where the assessment is carried out. Weigle (2002, p.173) distinguishes between in-class timed and out-of-class untimed writing assessments. The purpose of in-class timed writing is to test whether students can plan and write an essay on their own without recourse to external sources such as textbooks, dictionaries, or advice from peers and tutors. On the other hand, out-of-class untimed writing in academic settings aims to provide direct experience of the stages of the essay writing process such as

analysing the essay question, finding and evaluating sources, organising notes, drafting, receiving feedback, redrafting and editing (Weigle, 2002, p.174). Thus the completion of one extended writing assignment may take several weeks as opposed to the 40 to 60 minutes normally allocated in in-class timed tests. Two studies in the 1990s compared student performance on in-class timed and out-of-class untimed writing tasks. Caudery (1990) investigated the validity of timed essays, which, he felt, did not measure the ability to write extended texts required in academic or occupational settings. The study compares the performance of 24 secondary school students preparing for GCE O-level examination in English Language in Cyprus. The students produced two pieces of writing – a timed in-class essay written within a 40-minute time limit, and an out-of-class untimed essay, which they started in class and finished at home over a 2-day period. The study found no evidence that time affected the overall scores of the students. Among the explanations given, the author mentions the negative washback of preparing for timed writing and the age of the students, who, in his view, lacked maturity to take advantage of the increased amount of time. He speculates that a study involving adult students may yield different results. Kroll (1990) investigated the performance of first-year undergraduate foreign students at an American university on four writing tasks. Two of the tasks were timed in-class essays written within a 60-minute time limit and the other two were take-home essay written over a period of 10-14 days. This study also failed to find support for the hypothesis that more time will result in statistically significant differences between the in-class and out-of-class tasks, despite some improvement on the syntactic and rhetorical level. The explanations offered echo the ones by Caudrey (1990), namely students may be unaware of what 'constitutes good writing' and therefore were unable to make good use of the extra time. Moreover, Kroll (1990) points out that they had not been taught about the process of writing and concludes that explicit instruction is required in order to guide students in their development as competent academic writers. Therefore, students are likely to benefit from out-of-class untimed writing only if they have awareness of the academic genres relevant to their discipline and the most effective methods to approach academic writing tasks.

2.1.5 Product-oriented and process-oriented assessments

It is clear from the preceding discussion that the way writing is taught and assessed in EAP courses has direct relevance to student attainment at university. Traditionally, a distinction has been made between the product and process approach to teaching academic writing. Jordan (1997, p.164) describes the product approach as 'concerned with the finished product – the text'. The central concept of this method is the presentation of a model text whose characteristics are analysed with the aim of producing a parallel text. The process of analysis also includes a range of language functions such as explanation, definition, exemplification to name but a few. The product approach has been used to familiarise students with the wide range of academic genres they will encounter at university: 'essays, reports, case studies, projects, literature reviews, exam answers, research papers/articles, dissertations and theses' (Jordan, 1997, pp. 165-166).

On the other hand, the process approach focuses on the composing process which writers go through. It views writing as problem solving, which begins with a *thinking* stage that involves generating, grouping together and selecting relevant ideas. This is followed by the *process* stage that involves a cycle of writing a draft, receiving peer or tutor feedback, revising the draft and so on until satisfactory completion of the task (Dudley-Evans & St John, 1998, p. 117).

In terms of assessment, the majority of large scale tests such as IELTS, TOEFL and the Cambridge ESOL exams are product-oriented. Candidates are advised to plan and check their essays but the short time limit and lack of feedback from peers or tutors preclude revision, which means that such essays can only be considered first drafts (Weigle, 2002, p. 176). There have been some attempts to introduce process-oriented writing assessments. Weigle (2002, pp. 191-192) cites an example of a test developed at Georgia State University, where students receive a grade on how well they have edited and revised their final draft in the light of peer and tutor feedback. The grade was added to the assessment criteria in order to encourage students to adopt the process approach to writing, which according to Hartill (2000) is particularly appropriate in higher education settings.

Another study by Cho (2003) examined student performance on two placement tests based on either a product-oriented or a process-oriented approach to assessing writing. The first is a timed integrated listening-reading-writing task, where the examinees listen to a lecture, read an article and write an essay within a 70-minute time frame. The second is a two-day test, which the author describes as a 'work-shop based essay', whose aim is to simulate as closely as possible the real-life conditions in which academic writing is produced. On day one, the examinees listen to a lecture, read an article, engage in discussions and produce a first draft. On day two, the students read the drafts of two of their peers and give them feedback. After that they have 60 minutes to write their final draft taking into account the comments they received. Only the final draft is marked. The results show that, overall, the workshop essays received higher scores and 75% of the test-takers were placed in a higher level group. Also, comparison of the textual quality of the two essays revealed that the workshop essays ranked higher on all features excluding 'source attribution'. Organisation, content and source use showed the most improvement, while linguistic features such as grammar, vocabulary and sentence variety were least affected by the test method. These two examples demonstrate that it is feasible to design both timed and untimed assessments incorporating the process approach to writing, which is highly relevant in academic contexts.

2.1.6 Criterion-related validity

The concept of validity in language testing has undergone a marked change since the 1960s from being viewed as a property of the test to the current understanding of validity as a unitary concept (Chapelle, 1999). Messick (1980) defines validity as 'an overall evaluative judgment of the adequacy and appropriateness of inferences drawn from test scores'. Moreover, he argues that construct validity is 'a unifying concept of validity', which can be described as having two facets: (a) 'the source of justification' of the testing and (b) 'the function or outcome of the testing'. He also proposes two sources of justification – evidential and consequential, and two outcomes – interpretation and use (Messick, 1980). Chapelle (1999) explains that in terms of outcomes, tests allow us, on the one hand, to make inferences about the underlying abilities of the candidates (interpretation), and, on the other hand, to make decisions about the candidates on the basis of their test results (use). Furthermore, the labels 'evidential' and 'consequential' sources of justification pertain to the arguments

required to justify the testing outcomes (Bachman,1990, p.243) Messick (1995) proposes six aspects of construct validity, namely content, substantive, structural, generalizability, external, and consequential, and specifies the sources of evidence for each of them. What is of particular interest for this study is the external aspect of construct validity which aims to establish the relationship with other measures so that 'the constructs represented in the assessment should rationally account for the external pattern of correlations' (Messick, 1995). Thus evidence of criterion-related validity will support an argument for the construct validity of a test.

Criterion-related validity can be defined as 'the extent to which a person's performance on a criterion measure can be estimated from that person's performance on the assessment procedure being validated' (Salvia & Ysseldyke, 2001, p. 152). The criterion chosen can be one of a number of external variables such as a syllabus, teachers' judgements, performance in the real world, or another test (Davies et al, 1999). The criterion-related validity of an assessment procedure typically includes concurrent and predictive validities, which are determined statistically by using correlation.

The aim of concurrent validation is to investigate the relationship between an assessment procedure, e.g. a test, and another measure for the same group of test-takers obtained at approximately the same time (Weir, 2005, p. 207). In language testing, a well-established standardised test such as IELTS or TOEFL can be used as a criterion. However, Davies et al (1999, p. 30) caution that the test being validated may define the construct of ability differently. Therefore, such a test may correlate poorly with the standardised test chosen as criterion.

Predictive validity refers to how accurately a test score predicts performance on the criterion measure at some future point in time (Weir, 2005, p. 209). For example, the scores on an EAP test, whose aim is to assess students' readiness for university study, should correlate highly with the academic performance of the same students as measured by their grades on subsequent academic courses (Davies et al, 1999, p.149).

A number of predictive validity studies both of standardised and locally developed tests have been conducted to investigate the relationship between language proficiency

and academic achievement. Predictably, well-established tests such as IELTS and TOEFL, which undergo a rigorous test production process, have commissioned both internal and external validation research in an attempt to support the claims that they provide an adequate measure of candidates' readiness to 'begin studying or training in the medium of English' (IELTS, 2012) and their 'ability to use and understand English at the university level' (The TOEFL iBT Test, 2013).

2.1.7 Predictive validity of large-scale EAP tests

Predictive validity studies of language proficiency as measured by IELTS have produced inconclusive results. In most studies, the criterion measure of choice is the Grade Point Average (GPA) though some use first term or first year GPA to minimise the effect of the moderator variable 'elapsed time' (Cope, 2011). Some studies report significant, positive, but weak relationship between IELTS scores and GPA (Feast, 2002), or moderate correlation between overall IELTS scores and first semester GPA (Woodrow, 2006; Yen & Kuzma, 2009). However, some studies found either no positive correlations between IELTS scores and GPA (Cotton & Conrow, 1998), or little evidence for 'the validity of IELTS as a predictor of academic success' (Dooey & Oliver, 2002). On the other hand, Ingram and Bayliss (2007) investigated the relationship between IELTS scores and three different dependent variables: student perceptions of their own language proficiency, researcher ratings and staff observations of the students' language behaviour in their academic courses. They found that IELTS scores accurately predicted the students' language behaviour in the four skills during the first six month of university study.

The extent to which the IELTS sub-tests correlate with the selected criterion has also been explored by predictive validity studies. Cotton and Conrow (1998) found that the reading and writing scores correlated positively with staff ratings, while in Kerstjens and Nery's (2000) study only the reading score showed a significant positive correlation with the first-semester GPA. On the other hand, Yen and Kuzma (2009) provide evidence for significant correlations between the listening, reading and writing scores and the first-semester GPA among a group of Chinese students. However, Woodrow (2006) found positive correlations between all scores and first-semester GPA with the exception of reading. What is more, the level of English language

proficiency appeared to influence the achievement of students scoring 6.5 or lower, whereas with students scoring 7 and above, English proficiency has no effect on academic performance (Woodrow, 2006).

TOEFL has also been subject to predictive validity studies. Cho and Bridgeman (2012) found small predictive validity correlation coefficients between TOEFL iBT and GPA with average weighted correlations r=.16 for graduates and r=.18 for undergraduates across four broad disciplines – business, humanities and arts, science and engineering and social sciences. However, they also showed that when students' TOEFL iBT scores were in the top 25%, they were twice as likely to be in the top 25% GPA group as students in the bottom 25% TOEFL group.

Although IELTS and TOEFL seem to be the most widely recognised English language examinations for academic study, universities in English-speaking countries recognise a range of English language qualifications. A study by Oliver, Vanderford, and Grote (2012) investigated the predictive validity of 25 types of proficiency evidence accepted by one Australian university. The authors conclude that although English language proficiency tests have low predictive validity overall, standardised tests such as IELTS and TOEFL are the best indicators of future academic achievement. Still, the Pearson correlations between the IELTS scores and the Weighted Average Marks (WAM) were low. For undergraduates, the only significant correlation was between the IELTS reading sub-test and WAM (r=.27), while for postgraduates, there was a weak but significant relationship between WAM and the overall score and three sub-test scores with the exception of writing (the correlations ranged from r=.16 for speaking to r=.29 for reading).

The inconclusive results may be due to several factors. First, many of the studies are based on small sample sizes – 101 (Feast, 2002), 33 (Cotton, F., & Conrow, 1995), 65 (Dooey & Oliver, 2002), 113 (Kerstjens & Nery, 2000), 61 (Yen & Kuzma, 2009). Second, the criterion in each study is different. Third, participants included both undergraduate and postgraduate students from a range of disciplines. All of these make it difficult to generalise about the effect of proficiency on academic success.

2.1.8 Predictive validity of small-scale EAP assessments

The guidelines adopted by the European Association for Language Testing and Assessment recommend that language testers should provide evidence for the validity of the assessment instruments they develop (EALTA, 2006). Such evidence necessarily includes predictive validity studies and some researchers have examined the predictive power of locally-developed EAP performance tests. Cope (2011) focused on task-based assessments closely linked to the students' future academic programs at an Australian university. Black (1991) used the scores attained in two academic skills courses as the predictor variable: EASL 140, which focused on speaking skills, includes presentations, vocabulary, listening comprehension, and pronunciation, and EASL 143, which aimed to develop writing skills and includes library research techniques, writing term papers, and problematic grammatical structures. Lee (2005) analysed an EAP placement test which is described as 'a performance assessment that requires students to summarise and integrate content from articles and lectures'. The test also promotes a process approach to writing and involves pre-writing, receiving feedback, re-drafting and using a word processor.

These predictive validity studies produced a range of correlation coefficients. Cope (2011) reports statistically significant correlations between the task-based scores for two 'direct entry' programmes and GPAs of r=.361 and r=.418, and one correlation that did not achieve significance of r=.156. He interprets the statistics as showing weak to substantial relationships, where correlations of r=0.40–0.70 are considered substantial. Black (1991) found small to modest correlations between the skills scores and the students' GPA. The author notes that even the strongest correlation (r=.392, $p<.05$) for one of the groups only explains 15% of the variance, which means that the rest, or 85% may be the result of other factors, e.g. adaptability, motivation, organisational skills. In Lee's (2005) study, the correlation coefficients between the placement performance test and first semester GPA varied. It was positive for Business (r=.275) and Humanities (r=.350), but negative for Life Sciences (r= –.548) and Technology (r= –.213).

Robison & Ross (1996) examined the relationship between English language proficiency and performance on academic research tasks. They studied the extent to which an English language placement test and an indirect Library Skills Research Test

predicted performance on a direct Library Skills Research Test, which represents an authentic task for university students. The results showed that the language test alone was a weak predictor of 'the actual research skills of the EAP students'. However, performance on the direct Library Skills Research Test could be better predicted by a combination of a language proficiency test and an indirect Library Skills Research Test, which simulates the task of academic research.

On the other hand, Cho's (2003) investigation of the predictive validity of the 'work-shop based essay', which aimed to simulate as closely as possible the real-life aspects of academic writing, showed that its predictive power was not better than that of the timed essay when correlated with faculty evaluation of the students. As Cho (2003) notes, relying on the workshop test for placement purposes leads to more false positives, i.e. students are considered to have ability when they do not, whereas using the timed essay scores results in more false negatives, i.e. students are considered not to have ability when they do. As neither of these results is satisfactory, the author cautions against drawing inferences about test-takers' abilities based on a single piece of writing. As with the predictive validity studies of large-scale EAP tests, it is difficult to generalise from such a variety of assessment instruments, each defining EAP proficiency differently, using different criterion measures, and including different sample sizes and participants from various academic fields.

Moreover, other factors besides language proficiency are likely to contribute to academic attainment. Personal background, academic background, current teaching and support (Feast, 2002), motivation, learning strategies, quantitative skills (Cho & Bridgeman, 2012), intellect and acculturation (Cope, 2011) may affect achievement at university.

2.1.9 Summary of the literature review

In summary, strong performance assessments tend to operationalize a task/context focused construct, employ indigenous marking criteria, which might include non-linguistic factors, are often untimed in the sense that the assessment may be carried out over a longer time period – from a few hours to several weeks, and typically emphasise a process approach to task completion. On the other hand, weak performance assessments tend to operationalize a trait/ability focused construct based

on theories of language proficiency, and consequently make use of assessment criteria that are primarily linguistic, are timed in the sense that test-takers usually have 30-45 minutes to complete one writing task, and are product-oriented since there is no opportunity for the candidates to implement the process approach to writing except in a relatively superficial form and it is the final product, e.g. an essay or a report, that is assessed.

The aim of the present study is to compare the predictive validity of three second language performance assessments – two weak and one strong – in the context of an international college which offers foundation programmes to overseas students. One of the weak performance assessments is IELTS, which is the examination of choice for the majority of students attending the college, and the second one is an in-house language test. The strong performance assessment is also locally developed and consists of two end-of-course assignments. Detailed information about the assessments will be provided in the Methodology section.

There are four main reasons that make this study worthwhile. First, testing organisations need to provide evidence for the construct and consequential validity of the assessments they produce. Predictive validity studies contribute to the formulation of a convincing argument for the construct validity of assessments. Second, the results of the study may lead to improvements in the language and study skills courses at the college. Third, the study will address some of the limitations of previous predictive validity studies such as small sample sizes, range restriction, and heterogeneity of population. Fourth, the college presents an ideal opportunity to compare the predictive validity of strong and weak performance assessments for the same group of students. This is important as empirical evidence is necessary to ensure fairness of test use in educational settings where decisions based on test scores can have far-reaching implications for the individual test-taker.

2.2 Research context

The College is a private pathway provider in partnership with a UK university. It is part of a wider network of colleges, which offer preparation courses for international students who wish to pursue a bachelor's or a master's degree in the UK. At the

College, students can enrol on Foundation Certificate and Graduate Diploma courses either in Business, Law and Social Sciences, or Science and Engineering. The present study will focus the Foundation Certificate in Business, Law and Social Sciences.

From September 2010 to September 2013 the Foundation Certificate in Business, Law and Social Sciences offered two pathways which consisted of eight modules (see Figure 3).

Programme Structure and Pathways

Business Pathway

Term 1	Term 2	Term 3
Introduction to Business	Economics in an International Context	Business & the Business Environment
Skills for Study 1	Statistics for the Social Sciences	Business Enterprise
Language for Study 1	Skills for Study 2	Skills for Study 3
	Language for Study 2	Language for Study 3

Law and Social Sciences Pathway

Term 1	Term 2	Term 3
Introduction to Social Science	Economics in an International Context	The Individual, State and Society
Skills for Study 1	Statistics for the Social Sciences	Introduction to Legal Principles and Systems
Language for Study 1	Skills for Study 2	Skills for Study 3
	Language for Study 2	Language for Study 3

Figure 3. Programme structure and pathways at the College

Students with an IELTS overall score (or equivalent) below 5 take a one-term pre-sessional course in English language (12 weeks), after which they can start on the BLSS programme.

All subject and skills courses are credit-bearing and are included in the Final Academic Score. In order to complete the programme, the students need a Final Academic Score of 40% or higher. However, in order to progress to the host university, the students need to achieve a Final Academic Score of 60% or higher. In addition, students must pass an exit proficiency test with an overall score of 65% or higher with minimum 55% in each component – reading, writing, listening and speaking.

The Language for Study modules are non-credit bearing and therefore not included in the Final Academic Score. Language for Study 1 is taken by three-term students, whose English language proficiency is generally lower, with an IELTS entry test score (or equivalent) of 5. Language for Study 2 and 3 are taken by all 3-term and 2-term students with IELTS scores (or equivalent) of 5.5 to 6.5. Students whose IELTS entry score (or equivalent) is 6.5 or higher are exempt from language classes and do not take the exit proficiency test.

The Skills for Study and Language for Study modules have different objectives. The skills classes, as their name suggests, aim to equip students with the academic skills required for university study, ranging from time management and research skills to applying the process approach when writing extended essays and reports, and when preparing for presentations and group discussions. Skills for Study 1, in particular, is a foundation course during which students are introduced to academic conventions such as the use of academic sources, citing and referencing, the structure and organisation of academic essays and presentations. Through tutor and peer feedback, they learn how to improve their work by analysing the assignment question, maintaining the focus of their essay or presentation, and providing relevant arguments and data in support of their thesis statement. On the other hand, Language for Study 1 aims to introduce students to the grammar, vocabulary and functional language that they need in order to be able to complete the subject module assignments, for example, the use of the infinitive to express purpose, or signposting words and phrases that can be used in a presentation.

2.3 Research questions

a) In the college setting, to what extent can IELTS be considered an accurate predictor of the academic achievement of international students as measured by their Final Academic Score?
b) In the college setting, to what extent can the Language for Study 1 exam be considered an accurate predictor of the academic achievement of international students as measured by their Final Academic Score?
c) In the college setting, to what extent can the Skills for Study 1 assessments be considered an accurate predictor of the academic achievement of international students as measured by their Final Academic Score?

3. METHODOLOGY

3.1 Participants

In total 570 international students from over 40 different countries completed the Foundation Certificate in Business, Law and Social Sciences 2-term and 3-term programmes at the College between January 2012 and August 2013.

3.2 Instruments

The following instruments are used to measure the students' performance: (1) IELTS Overall, Listening, Reading, Speaking, and Writing scores; (2) Language for Study 1 (LS1) Raw scores, Reported Scores, Listening & Speaking sub-test, and Reading & Writing sub-test scores; (3) Skills for Study 1 (SS1) Total, Essay, and Presentation scores; (4) Final Academic Score – the average of the eight subject and skills modules; and (5) Final Academic Score minus Skills for Study 1 score (FAS-SS1).

3.2.1 International English Language Testing System (IETLS)

On enrolment students are required to present evidence of English language proficiency. IELTS is by far the most popular certificate submitted by students at the College. The IELTS test consists of four sub-tests: Academic Reading (60 minutes, 40 questions), Academic Writing (60 minutes, 2 tasks), Listening (approximately 40 minutes, 40 questions), and Speaking (up to 14 minutes, 3 parts). Marks are awarded in half bands from 0-9. Although no data is available, typically students take the IELTS test up to six months before arriving at the College.

3.2.2 Language for Study 1 (LS1)

The LS1exam is taken at the end of term 1 by 3-term students only. It is an example of weak second language performance assessment (McNamara, 1996) and consists of two sub-tests (see Appendices I and II for samples):

a) Essay – an in-class, timed, integrated reading-and-writing task. Students are allowed 60 minutes to read 3 short extracts and use them as sources for a 250-word essay. They must paraphrase and cite the sources, although citation skills are not assessed. The marking criteria include: Task Achievement, Range of

Language, Accuracy of Language, and Cohesion and Coherence measured on a 0-100 scale.

b) Oral exam – an in-class, timed, integrated listening-and-speaking task. Students listen to two short extracts, take notes, and summarise the speakers' opinion on two questions. After that, they answer up to 5 additional questions. Assessment criteria – Task Achievement, Fluency and Coherence, Language Range and Accuracy, and Pronunciation measured on a 0-100 scale. The exam lasts approximately 10 minutes.

3.2.3 Skills for Study 1 (SS1)

The SS1 assessments are submitted at the end of term 1. The assessment procedure is an example of strong second language performance assessment (McNamara, 1996). It consists of two tasks (see Appendices III and IV for samples):

a) Essay – out-of-class, untimed. Students are given an essay question accompanied by instructions. They are expected to analyse the essay question, identify appropriate sources by using the online catalogue of the host university library, skim, scan, read for detail, take notes, and write at least two drafts. They must paraphrase, cite and reference correctly by using the APA referencing style. The final draft of 750 words is submitted to Academic Services and an electronic copy is uploaded to TurnItIn to be checked for plagiarism. The assessment criteria include: Content, Structure, Support, and Clarity of Expression measured on a 0-100 scale. There are penalties for plagiarism (the final mark of the re-submitted essay is capped at 75% of the awarded mark) and for late submission (see Table 1). In addition, tutors can deduct up to 5 points for failure to follow assignment guidelines on presentation and formatting.

b) Paired Presentation – out-of-class, untimed research and preparation, but delivered in class within a 10-minute time limit. The topic is the same as the essay. The students must use PowerPoint, cite and reference correctly. The assessment criteria include: Content, Structure, Support, and Clarity of Expression and Delivery measured on a 0-100 scale.

Table 1. *Penalties for late submission.*

Number of working days late	Penalty Awarded
1	85% of original mark
2	80% of original mark
3	75% of original mark
More than 3	Zero mark awarded

3.2.4 Final Academic Score

An important aspect of both concurrent and predictive validity is that criterion measure should itself be valid (Salvia & Ysseldyke, 2001, p. 153). Hartnett and Willingham (1980) examine the pros and cons of a range of measures of academic success and comment that overall GPAs 'seem to represent a good composite of whatever kinds of academic performance are reflected in grades'. Therefore, in this study the Final Academic Score that the students are awarded at the end of their 2- or 3-term programme will be used as the criterion measure.

The Final Score is the average of the eight skills and subject modules. The students need 40% to complete the programme, but 60% to be able to progress to the host university. The subject module assessments comprise a mixture of 1500-to-2500-word essays, reports and business plans, and exams consisting of 10-20 MCQs plus 2-3 questions on case studies for which the students write 1-to-2-page answers. The assessment criteria include Research, Content and Argument, Structure, Coherence and Clarity, Presentation and Referencing. In the exam answers subject tutors look for factually correct content, analysis, critical discussion, and relevant support. Language is not part of the assessment criteria and grammar and vocabulary mistakes only affect the final mark if meaning is obscured.

3.2.5 Final Academic Score minus Skills for Study 1

Since Skills for Study 1 is part of the Final Academic Score, a Final Score minus SS1 was calculated to ensure that the two variables are independent of each other.

3.3 Data Collection and Procedures

3.3.1 Data collection

Electronic copies of the students' scores were obtained from the Academic Services office. As the record-keeping system changed in 2012, hard copies were obtained for the Language for Study 1 and Skills for Study 1 scores recorded on the old system, and were added to the spreadsheets.

The Final Academic Score (FAS) is available for all 570 students. The English Language Entry scores consist of 452 IELTS overall scores and 449 sub-scores. This is because 118 students either took other exams such as TOEFL iBT, PTE Academic, IGCSE, or their scores are missing. As some students were exempted from Language for Study 1 classes, only 279 LS1 total scores are available. However, 65 of those students took the new LS1 exam introduced in January 2013 and were therefore excluded from the analysis. The LS1 scores are rounded up or down (see Table 2) following the same convention as outlined in the IELTS Quality and Fairness Brochure (IELTS, 2012) and only the rounded scores are reported to the students. In this study the total LS1 Reported scores are used.

Table 2. *Rounding of LS1 raw average scores*

Raw score	**Reported score**	**Raw score**
≥ 37.5	**40**	< 42.5
≥ 42.5	**45**	< 47.5
≥ 47.5	**50**	< 52.5
≥ 52.5	**55**	< 57.5
≥ 57.5	**60**	< 62.5
≥ 62.5	**65**	< 67.5
≥ 67.5	**70**	< 72.5

For some students the LS1 total scores in the spreadsheets were wrong. Instead of calculating the average score of the Listening/Speaking and Reading/Writing, the sub-tests were weighted 40% and 60% respectively. However, this resulted in only 3 students being misclassified (e.g. awarded 60 instead of 55), who were therefore excluded from further calculations. In addition, another student, who received 0 for the Listening/Speaking sub-test, most likely because of absence, was excluded. Overall, 210 LS1 total scores, Listening/Speaking and Reading/Writing component scores are available. Out of 570 students, 569 SS1 total scores are available (1 student received 0 for their essay, but since it is not clear whether they resubmitted, their score was excluded). Out of these 569 scores, 563 consist of a total score plus essay and presentation sub-scores.

The two correlation coefficients most commonly used in correlation studies are the Spearman rank order correlation coefficient and the Pearson product-moment correlation coefficient. While both coefficients show the strength and direction of the relationship between two variables, they are underpinned by different assumptions regarding the normality of distribution. The Pearson product-moment correlation coefficient (r) is a parametric statistic, which assumes that the data is normally distributed. On the other hand, the Spearman rank order correlation coefficient (r_s) is a non-parametric statistic, which does not assume normal distribution of the data (Bachman, 2004, pp. 89-95). Another difference is that the Pearson (r) assumes that the two variables are continuous, while Spearman (r_s) assumes that the variables are rank orders, for example ratings of speaking and writing skills (Green, 2013, p.71). Finally, both the Pearson (r) and Spearman (r_s) work on the assumption that that data is linear, i.e. on a scatterplot it will appear as a more or less straight line, and independent, i.e. each score 'must have been generated in ways which could not have influenced the generation of the other' (Brown & Rodgers, 2002, p. 182). Although the Pearson product-moment and the Spearman rank correlation coefficients may produce different results if applied to the same data set, their interpretation is basically the same – they show the magnitude and direction of the relationship between the variables in question (Bachman, 2004, p.91). Cohen (1998, cited in Pallant, 2007, p.132) suggests the following interpretation of the strength of the relationship between two variables:

Small	r=.10 to .29
Medium	r=.30 to .49
Large	r=.50 to 1.0

3.3.2 Procedures

Data analysis was performed using SPSS Version 19. Box plots were generated to identify outliers and extreme cases. SPSS defines outliers as points that 'extend more than 1.5 box-lengths from the edge of the box', and extreme cases as points that 'extend more than three box-lengths from the edge of the box' (Pallant, 2007, p. 63). The outliers are represented by a small circle and extreme cases by an asterisk. There is no clear-cut answer as to whether to delete or include the outliers in the data analysis. Bachman (2004, p. 102) argues that deleting outliers is only justified if there are reasonable grounds to assume that the scores do not accurately represent the abilities measured by the test. He recommends 'going back through the data processing trail' in order to check if the scores are genuine and if administrative records can provide any explanation for extremely high or low scores. For example, in task-based EAP assessments low scores may be due to penalties incurred for plagiarism, late submission or no submission. Furthermore, Larson-Hall (2010, p. 91) points out that outliers are problematic for parametric statistical procedures as they assume normal distribution of the data, and therefore by deleting them we could be disregarding the fact that the distribution may be non-normal. Pallant (2007, p.129) recommends checking output from Explore for the values of the mean and the trimmed mean, i.e. the mean with the top and bottom 5% removed. If the values are very similar, the outliers may not be cause for concern.

Scatterplots for each data set were generated using SPSS and checked for linearity. In addition, the descriptive statistics and histograms were examined for normality of the distribution and for homogeneity of variance, which are assumptions for parametric tests (Pallant, 2007, p. 204). Normal Q-Q plots and Detrended normal Q-Q plots were also inspected. Bachman and Kunnan (2005, pp. 45-46) explain that in the normal Q-Q plot reasonably normally distributed scores should be closer to the diagonal line, which represents what the values would be if the distribution were normal. Also, in the Detrended normal Q-Q plots for a normal distribution the scores should cluster near the horizontal line. Finally, the Kolmogorov-Smirnov statistic was checked as a non-

significant result (Sig. > .05) indicates normality, while a significant result means that the assumption of normality has been violated (Pallant, 2007, p.62).

Correlation coefficients for each of the three sets of scores (including sub-scores) were calculated: IELTS and the Final Academic Score; LS1 and the Final Academic Score; SS1 and the Final Academic Score minus SS1.

Finally, histograms showing the distribution of the Final Academic Scores by IELTS band and LS1 band were produced and the results summarised in tables for easy reference. However, as the SS1 scores are not reported in whole or half bands, there are 106 separate S1 scores, which makes the procedure impracticable for reporting the distribution of the Final minus SS1 scores by SS1 score.

4. RESULTS

The following section presents the results of the data analysis. For each of the three sets of variables, the findings regarding the assumption of normality are described first as they determine the application of parametric or non-parametric measures. After that, the obtained correlation coefficients are presented.

4.1 IELTS & Final Academic Score

4.1.1 Box plots – outliers and extreme cases

Both box plots (see Appendix V) show a small number of outliers and the IELTS Overall Score box plot contains several extreme cases. However, there is some evidence that the scores accurately reflect the ability of the students so there is no justification for excluding them from further analysis. First, the IELTS mean score is 5.37, but the range is 5 (from 3.5 to 8.5). The majority of students studying at the College do not meet the host university English language entry requirements at the start of their programme. However, some students, particularly from Nigeria and Singapore, have high levels of English language proficiency, but for one reason or another do not meet the university academic requirements. The fact that IELTS has nine levels of competence – from non-user to independent user (IELTS, 2012) – means that in a pathway college it is not unusual to come across large differences in language proficiency. For this reason, there are 2- and 3-term courses. Moreover, IELTS has strict quality control procedures in place that ensure that the results are sufficiently reliable (IELTS, 2012). Therefore, the outliers and extreme cases will not be excluded from analysis.

4.1.2 Scatterplots

Examination of the scatterplot in Figure 4 shows that the relationship between the two variables is linear.

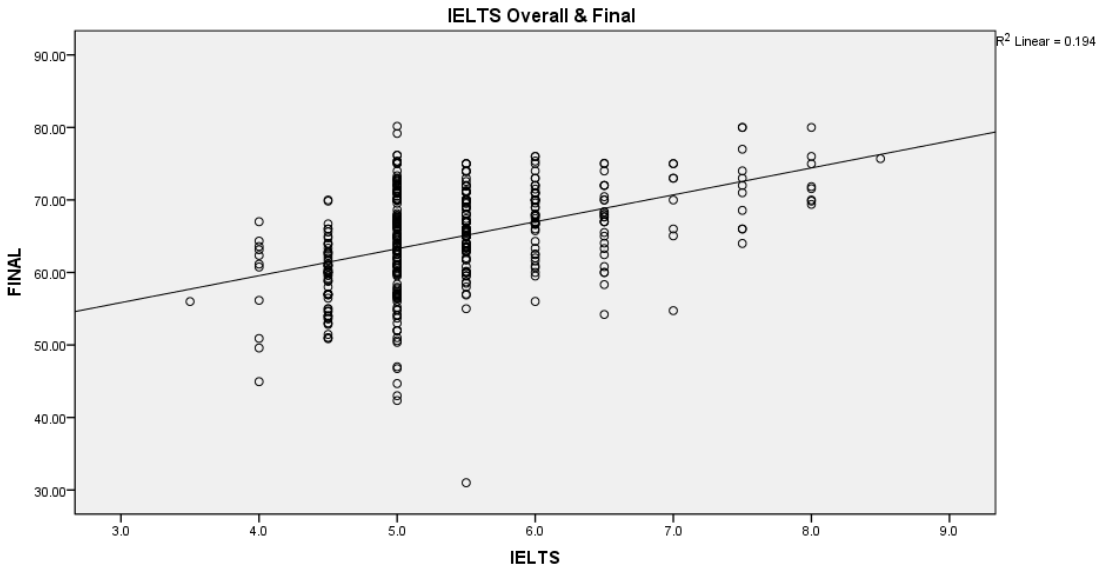

Figure 4. Scatterplot of IELTS and Final Academic Scores

4.1.3 Descriptive statistics

Table 3. *Descriptive statistics of IELTS and Final Academic Scores*

Statistic	IELTS	FINAL
N	452	452
Mean	5.37	64.65
Std. Error of Mean	.038	.317
Median	5.00	65.00
Mode	5.00	70.00
Std. Deviation	.80	6.75
Variance	.64	45.55
Skewness	1.238	-.615
$SE_{skewness}$	.115	.115
Skewness / $SE_{skewness}$	10.77	–5.35
Kurtosis	1.928	1.219
$SE_{kurtosis}$	.229	.229
Kurtosis / $SE_{kurtosis}$	6.45	4.08
Range	5.0	49.17
Minimum	3.5	31.00
Maximum	8.5	80.17

Table 3 shows that The IELTS and FAS skewness and kurtosis divided by their respective standard errors are outside the – 2 +2 range and therefore violate the assumption of normality. The histograms (see Appendix VI) also show that IELTS

scores are positively skewed, while the FAS scores are negatively skewed. The distribution of both sets of scores is leptokurtic.

4.1.4 Q-Q plots and test of normality

The Normal Q-Q plots and Detrended normal Q-Q plots generated for both IELTS and the Final Academic Score show deviations from the expected values if the distribution was normal (see Appendix VII).

A Kolmogorov-Smirnov test returned Sig. value .000 for IELTS overall and all subtests, and .010 for the Final Academic Score. A non-significant result (Sig. > .05) indicates normality (Pallant, 2007, p.62). Therefore, the distribution of IELTS and FAS scores is not normal.

The analysis presented above indicates that a non-parametric test will be more appropriate, therefore Spearman rank order correlations (r_s) were calculated for IELTS (including the Reading, Writing, Speaking and Listening sub-scores) and the Final Academic Score.

4.1.5 Non-parametric correlations

Table 4. Non-parametric correlations (Spearman r_s) between IELTS Overall, Reading, Writing, Speaking, Listening and the Final Academic Score

	Overall	**Reading**	**Writing**	**Speaking**	**Listening**
Reading	.740**	-			
	(n=449)				
Writing	.739**	.496**	-		
	(n=449)	*(n=449)*			
Speaking	.797**	.475**	.511**	-	
	(n=449)	*(n=449)*	*(n=449)*		
Listening	.855**	.605**	.561**	.638**	-
	(n=449)	*(n=449)*	*(n=449)*	*(n=449)*	
Final Academic	.467**	.397**	.401**	.313**	.453**
Score	*(n=452)*	*(n=449)*	*(n=449)*	*(n=449)*	*(n=449)*

** Correlation is significant at the 0.01 level (2-tailed).

Significant medium positive correlations were found between the Final Academic score and the IELTS overall score (r_s=.467, n=452, p< .01), Reading (r_s=.397, n=449, p< .01), Writing (r_s=.401, n=449, p< .01), Speaking (r_s=.313, n=449, p< .01), and Listening (r_s=.453, n=449, p< .01).

Since the students need to achieve a Final Academic Score of 60 and above to progress to the host university, the percentage of students who obtained such a score is presented in Table 5 by IELTS band.

Table 5. Percentage of students achieving a Final Academic Score of 60 and above by IELTS band

IELTS Band	**N=452**	**Minimum Final**	**Maximum Final**	**Range Final**	**% achieving 60+**
3.5	1	55.99	55.99	0	0%
4.0	11	44.94	67.00	22.06	64%
4.5	63	50.86	70.00	19.14	52%
5.0	175	42.33	80.17	37.83	77%
5.5	96	31.00	75.00	44.00	91%
6.0	50	56.00	76.00	20.00	96%
6.5	28	54.21	75.06	20.85	93%
7.0	8	54.74	75.00	20.26	87.5%
7.5	11	64.00	80.00	16.00	100%
8.0	8	69.39	80.00	10.61	100%
8.5	1	75.69	75.69	0	100%

Table 5 shows that although higher language proficiency (as measured by IELTS) generally results in a higher percentage of students achieving a final score of 60 and above, there is still a significant proportion of students with an IELTS score below 5.5 who complete their academic programme with a Final Academic score of 60+. For instance, 77% of the students who start their programme with an overall IELTS score of 5.0 progress to the host university. The percentage for students arriving with IELTS 4.5 and 4.0 is 52% and 64% respectively.

4.2 LS1 & Final Academic Score

4.2.1 Box plots – outliers and extreme cases

The box plot for LS1 Reported shows no outliers or extreme cases, but the box plot for Final Academic Score (FAS) shows two outliers and one extreme case – 155 (see Appendix VIII). Following the 'data processing trail' revealed that student 155 dropped out of college before the resubmission date after failing some of the modules, and was therefore excluded from further analysis.

4.2.2 Scatterplots

Examination of the scatterplot in Figure 5 shows that the relationship between the two variables is linear.

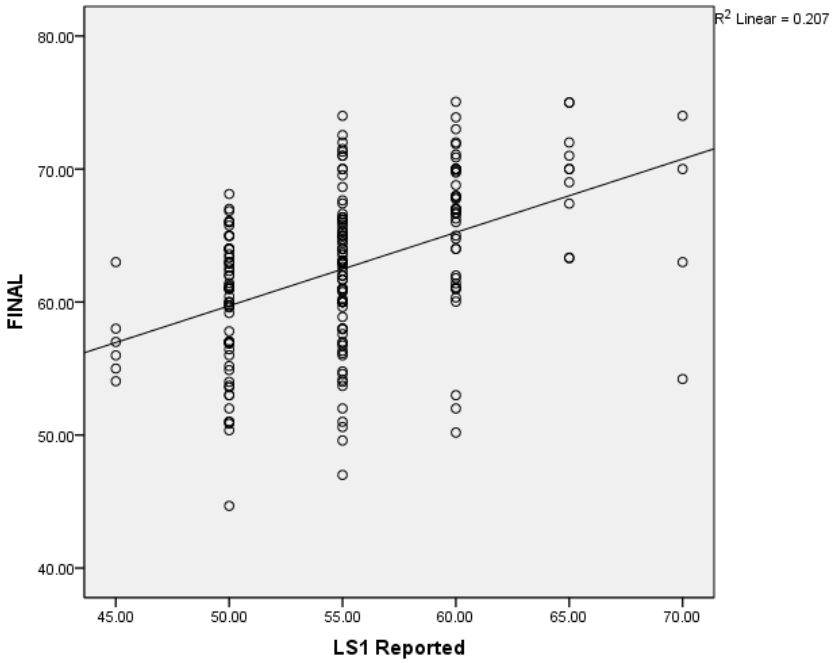

Figure 5. Scatterplot of LS1 Reported and Final Academic Scores

4.2.3 Descriptive statistics

Table 6. *Descriptive statistics of LS1 Reported and Final Academic Scores*

Statistic	LS1 Reported	FINAL	LS1L&S	LS1&W
N	209	209	209	209
Mean	55.05	62.50	55.19	54.82
Std. Error of Mean	.343	.416	.395	.330
Median	55.00	63.00	54.50	54.00
Mode	55.00	64.00[a]	51.00[a]	53.00
Std. Deviation	4.95	6.01	5.71	4.77
Variance	24.517	36.124	32.617	22.713
Skewness	.581	-.253	.758	.404
$SE_{skewness}$	.168	.168	.168	.168
Skewness / $SE_{skewness}$	3.08	–1.51	4.51	2.40
Kurtosis	.452	-.230	1.624	.302
$SE_{kurtosis}$	.335	.335	.335	.335
Kurtosis / $SE_{kurtosis}$	1.35	–0.69	4.85	0.90
Range	25.00	30.38	37.50	26.80
Minimum	45.00	44.68	41.50	42.00
Maximum	70.00	75.06	79.00	68.80

a. Multiple modes exist. The smallest value is shown

Table 6 shows that FAS skewness and kurtosis divided by their respective standard errors are within the –2 +2 range and therefore the scores are normally distributed.
LS1 Reported kurtosis divided by its standard error is within the –2 +2 range. However, LS1 Reported skewness divided by its standard error is not within the –2 +2 range. Furthermore, both LS1 L&S kurtosis divided by its standard error and LS1 L&S skewness divided by its standard error are not within the –2 +2. LS1 R&W kurtosis divided by its standard error is within the –2 +2 range. However, LS1 R&W skewness divided by its standard error is not within the –2 +2.

Therefore, the assumption of normality is violated for some of the variables. Visual inspection of Table 6 also shows that there is no homogeneity of variance.

4.2.4 Q-Q plots

The Normal Q-Q plots and Detrended normal Q-Q plots generated for LS1 Reported, LS1 R&W, and LS1 L&S show deviations from the expected values if the distribution was normal (see Appendix IX).

The analysis presented above indicates that a non-parametric test will be more appropriate, therefore Spearman rank order correlations (r_s) were calculated for LS1 Reported (including the Reading/Writing, Speaking/Listening sub-scores) and the Final Academic Score.

4.2.5 Non-parametric correlations

Table 7. Non-parametric correlations (Spearman r_s) between LS1 Reported, Reading & Writing sub-test, Speaking & Listening sub-test and the Final Academic Score

N=209	LS1 L&S	LS1 R&W	LS1 Reported	FINAL
LS1 R&W	.579**	-		
LS1 Reported	.839**	.823**	-	
FINAL	.342**	.501**	.479**	-

** Correlation is significant at the 0.01 level (2-tailed).

Significant medium positive correlations were found between the Final Academic score and the LS1 Reported score (r_s=.479, n=209, p< .01), and between the Final Academic score and the LS1 Listening & Speaking sub-test score (r_s=.342, n=209, p< .01). A significant large positive correlation was found between the Final Academic score and the LS1 Reading and Writing sub-test score (r_s=.501, n=209, p< .01).

Table 8. Percentage of students achieving a Final Academic Score of 60 and above by LS1 Reported score

LS1 Reported	**N=209**	**Minimum Final**	**Maximum Final**	**Range Final**	**% achieving 60+**
45.00	6	54.05	63.00	8.95	17%
50.00	61	44.68	68.13	23.45	56%
55.00	85	47.00	74.00	27.00	67%
60.00	43	50.19	75.06	24.87	93%
65.00	10	63.30	75.00	11.70	100%
70.00	4	54.21	74.00	19.79	75%

Table 8 shows that although higher language proficiency (as measured by LS1 overall reported score) generally results in a higher percentage of students achieving a final score of 60 and above, there is still a sizable proportion of students with scores of 50 to 55, who complete their academic programme with a score of 60+, which allows them to progress to the host university.

4.3 SS1 and Final minus SS1

As the students' Final Academic Score includes the Skills for Study 1 score, which may affect the correlation coefficient, a new FAS minus SS1 score was calculated. However, since significant large positive correlations were found between the Final Academic scores and the Final minus SS1 scores (r_s=.996, n=568, p< .01) with the correlation coefficient very close to 1, we can be certain that 'all that may be known about a person with regard to the one variable is perfectly revealed by the other variable' (Henning, 1987, p. 59), or in other words FAS and FAS minus SS1 measure essentially the same construct. Therefore, the Final minus SS1 score was used in assessing the predictive validly of the SS1 scores.

4.3.1 Box plots – outliers and extreme cases

Both box plots (see Appendix X) show a small number of outliers and one extreme case (432) among the Final minus SS1 scores. Following the 'data processing trail' revealed that 432 was the same student who dropped out of college before the

resubmission date after failing some of the modules, and was therefore excluded from further analysis. Individuals who receive low scores in SS1 or the subject modules usually do so for one of three reasons: (a) they committed plagiarism and their final mark was capped at 75% after resubmission, i.e. a student who would have obtained a score of 60, actually received 45; (b) they submitted their assignment late and as a result had their score reduced by 15%-25% in line with the regulations mentioned in 3.2.3; (c) they submitted work which was of poor quality. On the other hand, individuals with high scores typically submit work that is above average. Since SS1 is a performance assessment in the 'strong' sense, its criteria are similar to ones applied by the subject specialists at the College, which include penalties for plagiarism and poor time management skills. Thus, it can be said that the results accurately reflect the abilities of the students to complete the tasks and therefore the outliers should not be excluded from further analysis.

4.3.2 Scatterplot

Examination of the scatterplot in Figure 6 shows that the relationship between the two variables is linear.

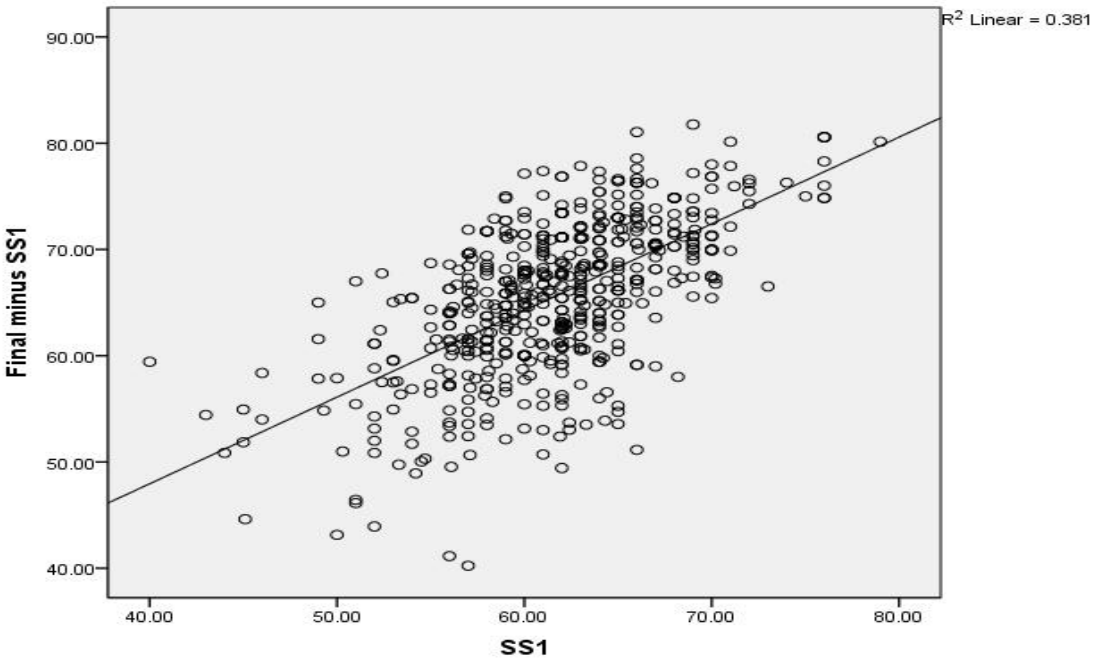

Figure 6. Scatterplot of SS1 and Final minis SS1 scores

4.3.3 Descriptive statistics

Table 9. Descriptive statistics of SS1 and Final minus SS1

	SS1	SS1E	SS1P	F-SS1
N	568	562	562	568
Mean	61.57	61.30	61.86	65.55
Std. Error of Mean	.226	.262	.228	.298
Median	62.00	61.50	62.00	66.14
Mode	62.00	60.00[a]	62.00	71.14[a]
Std. Deviation	5.38	6.22	5.40	7.10
Variance	28.896	38.630	29.155	50.440
Skewness	-.232	-.354	-.157	-.469
$SE_{skewness}$	.103	.103	.103	.103
Skewness /$SE_{skewness}$	-2.252	-3.44	-1.52	-4.553
Kurtosis	1.133	1.500	.544	.180
$SE_{kurtosis}$	.205	.206	.206	.205
Kurtosis /$SE_{kurtosis}$	5.527	7.28	2.64	0.878
Range	39.00	44.50	37.00	41.52
Minimum	40.00	35.50	42.00	40.24
Maximum	79.00	80.00	79.00	81.76

a. Multiple modes exist. The smallest value is shown

Table 9 shows that SS1 skewness and kurtosis divided by their respective standard errors are not within the –2 +2 range and therefore violate the assumption of normality. Although the Final minus SS1 kurtosis divided by its standard error is within the –2 +2 range, the result for skewness shows that the Final minus SS1 scores are negatively skewed. Moreover, there is no homogeneity of variance as the two statistics are very different.

4.3.4 Q-Q plots and test of normality

The Normal Q-Q plots and Detrended normal Q-Q plots generated for both SS1 and Final minus SS1show deviations from the expected values if the distribution was normal (see Appendix XI).

A Kolmogorov-Smirnov test returned Sig. value .004 for F-SS1 and .001 for SS1. As a non-significant result (Sig. > .05) indicates normality (Pallant, 2007, p.62), the distribution of scores for SS1 and Final minus SS1 is not normal.

The analysis presented above indicates that a non-parametric test will be more appropriate, therefore Spearman rank order correlations (r_s) were calculated for SS1 (including the essay and presentation sub-scores) and Final minus SS1.

4.3.5 Non-parametric correlations

Table 10. Non-parametric correlations between SS1 Overall, SS1 Essay, SS1 Presentation, and Final minus SS1

	SS1	**SS1Essay**	**SS1Presentation**
SS1Essay	.925**		
	(n=562)		
SS1Presentation	.823**	.579**	
	(n=562)	*(n=562)*	
Final Minus SS1	.593**	.529**	.529**
	(n=568)	*(n=562)*	*(n=562)*

**. Correlation is significant at the 0.01 level (2-tailed).

Significant large positive correlations were found between the Final Minus SS1 score and the SS1 Overall score (r_s=.593, n=568, p< .01), the SS1 Essay (r_s=.529, n=562, p< .01), and the SS1 Presentation (r_s=.529, n=562, p< .01).

5. DISCUSSION OF RESULTS

5.1 Research question 1

In the college setting, to what extent can IELTS be considered an accurate predictor of the academic achievement of international students as measured by their Final Academic Score?

The results of this study confirm the findings of Woodrow (2006) and Yen and Kuzma (2009), who also reported correlations between overall IELTS scores and 1st semester GPA of r=.40 and r=.46 respectively. One possible reason for the medium positive correlations may be the homogeneity of the sample. Studies that found no or little evidence of the predicative power of IELTS (Cotton & Conrow, 1998; Dooey & Oliver, 2002) focused on heterogeneous populations. For example, Cotton & Conrow's (1998) participants came from the schools of Science and Technology, Architecture, Engineering, Health science, Humanities, Social Sciences, Visual and Performing Arts, Commerce and Law. Dooey and Oliver (2006) recruited participants from three disciplines – business, science and engineering. However, different disciplines most likely require different levels of English language proficiency. Not surprisingly, the IELTS test score guidelines divide university courses into two categories: linguistically demanding and linguistically less demanding, and suggest different IELTS entry scores for each (IELTS, 2011, p.13). Indeed, Cotton and Conrow (1998) argue that the fact that 45% of the students in their study attended linguistically less demanding courses such as science and engineering weakened 'any link between language proficiency measures and academic outcomes'. For such students, mathematical ability, for example, may be more important for academic success than language skills. Moreover, when selecting the criterion measure it should be borne in mind that final grades may show substantial variations within and across disciplines (Hartnett & Willingham, 1980). The present study circumvents these limitations because, first, all the participants were foundation certificate students enrolled on the Business, Law and Social Sciences programme, and, second, as many of the subject tutors at the college teach on more than one course there is bound to be more agreement among tutors about what constitutes academic success and greater comparability of the assessment criteria. Therefore, it seems reasonable to assume that

the medium positive correlations reported above may partially be due to the relative homogeneity of the population.

Another contributing factor may be the wider range of IELTS scores in the present study – from 3.5 to 8.5. A truncated sample may show weak correlations and lead to incorrect conclusions about the relationship between language proficiency and academic achievement (Henning, 1987, p.66; Bachman, 2004, p.96). Indeed, many predictive validity studies emphasise the effect of range restriction on their results (Cotton & Conrow, 1998; Dooey & Oliver, 2002; Woodrow, 2006; Cho & Bridgeman, 2012). This is not surprising as universities only enrol students who have met their minimum English language entry requirements, commonly set at IELTS 5.5 to 7 for undergraduates at UK universities (Hayatt & Brooks, 2009, p14). An advantage of the current study is that the students attended a pathway programme **before** starting university and therefore were accepted even if their IELTS score was lower than 5.5. In fact, 250 out of the 452 students, or 55%, had an IELTS score of 5.0 or less on entry. As mentioned in section 2.2, such students are required to take a 12-week pre-sessional English language course and it could be argued that their level of English at the start of their academic programme was higher than when they took their IELTS test. However, two studies cast some doubt about the likelihood of quick score gain as a result of intensive study. First, Elder and O'Loughlin (2003) investigated whether attending a 10-12-week intensive English course had an effect on the IELTS scores of 112 international students in Australia and New Zealand. The found that the maximum score gain was 2 bands and the minimum -1, with an 'average overall gain […] slightly more than half a band'. The second study by O'Loughlin and Arkoudis (2009) examined score gain over a longer period. They compared the IELTS scores of 30 undergraduate and 33 postgraduate students before and after graduating from an Australian university. The results reveal that the highest average increase in scores was in Reading (0.532), followed by Listening (0.500), Speaking (0.444) and Writing (0.206). This suggests that the language proficiency of the students in the present study was unlikely to have been dramatically affected by their attending a one-term pre-sessional language course. Although we need to bear in mind that lower level students demonstrate greater progress in language learning than their more proficient counterparts within the same time period (O'Loughlin & Akkoudis, 2009), it is still

the case that in the present study there is less range restriction compared with the majority of the predictive validity studies discussed in the literature review.

Despite the fact that there were significant medium positive correlations between the Final Academic score and the IELTS overall and subtest scores, these are not accurate enough to allow for precise group predictions. As can be seen from Table 5 in 4.1.5, a substantial percentage of students who started their programme with an IELTS score between 4.0 and 5.5 managed to achieve a Final Academic score of 60% and above, which allowed them to progress to the host university. To account for this we need to examine the view put forward in the literature review that IELTS is a timed, product-oriented, weak performance assessment, which simulates real-life tasks with the aim of sampling a candidate's language ability and whose construct definition is ability/trait focused.

First, there are some major differences between the IELTS speaking construct and the oral assessments at the College. While the IELTS speaking tasks attempt to simulate real-life situations that students may encounter at university, such as giving personal information, making a presentation and stating an opinion, all of this is done within a 14-minute period (IELTS, 2011). The candidates are expected to use their general knowledge and discuss topics that have been carefully selected to avoid bias against any group of candidates (IELTS, 2012). The marking criteria are predominantly linguistic: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation (IELTS, 2013). The oral presentations in the college assessments, on the other hand, emphasise research skills, use of academic sources, preparation and practice. Moreover, the assessment criteria include non-linguistic factors such as appropriate use of gestures, eye contact, tone of voice and body language, in addition to content, logical structure and support from sources. When Yen and Kuzma (2009) found a small negative correlation between Speaking and 1st semester GPA, they suggested as a possible reason that the GPA scores were primarily based on written assignments. In the present study, oral assessments also form a small part of the Final Academic Score; in fact, there are only three oral assessments – one presentation for Business Enterprise, a group discussion for Skills for Study 2, and an oral presentation for Skills for Study 3. Therefore, the different abilities measured by IELTS Speaking and the oral assessments at the College and the relatively small number of the latter

might explain why the correlation between the Final Academic score and the IELTS Speaking sub-scores (r_s=.313) is at the lower end of the medium correlation range.

Furthermore, the IELTS writing construct has also been criticised as not being reflective of academic writing at university. A comparison of the marking criteria shows that while IELTS scripts receive a grade for task achievement, coherence and cohesion, lexical resource, and grammatical range and accuracy, subject module criteria at the College include research, content and argument, structure, coherence and clarity, and presentation and referencing. Moore and Morton (2007, p.228) conclude that despite some similarities between academic writing and the IELTS essay, the latter, with its emphasis on personal opinion, is 'suggestive of such public, non-academic genres as the letter to the editor or the newspaper editorial'. Leki (2011, p. 103) even argues that English language university screening tests like IELTS have had a negative washback leading to the creation of 'the monster of the English L2 essay exam genre', which reduces writing skills to learning certain formulas and structures in the hope that slotting them neatly together would guarantee success. The medium correlation coefficient (r_s=.401) that emerged from the present study may be seen as providing further evidence that the writing construct operationalized by IELTS is quite district from the construct of academic writing at university. The findings also appear to support Leki's (2011, p.106) view that once students are introduced to the conventions and expectations of university writing, they have 'no trouble setting aside the essay exam as, at best, marginal to their new writing tasks'.

Finally, IELTS is timed and product-oriented. Candidates are given 20 minutes to complete Writing task 1 and 40 minutes for Writing task 2. Although they are encouraged to spend a few minutes planning their response and a few minutes checking it at the end, it is clear that there is no time to produce more than one draft. This contrasts with the process approach adopted by the course tutors at the College, where students work on their assignments for most of the term, hand in detailed plans or first drafts, and receive feedback before formally submitting their final draft for marking.

As far as reading skills are concerned, the medium correlation (r_s=.397) between the IELTS Reading sub-scores and the Final Academic Score may be due to that fact that

in an IELTS exam the candidates are restricted not only by the time available but also to reading three pre-selected texts, while in academic settings they have access to a range of sources and even if they fail to understand some points in the text, they may still be able to identify and use appropriate arguments in their written assignments.

Among the four skills assessed by IELTS Listening correlates the highest with the Final Academic Score (r_s=.453) in the present study. This finding is somewhat surprising, although Yen and Kuzma (2009) and Woodrow (2006) found similar correlations of r=.45 and r=.35 respectively between IELTS Listening and 1st semester GPA. Obviously, listening is an important skill in lectures and seminars, so although it is not specifically assessed in the subject modules, it must contribute to the students' ability to engage with the course content.

Commenting on the meaningfulness of correlation coefficients, Hughes (2003, p.32) notes that 'if the coefficients between scores on the same construct are consistently higher than those between scores on different constructs, then we have evidence that we are indeed measuring separate and identifiable constructs'. This raises the question of whether the abilities measured by IELTS can form the basis for making inferences about the students' ability to complete the academic tasks required of them by subject tutors.

5.2 Research question 2

In the college setting, to what extent can the Language for Study 1 exam be considered an accurate predictor of the academic achievement of international students as measured by their Final Academic Score?

The Language for Study 1 exam, like IELTS, is a weak performance test whose construct is trait/ability focused. However, there is one major difference – the LS1 test contains integrated Reading/Writing and Listening/Speaking tasks. In this respect, it implements one of the recommendations put forward by Moore and Morton (2007, p. 235), namely the inclusion of source texts, which the test-takers read in order to identify relevant information which they must then incorporate in their response. Moore and Morton (2007) argue that such a change to the current IELTS writing task

will increase its authenticity as it is more representative of university writing. The results from the present study seem to support this view as the correlation coefficient between the Final Academic scores and the LS1 Reading & Writing sub-test scores (r_s=.501, n=209, p<.01) is higher than the correlation coefficient between the Final Academic scores and the IELTS Writing scores (r_s=.401, n=449, p<.01). Although the reading-to-write task assessment criteria are almost identical to the ones used to grade the IELTS essay task – Task achievement, Range of Language, Accuracy of Language, Cohesion and coherence – the Task achievement criterion includes the extent to which the arguments presented in the essay are supported by incorporating information from at least two of the excerpts provided. Students are assessed on their ability to paraphrase and penalised if they lift phrases from the source texts. Therefore, although the LS1 Reading & Writing task can be considered an instance of 'role-playing' (McNamara, 1996), the context provided by task is realistic and simulates some aspects of academic writing at university such as summarising, paraphrasing and using in-text citations**.**

On the other hand, the correlation between the Final Academic score and the LS1 Listening & Speaking task score, although significant, is at the lower end of the medium correlations (r_s=.342, n=209, p<.01), which is comparable to the correlation between the Final Academic score and the IELTS Speaking score (r_s=.313, n=449, p< .01). The fact that these two correlations are the lowest in the present study suggests that the Listening/Speaking construct in LS1 and the Speaking construct in IELTS may be under-representations of the traits and abilities that are important for successful oral performance in academic contexts. Indeed, the assessment criteria for the integrated Listening & Speaking tasks – Task Achievement, Fluency and Coherence, Language Range and Accuracy, and Pronunciation – strongly resemble the IELTS Speaking marking criteria. The only difference is the introduction of a Task achievement criterion, which evaluates the extent to which examinees correctly summarise the opinions of the speakers in the listening section. These criteria differ significantly from the ones used to assess academic presentations, which include Support, i.e. the appropriate use of academic sources and correct referencing, and Clarity of delivery, i.e. suitable body language, gestures, speech rate and tone of voice.

The LS1 exam is also timed and product-oriented. Students are given 60 minutes to complete the Reading & Writing task without sufficient time to revise and make alterations, so the final product can only be viewed as a first draft (Weigle, 2002; Hartill, 2000). The Listening & Speaking part lasts approximately 10 minutes. The recording is heard once only, and the students have little time to organise or think about their notes before they are asked to respond to the questions. In the interest of fairness to all examinees, the tutors are not allowed to rephrase the questions. Although this may enhance the reliability of the test, it hardly resembles what happens in real-life communication. In a lecture, if students miss a particular point, they may reasonably expect that it will be repeated in the summary at the end. And even if it is not, they can rely on the lecture notes or the coursebook. In a conversation, the speaker will most likely be willing to paraphrase and explain if their interlocutor does not understand.

The LS1 exam is modelled on large-scale tests such as IELTS, which, according to Weigle (2002, p.175) are mostly concerned with reliability and practicality, i.e. providing reliable scores within a short period of time to a large number of test-takers. The other aspects of test usefulness suggested by Bachman and Palmer (1996), namely construct validity, authenticity, interactiveness, and impact, are not neglected, of course, but Weigle (2002, p.175) argues that they are more likely to be prioritised by achievement assessments as teachers are more concerned with achieving the course objectives, preparing students for real-world tasks and ensuring they engage in the process of writing.

Overall, scores on the LS1 exam, and especially the reading-to-write task, seem a better predictor of academic success at the College than IELTS scores. However, it should be borne in mind that the sample is twice as small (n=209) as the IELTS sample (n=452), and it is truncated since all the students who take the course are selected because their entry IELTS score is below 6.5. The correlation coefficient may be different, if higher level students took the exam as well. The results also show a number of false negatives, i.e. some students who obtained low marks still achieved a score of 60+ after 1 or 2 more terms at the College. Therefore, it would be difficult to base individual predictions on the scores of the LS1 exam only.

5.3 Research question 3

In the college setting, to what extent can the Skills for Study 1 assessments be considered an accurate predictor of the academic achievement of international students as measured by their Final Academic Score?

The positive correlations observed in the present study between the SS1 scores and the Final minus SS1 scores are higher than the correlations reported by the predictive validity studies cited in the literature review. It was noted earlier that comparisons between the predictive validity of different EAP assessment instruments need to take into account the way the construct is defined and operationalized in each test. A higher correlation would imply a stronger similarity between the predictor and the criterion variables.

The SS1 assessments tend towards the 'strong' end of McNamara's (1996) scale of second language performance assessments as their construct is task/context focused. The main aim of the SS1 assessments is to answer the question: Are the students able to write an academic essay and give an academic presentation in English? The assessment criteria of SS1 are modelled on the criteria used by subject tutors when evaluating student performance in academic contexts. Content, structure, support and clarity of expression are each given equal weighting, with the first three criteria reflecting the factors cited by Errey (2000) as the ones that influence subject tutors the most. A comparison with the marking criteria for some of the BLSS modules reveals that students are assessed on research, content and argument, structure, coherence, and referencing. Moreover, for the oral presentation, non-linguistic factors such as body language, eye-contact, and voice projection are important for successful completion of task. Such criteria are combined with content, which subsumes research skills, structure and support from academic sources to conform to the indigenous assessment criteria of the academic community.

In addition, the SS1 assessments are in essence achievement tests based on course outcomes, which comprise a range of skills that students need in order to succeed in

their academic studies – analysing assignment tasks, performing library searches, identifying relevant sources, skimming, scanning, reading for detail, note-taking, synthesising information, planning, writing drafts that adhere to specific academic genre conventions, using sources, citing and referencing correctly. Therefore, it should not be surprising that the SS1 scores correlate well with the final academic scores as these skills are transferrable across academic disciplines. As Leki (2011, p.104) points out students 'are less likely to transfer knowledge to new situations perceived as unlike the old ones', but if students understand that their subject tutors value the same skills they are taught in their SS1 classes, such a transfer is more likely to occur.

Another factor that may contribute to the higher predictive validity of the SS1 assessments is that they are out-of-class and untimed. Hartill (2000) notes that the Law Department at a UK university introduced take-home papers because they realised that even non-native speakers of English who generally produced superior work to their native-speaker counterparts often underperformed under exam conditions. Slaght and Howell (2007, p. 255) also point out that grades from out-of-class assessments should be part of the overall evaluation of student performance as timed tests may not always be the best measure of a person's abilities. The SS1 assessments allow the students not only enough time to research a topic thoroughly, but also the opportunity to revise and improve the language and style of their essay. Undoubtedly, when working on out-of-class assignments, students have access to dictionaries, thesauruses and translation software, although there is no guarantee that these necessarily improve the quality of lexis and grammar in student essays. For example, one study of dictionary use by international students in a UK university found that students often experienced problems with identifying the appropriate sense of the word they were looking up, which sometimes resulted in a serious misconstruction of the text (Nesi, 2002). Moreover, the product of translation software, regardless of whether it is used in reading comprehension or as a writing tool, is by no means perfect, especially between languages with 'striking lexico-sematic and structural differences' (Niño, 2009). Therefore, it would be an oversimplification to argue that access to reference books and tools such as Google translate automatically improves the quality of student writing.

In fact, a more important determinant of the better predictive validity of strong performance assessments may be the process approach to writing utilised by the Skills for Study 1 course. Students receive support from their tutors in the form of feedback on an essay plan and a first draft before they submit their essay for grading. Although some authors have questioned the validity of assessments completed with some assistance from peers or tutors (Gearhart & Herman, 1998), Weigle (2002, p. 177) argues that in the real world, not least in academic settings, the 'social aspects of writing' are just as important as the individual ability to produce texts. She even points out that as an author she has benefited from the feedback and suggestions received from colleagues and editors. Thus, it could be argued then the SS1 assessments are more reflective of writing in the real world than a timed impromptu test.

Of course, the reliability of out-of-class assessments should not be neglected. Just as with large-scale language tests, it is necessary to implement proper scoring procedures (Weigle, 2002, p.182) so that all stakeholders, including the examinees and the host university, can be assured that the assessments are a fair and reliable measure of ability. In the College, this is achieved by means of termly standardisation meetings and the double marking of 10-20% of the essay scripts and presentations. Another threat to reliability is plagiarism. If students submit assignments that have been partially or wholly plagiarised, the score awarded will not be a reflection of their ability. Chen and Ku (2008, p.87) summarise the factors that contribute to plagiarism on the part of international students: cultural attitudes, misunderstanding of what constitutes plagiarism, lower English language proficiency, and inadequate institutional policies. In the College, this problem is addressed by (a) in-class activities aimed at raising awareness of plagiarism and ways to avoid it, and (b) the use of TurnItIn, defined as an 'electronic detection service' (Salmons, 2008, p. 210), to locate passages from electronic sources that have been reproduced word-for-word.

One of the possible reasons why the correlation, although large at r_s=.593, is not even higher is that at the end of term one, the students are still adapting to the requirements of academic study in the UK. After submitting their SS1 assignments, they have one or two terms to further develop the skills required for academic success. Another factor may be the extent to which the students master the content of their subject modules. Although SS1 can be classified as strong performance assessment, the topic of the

essay and the presentation is not directly related to the students' area of study. This is because (1) the assessments are designed for students on both the Business, Law and Social Sciences, and Science and Engineering programmes, and (2) the tutors on the course are not necessarily subject specialists in business or social sciences. Therefore, it is conceivable that students may acquire the necessary research and academic writing skills, but still perform unsatisfactorily in their subject modules due to inadequate content knowledge.

Overall, the large positive correlation between SS1 and the Final minus SS1 scores is to a certain extent due to the observation that as a strong performance assessment SS1 incorporates a range of the real-life skills required for academic success. These are reflected both in the assessment setting and in the marking criteria. Out-of-class, untimed assessments are the norm in higher education. Moreover, subject specialists employ marking criteria that focus primarily on content, text structure, logical argumentation, and use of academic sources, and in that sense SS1 represents an authentic assessment of students' readiness for academic study.

6. CONCLUSIONS AND IMPLICATIONS FOR FUTURE RESEARCH

The present study investigated the predictive validity of weak and strong second language performance assessments in a university pathway programme. Spearman rank order correlations were calculated between the scores of two weak performance assessments – IELTS and Language for Study 1 (LS1), which is an in-house end-of-term language examination – and the Final Academic Score, which is a composite of eight scores that students obtain at the end of their studies and which determines their progression to the host university. In addition, Spearman rank order correlations were calculated between the scores from one strong performance assessment – Skills for Study 1 (SS1), which consists of an academic essay and presentation – and the Final Academic Score minus SS1.

The findings appear to confirm the views expressed in the literature that assessments based on the view of language proficiency as 'pragmatic ascription' (Bachman, 1990, p.254), i.e. someone is 'able to do X' in the language attain a greater degree of success

when predicting real-life performance (Skehan, 1998). Significant large positive correlations were found between the SS1 Overall score and the Final Minus SS1 score (r_s=.593, n=568, p< .01), while although significant the correlations between the Final Academic score and the LS1 Reported score (r_s=.479, n=209, p< .01), and the Final Academic score and the IELTS overall score (r_s=.467, n=452, p< .01) are smaller and can be described as medium (Cohen, 1998, cited in Pallant, 2007, p.132).

It was argued that strong second language performance assessments are better predictors of success in university pathway programmes as they are likely to simulate the conditions of the target language use situation more accurately than weak performance assessments, which typically bear only a superficial resemblance to real-world academic tasks. The main features of strong second language performance assessments that increase their predictive power are: the task/context approach to construct definition, the use of indigenous marking criteria, and their untimed and process-oriented nature. On the other hand, the characteristics of weak second language performance assessments – the trait/ability focused approach to construct definition, the use of primarily linguistic marking criteria, and their timed, product-oriented nature – stem from the view of language proficiency as a 'theoretical construct' (Bachman, 1990, p.254), i.e. someone 'has ability X'. Such assessments are more appropriate for determining the level of ability of individuals.

One interesting area for future research would be to establish the link between 'someone has ability Y' and 'someone is able to do X'. Bachman (1990, p. 253) argues that an assessment that predicts future behaviour cannot be seen as an indicator of ability. On the other hand, weak performance assessments that indicate ability are often used as predictors of future behaviour. The present study demonstrates that this may not be a valid use of such tests. Perhaps, in the interest of fairness, it is advisable to place test-takers on both a language ability scale and a successful task completion scale. It is conceivable, in my view, that people may exhibit different levels of, say, reading comprehension in a timed test designed to measure individual language ability, and in an untimed assessment which allows access to external resources. The question for educational institutions will be: What is valued more – spontaneous impromptu demonstration of receptive and productive skills, or unhurried and

rehearsed performance? If the answer is the former, only students approaching native-like ability should be admitted to university to succeed or fail just like their native-speaker counterparts. If the answer is the latter, then students who, through strong performance assessments, show ability to adapt and develop should not be denied access to higher education on the basis of weak second language performance assessments.

Another point that merits consideration is the extent to which the findings of the present study have any meaningful practical applications. For example, do they enhance our ability to predict academic achievement? Unfortunately, according to Cohen, Manion, and Morrison (2000, p.202) only correlations over 0.85 allow for accurate individual or group predictions. The authors note, however, that such high correlations are rare in education studies. In the present study the correlations are within the 0.35 to 0.65 range and significant at the 0.01 level. Cohen, Manion, and Morrison (2000, p.202) state that correlations in this range are 'of little use for individual prediction' and can be used to make only 'crude group predictions'. However, they point out that their value may increase if used as part of multiple regression analysis. As mentioned in the literature review, a number of factors have been identified as contributing to academic success: motivation, learning strategies, quantitative skills (Cho & Bridgeman, 2012), intellect and acculturation (Cope, 2011), personal background, academic background, teaching and support (Feast, 2002). Therefore, future research making use of multiple regression analysis may focus on what combination of factors most successfully predicts academic achievement in the College.

The study also has some limitations. For example, although the majority of the assessments that comprise the Final Academic Score involve extensive writing, one subject – Statistics for the Social Sciences – assesses completely different skills. This study did not focus on the individual grades that count towards the Final Academic Score, but there is some anecdotal evidence that Chinese students, in particular, perform better in the Statistics module, which may boost their final score. As some predictive validity studies of IELTS have shown, the subject courses that students attend may have an effect on the predictive power of second language performance assessments (Cotton & Conrow, 1995; Dooey & Oliver, 2002). Therefore, it may be useful to explore the extent to which the inclusion of Statistics for the Social Sciences

in the Final Academic Score affects the strength of the correlation. Related to this issue is the focus on Foundation Certificate of Business, Law and Social Sciences programme in the present study. However, a certain proportion of the students in the College study on the Foundation Certificate of Science and Engineering. In addition, a number of students are enrolled in the Graduate Diploma programmes both in Business, Law and Social Sciences, and Science and Engineering. All these students take the same Language for Study and Skills for Study modules and therefore complete the same assessments. It is important to ascertain the extent to which the LS1 and SS1 assessments predict the academic achievement as measured by the FAS for these groups of students as well. It is not inconceivable that the assessments may be valid predictors for one group, but not for another. Such a study may lead to changes in the curriculum in order to ensure that the skills students acquire and are assessed on are relevant.

Finally, a positive relationship between two variables X and Y 'does not imply in any way that X *influences*, *affects*, or *causes* Y' (Chen & Popovich, 2002). What the correlations reported in the present study demonstrate is that there is a tendency for students who perform better in second language performance assessments to achieve higher marks at the end of their Foundation Certificate programme, and that this tendency is more pronounced for strong performance assessments.

7. REFERENCES

Alderson, J. (2000). Testing in EAP: Progress? Achievement? Proficiency? In G. M. Blue, J. Milton, & J. Saville (Eds.), *Assessing English for Academic Purposes* (pp. 21-47). Bern: Peter Lang.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Bachman, L. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. (2004). *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press.

Bachman, L. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language Testing Reconsidered* (pp. 41-71). Ottawa: University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Bachman, L., & Kunnan, A. (2005). *Statistical Analyses for Language Assessment: Workbook and CD-ROM.* Cambridge: Cambridge University Press.

Banerjee, J., & Wall, D. (2006). Assessing and reporting performances on pre-sessional EAP courses: Developing a final assessment checklist and investigating its validity. *Journal of English for Academic Purposes, 5*, 50-69.

Black, J. (1991). Performance in English Skills Courses and Overall Academic Achievement. *TESL Canada Journal, 9(1)*, 42-56.

Brindley, G. (2009). Task-centred language assessment in language learning. In K. Van den Branden, M. Bygate, & J. Norris (Eds.), *Task-Based Language Teaching: A Reader* (pp. 435-454). Amsterdam: John Benjamins Publishing Company.

Brown, J. D. (2004). Performance Assessment: Existing Literature and Directions for Research. *Second Language Studies, 22(2)*, 91-139.

Brown, J. D., & Rodgers, T. S. (2002). *Doing Second Language Research.* Oxford: Oxford University Press.

Brown, J., Hudson, T., Norris, J., & Bonk, W. (2002). *An Investigation of Second Language Task-Based Perfomance Assessments.* Honolulu: University of Hawaii Press.

CAEL. (2013). *What does the CAEL Assessment measure?* Retrieved from http://www.cael.ca/edu/format.shtml

Caudery, T. (1990). The validity of timed essay tests in the assessment of writing skills. *ELT Journal, 44(2),* 122-131.

Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19,* 254-272.

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and Nonparametric Measures.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-139. Thousand Oaks, CA: Sage Publications.

Chen, T., & Ku, N. K. (2008). EFL Students: Factors Contributing to Online Plagiarism. In T. Roberts (Ed.), *Student Plagiarism in an Online World: Problems and Solutions* (pp. 77-91). Hershey, PA: Information Science Reference.

Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing Writing, 8,* 165-191.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing, 29(3),* 421-442.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research Methods in Education, 5th Edition.* London: RoutledgeFalmer.

Cope, N. (2011). Evaluating Locally-Developed Language Testing: A Predictive Study of 'Direct Entry' Language Programs at an Australian University. *Australian Review of Applied Linguistics, 34(1),* 40-59.

Cotton, F., & Conrow, F. (1995). An Investigation into the predictive validity of IELTS amongst a group of international students studying at the university of Tasmania. *IELTS Research Reports, 1, Report 4,* 72-115.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of Language testing.* Cambridge: Cambridge University Press.

Dooey, P., & Oliver, R. (2002). An investigation into the predictive validity of the IELTS Test as an indicator of future academic success. *Prospect, 17(1),* 36-54.

Douglas, D. (2000). *Assessing Languages for Specific Purposes.* Cambridge : Cambridge University Press.

Douglas, D. (2010). *Understanding Language Testing.* Oxon: Hodder Education.

Dudley-Evans, T., & St John, M. (1998). *Developments in English for Specific Purposes.* Cambridge: Cambridge University Press.

EALTA. (2006). *EALTA Guidelines for Good Practice in Language Testing and Assessment.* Retrieved from http://www.ealta.eu.org/guidelines

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. *IELTS Research Reports, 4, report 6*.

Errey, L. (2000). Stacking the Decks: What does it Take to Satisfy Academic Readers' Requirements? In G. M. Blue, J. Milton, & J. Saville (Eds.), *Assessing English for Academic Purposes* (pp. 147-167). Bern: Peter Lang.

Feast, V. (2002). The Impact of IELTS Scores on Performance at University. *International Education Journal, 3(4)*, 70-85.

Gearhart, M., & Herman, J. (1998). Portfolio Assessment: Whose Work Is It? Issues in the Use of Classroom Assignments for Accountability. *Educational Assessment, 5(1),* 41-55.

Gebril, A., & Plakans, L. (2013). Toward a Transparent Construct of Reading-to-Write Tasks: The Interface Between Discourse Features and Proficiency. *Language Assessment Quarterly, 10*, 9-27.

Green, A. (2007). *IELTS Washback in Context: Preparation for academic writing in higher education.* Cambridge: Cambridge University Press.

Green, R. (2013). *Statistical Analyses for Language Testers.* Basingstoke: Palgrave Macmillan.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 - Writing: Composition, Community, and Assessment.* Princeton, NJ: ETS.

Hartill, J. (2000). Assessing Postgraduates in the Real World. In G. M. Blue, J. Milton, & J. Saville (Eds.), *Assessing English for Academic Purposes* (pp. 117-130). Bern: Peter Lang.

Hartnett, R., & Willingham, W. (1980). The Criterion Problem: What Measure of Success in Graduate Education? *Applied Psychological Measurement, 4*, 281-291.

Hayatt, D., & Brooks, G. (2009). Investigating stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. *IELTS Research Reports, 10, Report 1*.

Henning, G. (1987). *A guide to language testing: development, evaluation, research.* Cambridge, MA: Newbury House Publishers.

Howell, B., Nathan, P., Schmitt, D., Sinclair, C., Spencer, J., & Wrigglesworth, J. (2012). *BALEAP Guidelines on English Language Tests for University Entry.* Retrieved from http://www.baleap.org/projects/testing-working-party

Hughes, A. (2003). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

IELTS. (2011). *Guide for educational institutions, governments, professional bodies and commercial organisations.* Retrieved from www.ielts.org

IELTS. (2012). *Ensuring quality and fairness in international language testing.* Retrieved from http://ielts.org/pdf/IELTSQualityFairnessBrochure_2012.pdf

IELTS. (2013). *Band descriptors, reporting and interpretation*. Retrieved from http://www.ielts.org/researchers/score_processing_and_reporting.aspx

Ingram, D., & Bayliss, A. (2007). IELTS as a predictor of academic language performance. *IELTS Research Reports, 7*, Report 3.

International Study and Language Institute. (2013). *TEEP: Candidate's Handbook.* Retrieved from http://www.reading.ac.uk/web/FILES/ISLC/TEEP_candidate%27s_handbook.pdf

Jordan, R. (1997). *English for Academic Purposes: A Guide and Resource Book for Teachers.* Cambridge: Cambridge University Press.

Kerstjens, M., & Nery, C. (2000). Predictive Validity in the IELTS Test: A Study of the Relationship Between IELTS Scores and Students' Subsequent Academic Performance. *IELTS Research Reports, 3, Report 4*.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*, 81-96.

Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom* (pp. 140-154). Cambridge: Cambridge University Press.

Larson-Hall, J. (2010). *A Guide to Doing Statistics in Second Language Research Using SPSS.* New York: Routledge.

Lee, Y. (2005). A Summary of Contsruct Validation of an English for Academic Purposes Placement Test. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, Volume 3*, 113-131.

Leki, I. (2011). Learning to write in a second language: Multilingual graduates and undergraduates expanding genre repertories. In R. M. Manchon (Ed.), *Learning-to-Write and Writing-to-learn in an Additional Language* (pp. 85-109). Amsterdam: John Benjamins Publishing Company.

Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham, & D. Corson (Eds.), *Encyclopaedia of language and education, vol. 7: Language testing and assessment* (pp. 121-130). Dortrecht, The Netherlands: Kluwer.

Long, M., & Norris, J. (2009). Task-based teaching and assessment. In K. Van den Branden, M. Bygate, & J. Norris (Eds.), *Task-Based Language Teaching: A Reader* (pp. 135-142). Amsterdam: John Benjamins Publishing Company.

McNamara, T. (1996). *Measuring Second Language Performance.* Harlow: Longman.

McNamara, T. (1997). Performance Testing . In C. Clapham, & D. Corson (Eds.), *Encyclopedia of Language and Education, Volume 7: Language Testing and Assessment* (pp. 131-139). Dordrecht: Kluwer Academic Publishers.

Messick, S. (1980). Test Validity and the Ethics of Assessment. *American Psychologist, 35(11)*, 1012-1027.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14(4)*, 5-8.

Moore, T., & Morton, J. (2007). Authenticity in the IELTS Academic Module Writing test: a comparative study of Task 2 items and university assignments. In L. Taylor, & P. Falvey (Eds.), *IELTS Collected Papers: Research in speaking and writing assessment* (pp. 197-248). Cambridge: Cambridge University Press.

Nesi, H. (2002). A study of dictionary use by international students at a British university. *International Journal of Lexicography, 15(4)*, 277-305.

Niño, A. (2009). Machine translation in foreign language learning: language learners' and tutors' perceptions of its advantages and disadvantages. *ReCALL, 21(2)*, 241-258.

Norris, J. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing, 19(4)*, 337-346.

Oliver, R., Vanderford, S., & Grote, E. (2012). Evidence of English language proficiency and academic achievement of non-English-speaking background students. *Higher Education Research and Development, 31(4)*, 541-555.

O'Loughlin, K. (2011). The Interpretation and Use of Proficiency Test Scores in University Selection: How Valid and Ethical Are They? *Language Assessment Quarterly, 8(2)*, 146-160.

O'Loughlin, K., & Arkoudis, S. (2009). Investigating IELTS exit score gains in higher education. *IELTS Research Report, Volume 10*.

Pallant, J. (2007). *SPSS Survival Manual.* Maidenhead: Open University Press.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*, 111-129.

Rees, J. (1999). Counting the Cost of International Assessment: why universities may need to get a second opinion. *Assessment and Evaluation in Higher Education, 24(4)*, 427-438.

Reinders, H., Moore, N., & Lewis, M. (2008). *The International Student Handbook.* Basingstoke: Palgrave Macmillan.

Robinson, P., & Ross, S. (1996). The Development of Task-Based Assessment in English for Academic Purposes Programs. *Applied Linguistics, 17(4)*, 455-476.

Routes into university and higher education. (2013). Retrieved from http://www.nidirect.gov.uk/routes-into-university-and-higher-education

Salmons, J. (2008). Expect Originality! Using Taxonomies to Structure Assignments that Support Original Work. In T. Roberts (Ed.), *Student Plagiarism in an Online World* (pp. 208-226). Hershey, PA: Information Science Reference.

Salvia, J., & Ysseldyke, J. (2001). *Assessment.* Boston, MA: Houghton Mifflin .

Schmitt, D. (2012). *EAP Assessment in the UK.* Retrieved from http://www.baleap.org/projects/testing-working-party

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18(4)*, 373-391.

Skehan, P. (1998). *A Cognitive Approach to Language Learning.* Oxford: Oxford University Press.

Slaght, J., & Howell, B. (2007). TEEP: A Course-driven Assessment Measure. In O. Alexander (Ed.), *New Approaches to Materials Development for Language Learning: Proceedings of the 2005 joint BALEAP/SATEFL conference* (pp. 253-264). Bern: Peter Lang AG.

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes*, 89-101.

The TOEFL iBT Test. (2013). Retrieved from TOEFL: http://www.ets.org/toefl

UK Border Agency. (2013). *English language ability.* Retrieved from http://www.ukba.homeoffice.gov.uk/visas-immigration/studying/adult-students/can-you-apply/english-language/

UK Council for International Student Affairs. (2013). *International student statistics: UK higher education.* Retrieved from http://www.ukcisa.org.uk

Weigle, S. C. (2002). *Assessing Writing.* Cambridge: Cambridge University Press.

Weigle, S. C. (2004). Integrated reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*, 27-55.

Weir, C. (2005). *Language Testing and Validation.* Basingstoke: Palgrave Macmillan.

Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy, & N. Hornberger (Eds.), *Encylopedia of Language and Education, 2nd Edition, Volume 7: Language Testing and Assessment* (pp. 111-122). Boston, MA: Springer Science+Business Media LLC.

Woodrow, L. (2006). Academic Success of International Postgraduate Education Students and the Role of English Proficiency. *University of Sydney Papers in TESOL, 1*, 51-70.

Yen, D., & Kuzma, J. (2009). Higher IELTS Score, Higher Academic Performance? The Validity of IELTS in Predicting the Academic Performance of Chinese Students. *Worcester Journal of Learning and Teaching (3)*, 1-7.

8. APPENDICES

Appendix I: Sample LS1 Reading & Writing Exam

Language for Study Level 1
End of Term Writing Examination Question Book

Time: 60 minutes

Instructions to students:

1. Read all instructions very carefully.

2. At the start of the examination, you should have:
   i. this question book
   ii. an answer book.

3. Read all of the exam questions carefully before you begin to write.

4. Write your answer(s) in the answer book.

5. Write in black or blue ink.

6. Cross out any mistakes.

7. You cannot use a dictionary.

8. If you have questions at any time during the examination, you must raise your hand and wait for an invigilator. Do not attempt to communicate, by any means, with any other candidate at any time before or during the examination.

9. You may not leave the examination hall during the first hour. If the examination is less than one hour, you must stay for the entire duration of the exam (toilet breaks are permitted if necessary).

10. At the end of the exam, do not speak or leave your seat until you have handed both the question book and the answer book and any additional sheets of paper you may have used to an invigilator, and you have been told by an invigilator that you may leave the room.

TASK:

You are a student at a UK university. In one of your core modules you have recently been discussing issues related to **Education and technology**. Read the notes you have made below (page 2) and the related quotations (page 3). Then **using the information appropriately**, write a discussion essay for your tutor in response to the following statement:

"In the 21st century, distance learning is a more attractive option than attending a university in person."

In your essay you must **discuss** the statement above, presenting **arguments for** and **against** it. You should support your ideas by making reference to the reading extracts provided on page 3. **You should not copy sentences directly from the extracts but paraphrase the information using your own words.** You should provide in-text citations every time you use the provided sources. **You should discuss both positive and negative aspects of the essay topic and use ideas from at least 2 of the extracts provided. You may also use your own ideas.**

You have 60 minutes to complete this task. You should write a maximum of 250 words.

NOTES:

Issues about Distance Learning

Positive Aspects

- *More choice and flexibility*

Negative Aspects

- *Lack of self-discipline may lead to failure*

Students have the opportunity to choose from various schools, programs and courses which are not available in the area where they live. This is especially beneficial for those who live in rural areas that only have one or two educational facilities, which may offer limited course and program options for students. Furthermore, the flexibility offered by online learning benefits not only undergraduate students, but also individuals who already have full-time jobs or family commitments.

Doe, J. (2012). *Online Education.* Retrieved from http://seacstudentweb.org/top-benefits-of-online-education.php

The flexibility advantage of distance learning does not reduce the most significant demand on a student's time because distance learning still requires students to log on, study, do homework and write exams. The drop-out rate for online students is higher partly due to a common misunderstanding about the amount of time required of a distance student. Moreover, the traditional university environment in which inquiring students learn through interaction with other inquiring students is replaced by an uninspiring electronic environment experienced in the isolation of the home.

Simmons, J. R. (2001). Distance Learning: Education or Economics? *International Journal of Value-Based Management,* 14,157–169

Online technologies are attractive because they provide the opportunity to create rich learning environments consisting of multimedia resources and facilities for communication and interaction. What online technologies do, in addition to common access to learning resources, is promote student–teacher and student–student interaction whatever the participants' location.

Calvert, J. (2005). Distance Education at the Crossroads. *Distance Education,26*(2), 227–238

*****This is the end of the exam paper*****

Appendix II: Sample LS1 Listening & Speaking Exam

LANGUAGE FOR STUDY 1
END OF TERM SPEAKING EXAMINATION (SAMPLE SET)

EXAMINER'S VERSION

Task Instructions:

This exam task consists of two parts:

- In the first part of the task the candidate listens to two speakers and answers questions about what they have heard (2 questions for each speaker). (2- 3 minutes)
- In the second part of the task, the examiner will ask the candidate up to 3 more questions about the topic of healthy eating. The candidate needs to answer all questions they are asked with a brief response. This part of the exam should last 2 – 3 minutes.

Total time: approx **10 minutes** including turnaround time and completing mark sheets.

Part 1.

Give the candidate their tasks, some paper to make notes on and ask them to read the instructions. Explain what will happen in the exam and give them 30 seconds to go through the questions for the listening. Before you play the recordings, ask the candidate if they understand what they need to do, explain again if necessary. Point to the paper provided and encourage the candidate to make notes whilst they listen.

Questions for the extracts:

Speaker 1:

1. **Why do supermarkets sell fat-free foods?**
2. **Why are fat-free foods not always healthier than full-fat products?**

Speaker 2:

1. **Why do most people nowadays choose fat-free products?**
2. **What trend in people's weight has been observed?**

Ask the candidate if they are ready to listen. Then play the recordings of both speakers **once only. Do not give the candidate extra time between the speakers.**

Audio scripts:

Speaker One:

(male/female lecturer's voice): Have you noticed how supermarket shelves have gradually been stacked with more and more "fat free" or "low-fat" food? Almost every food product comes in two versions nowadays: every regular, that is, full-fat product also has a low-fat or fat-free equivalent. You might ask why? Well, the answer is simple, the food industry has responded to their customers' demand. These customers seem to believe that eating 'low-fat' or 'fat-free' products is healthier than eating regular, full-fat foods. The problem is, however, that this is not exactly true. Let me explain why. You see, to claim that a product is 'low fat', the amount of fat in this product must be at least 25% lower than in the standard product. This fat is replaced with other ingredients, which are not necessarily healthier than fat! For example, fat is replaced with sugar or, worse still, with some chemicals which improve the taste of the product. And so, in result, low-fat products have got twice as many ingredients as the regular versions and can be much worse for our diet!

Speaker Two:

(male/female lecturer's voice): Recent research shows that most people nowadays opt for low-fat or fat-free products. The cause of this trend is relatively easy to understand. People believe that if they buy fat-free foods then they should not gain weight or should be able to lose weight because they are eating less fat. Although it seems logical, it is, unfortunately, not true at all and here's why. The fact is that there is no connection between eating fat and the fat produced by your body. The body can take any type of food and turn it into body fat. For example, if you eat 2 lbs. of sugar every day then you'll probably gain some fat tissue after a while – yet sugar has zero fat in it! So, eating 2 lbs of sugar every day is a "fat free" diet, but yet you'd still gain fat. It's simple, eating fat-free products doesn't mean you will not put on weight!

This brings me to my second point, namely: what does this all tell you about fat-free products? You see, what happens is that most people hear the term 'fat-free' and see it as a green light to eat as much of it as they want, thinking they will not put on extra weight. And that's why despite the fact that we have all these low-fat and fat-free foods, we are fatter than at any time in history and the trend is only continuing to go higher!

Stop the recording. Ask the candidate the listening questions. Elicit an answer after each question.

PART 2.

1. **Choose <u>three questions</u> from the list below. Ask ONE question at a time. You can repeat the question if the candidate asks you to. This part of the exam should last 2 – 3 minutes.**

    1. *What kind of food is popular in your country?*
    2. *How important is it to eat healthily?*
    3. *What is your typical diet?*
    4. *How important is it to teach children about healthy food and who's responsibility is it (school, parents)?*
    5. *Where can people find information about and get advice on healthy eating in your country?*
    6. *How different is British food to food from your country?*

2. **After the candidate has answered the last question (or after the 3 minutes have passed), thank the candidate and inform them this is the end of the exam and they can leave the examination room. Collect all the notes and exam materials from the candidate before they leave the room.**

5. **Fill in the candidate's mark sheet using the LS Speaking Descriptors.**

Appendix III: Sample SS1 Essay Task

SS1 （Foundation Level- 2 Term）

Essay （60%）

（Autumn 2012）

This assessment is an **essay writing task**. The following gives you more information about the task.

Question:

Outline the factors which could contribute to low health expectancy in developed countries. Discuss possible solutions to reduce this problem.

In your **essay**, you should refer to a number of sources and can include your own ideas on this topic （you can use the sources provided in Unit 4 of the SS1 textbook, but must also include a minimum of **two** other relevant sources as well）.

You must write in paragraphs and make sure you provide in-text citations for all the sources you refer to in your essay and a list of references at the end of your essay.

The essay should be 750 words.

What you will be assessed on:

- The relevance of your ideas to the essay task.
- The structure of your essay （how well you planned your essay and how well you link your ideas within and between paragraphs）.
- How well you support your ideas using sources and how accurate your in-text referencing and final references are.
- How appropriate and accurate your use of English language is, especially register and style.

Submission checklist:

- Checked your essay several times for accuracy
- Typed and printed essay
- Used Times New Roman or Arial font type, size 12
- Double spaces between lines.
- Printed single-sided only.
  Included a title page with the following information:

  i. Module Code (e.g. FC501 2T)
  ii. Class/Group: (e.g. Group A)
  iii. Module Title (Skills for Study 1)
  iv. Assessment Title (e.g. Essay, Presentation etc.)
  v. Assignment Title: (e.g. *Discuss.....)*
  vi. Tutor Name: (name of tutor)
  vii. Student ID Number: (please add your ID number only and NOT your name)
  viii. Date of Submission: (date)
  ix. Word count (the number of words YOU used)

- Used a header on each page of your assignment with your ID number and module code (FC501 2T).
- Numbered all pages.
- Stapled all pages together
- Printed and submit two copies of your assignment.
- Filled in the official submission form, clearly stating: essay title, your name and ID number, your tutor's name.
- Kept the receipt for your submission.
- Sent a copy of your assignment to the Turnitin class set up by your tutor.

- Failure to follow these guidelines could result in you losing 5 marks.

Submission Deadline:

Appendix IV: Sample SS1 Presentation Task

SS1 (Foundation Certificate Level)

2 Term

Oral Presentation (40%)

(Autumn 2012)

You will work with one other student to prepare a presentation using the following title.

Outline the factors which could contribute to low health expectancy in developed countries. Discuss possible solutions to reduce this problem.

You need to:

1- Discuss what you are trying to persuade your audience of.
2- Work with your partner to create a PowerPoint presentation.
3- Deliver the presentation with your partner.

What you will be assessed on:

- The way you deliver your presentation (including how well you cooperate with your partner and how accurate and appropriate is your use of English language)
- The content of your presentation and the design of the PowerPoint slides to support what you say.
- How well you organize your ideas and link/signpost them.
- How well you present your thesis and support it using data or examples in appropriate detail.
- How well you deal with questions from the audience.

Further Guidelines:

- Presentations must be between **6 and 8 minutes**. Up to 3 additional minutes will be allowed for answering questions from the audience.
- Any sources you use to support points should be acknowledged.

Submission (Presentation) Date:

Appendix V: Box-and-whisker plots of frequency distribution output: IELTS Overall Score and Final Academic Score

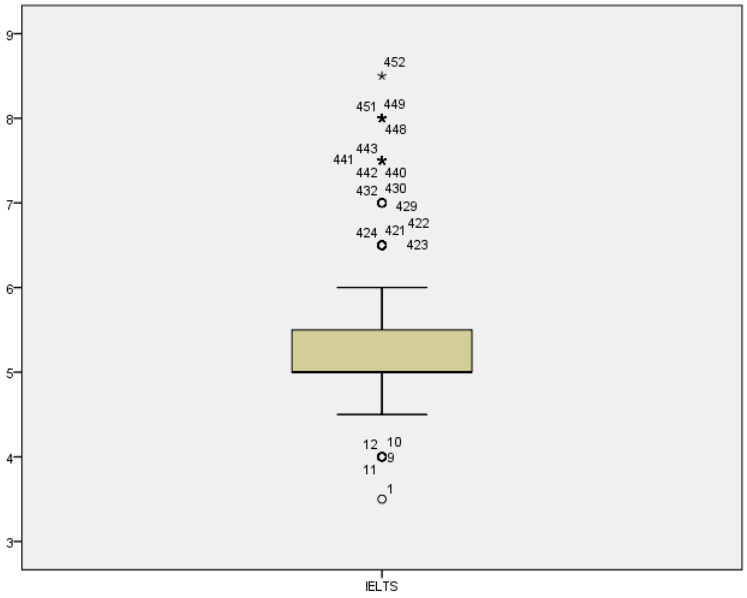

Figure 1. Box-and-whisker plot of frequency distribution output: IELTS Overall Score

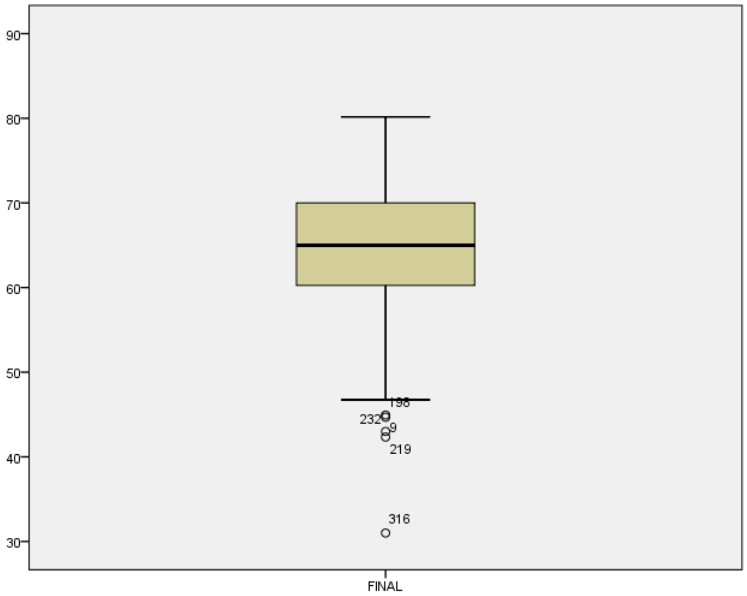

Figure 2. Box-and-whisker plot of frequency distribution output: Final Academic Score

Appendix VI: Histograms of IELTS and Final Academic Score distributions

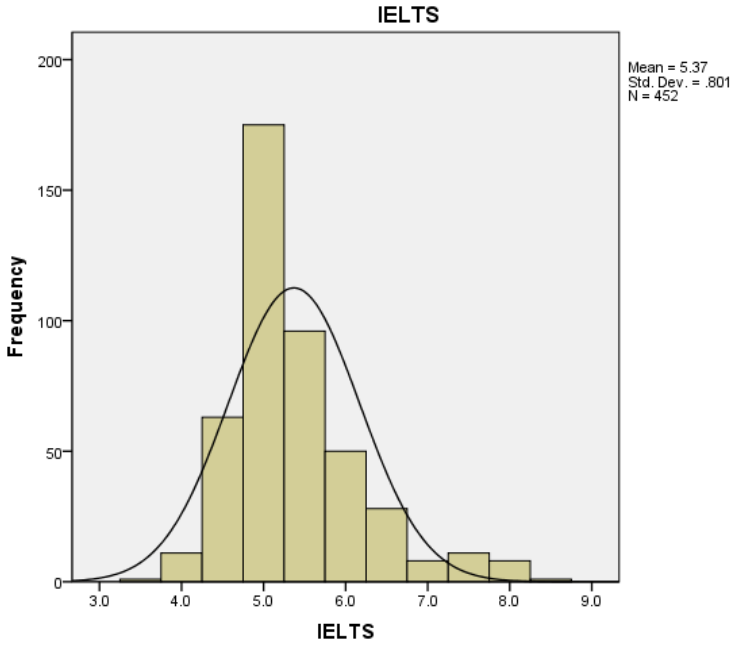

Figure 1. Histogram of IELTS score distribution

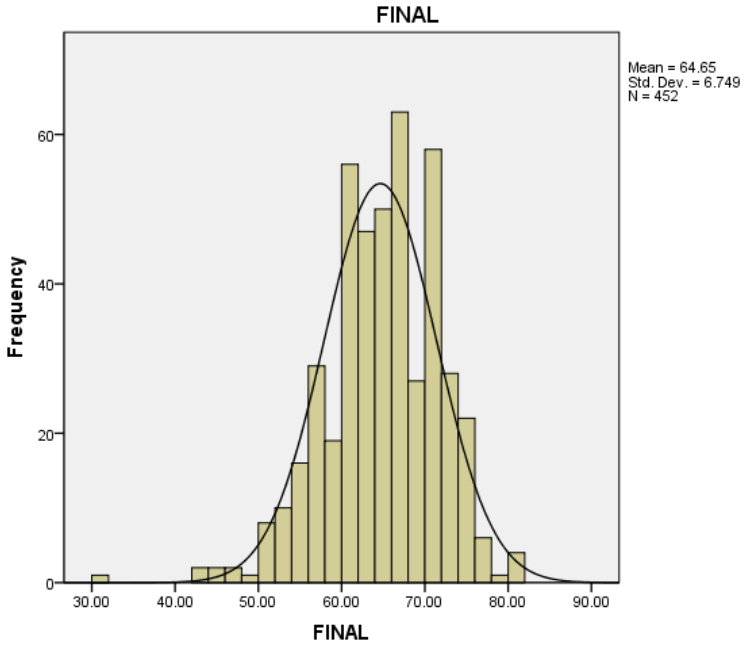

Figure 2. Histogram of Final Academic score distribution

Appendix VII: Normal and Detrended Q-Q plots of IELTS Overall and Final Academic Score

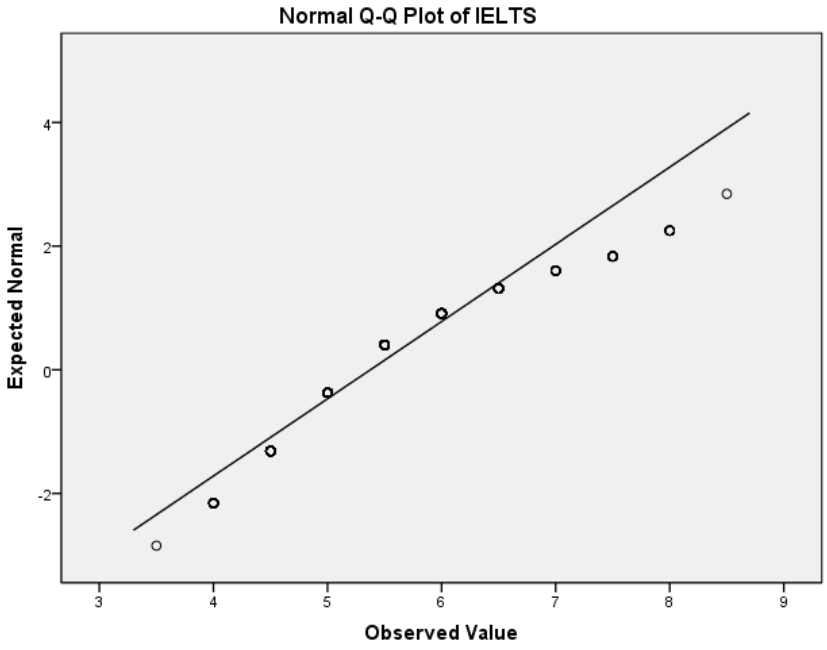

Figure 1. Normal Q-Q plot of IELTS Overall score

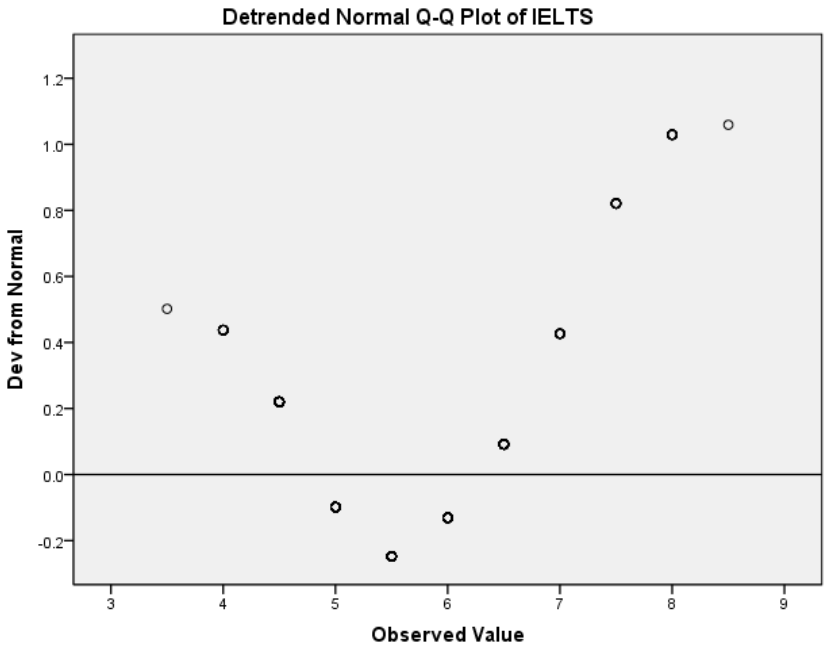

Figure 2. Detrended Q-Q plots of IELTS Overall score

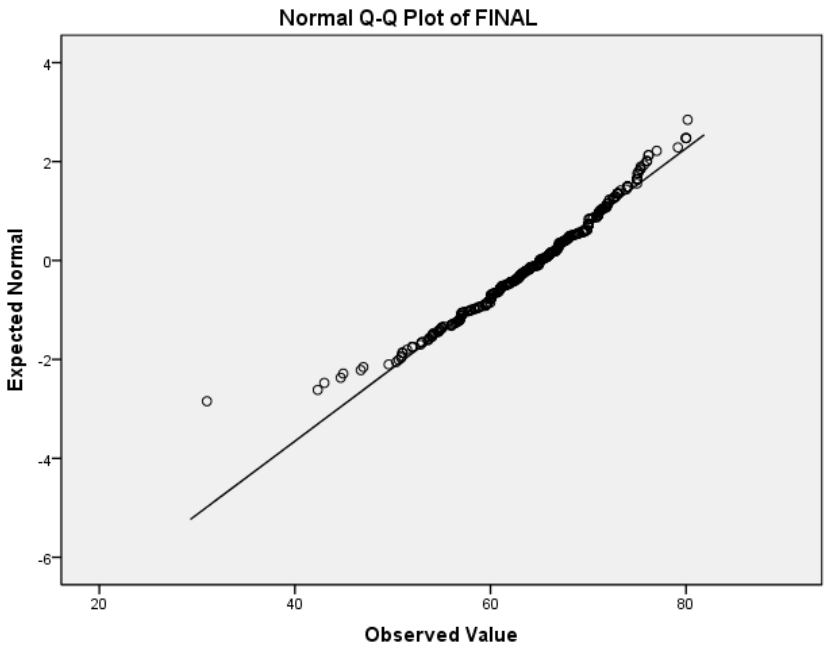

Figure 3. Normal Q-Q plot of Final Academic Score

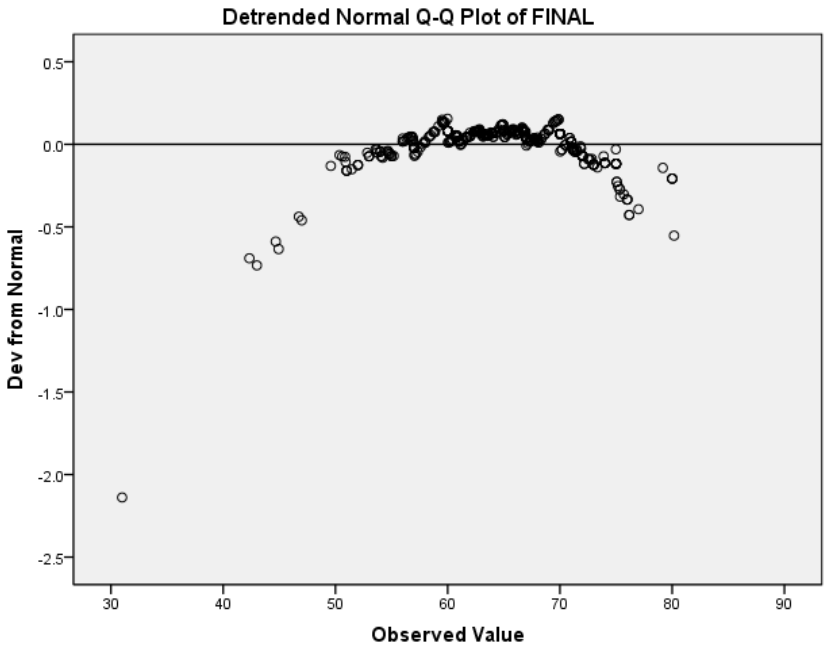

Figure 4. Detrended Q-Q plots of Final Academic Score

Appendix VIII: Box-and-whisker plots of frequency distribution output: LS1 Reported Overall Score and Final Academic Score

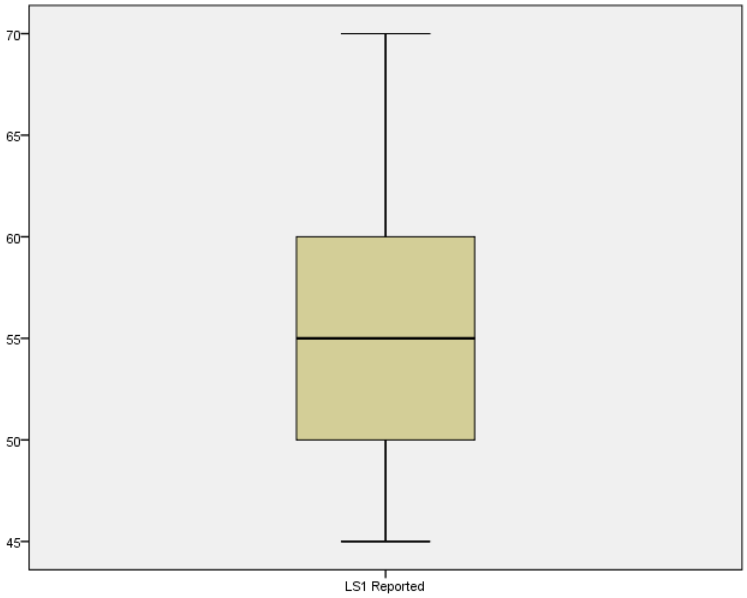

Figure 1. Box-and-whisker plot of frequency distribution output: LS1 Reported Overall Score

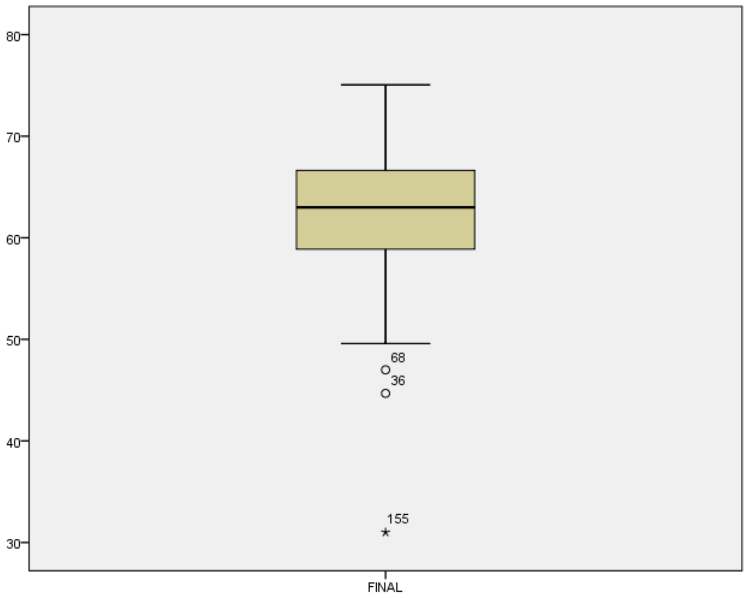

Figure 2. Box-and-whisker plot of frequency distribution output: Final Academic Score

Appendix IX: Normal and Detrended Q-Q plots of LS1 Reported, LS1 R&W, LS1 S&L, and Final Academic Score

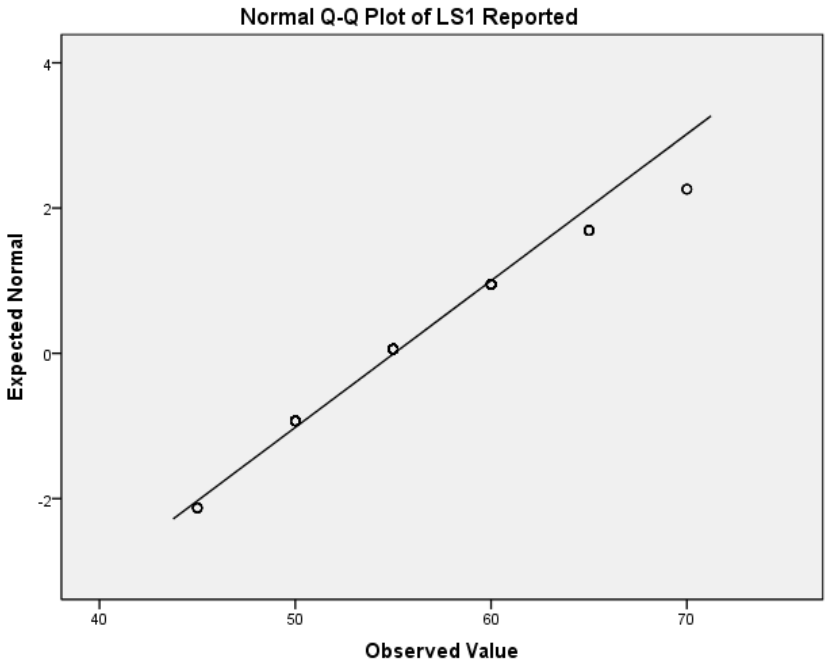

Figure 1. Normal Q-Q plot of LS1 Reported

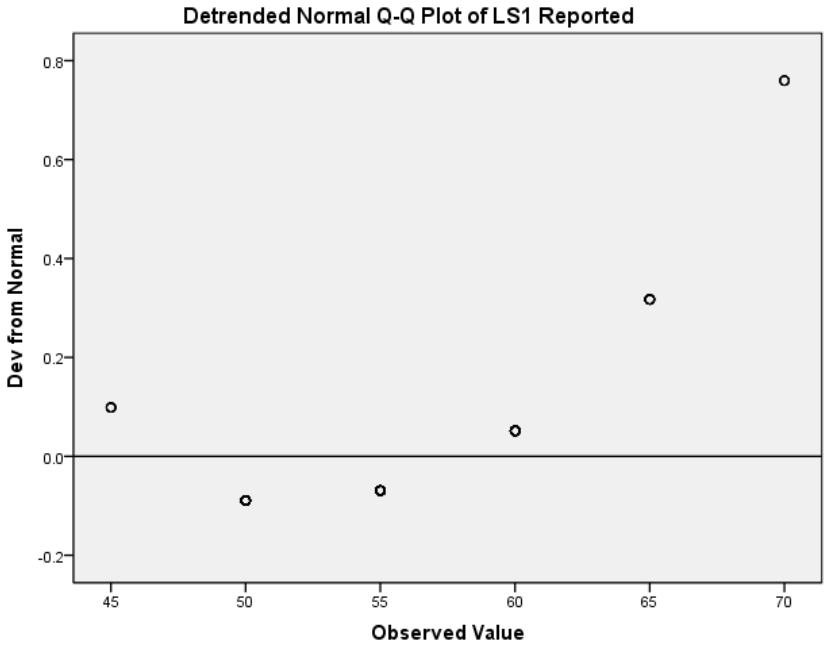

Figure 2. Detrended Q-Q plot of LS1 Reported

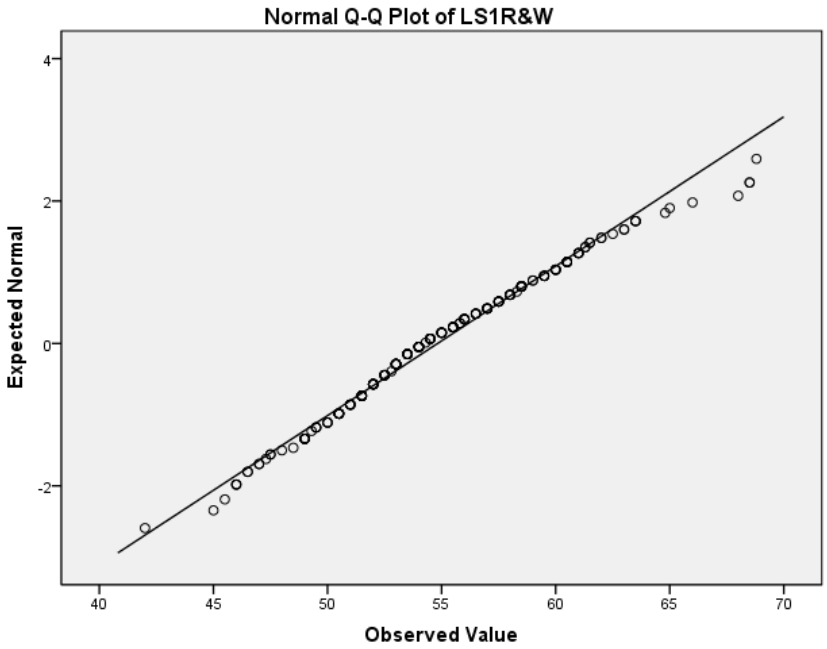

Figure 3. Normal Q-Q plot of LS1 R&W

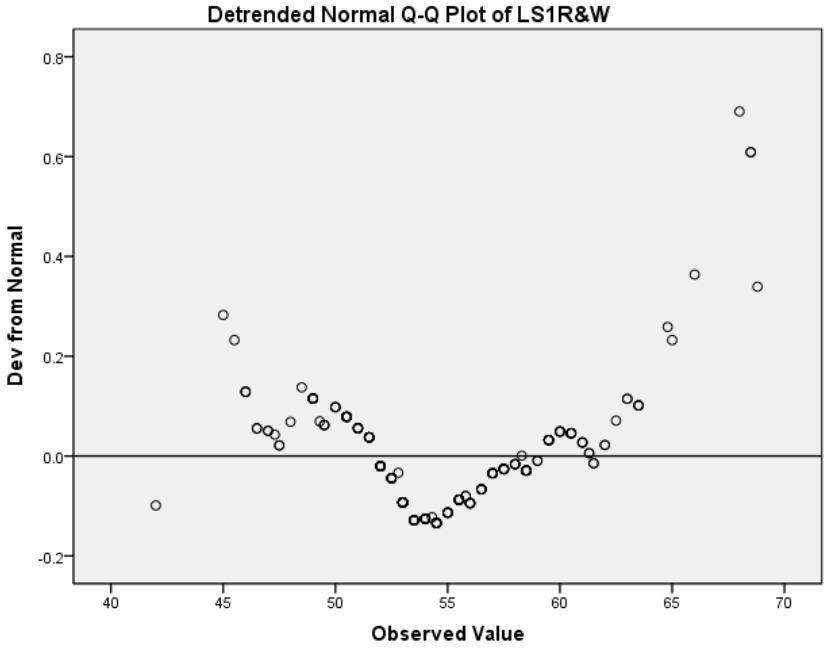

Figure 4. Detrended Q-Q plot of LS1 R&W

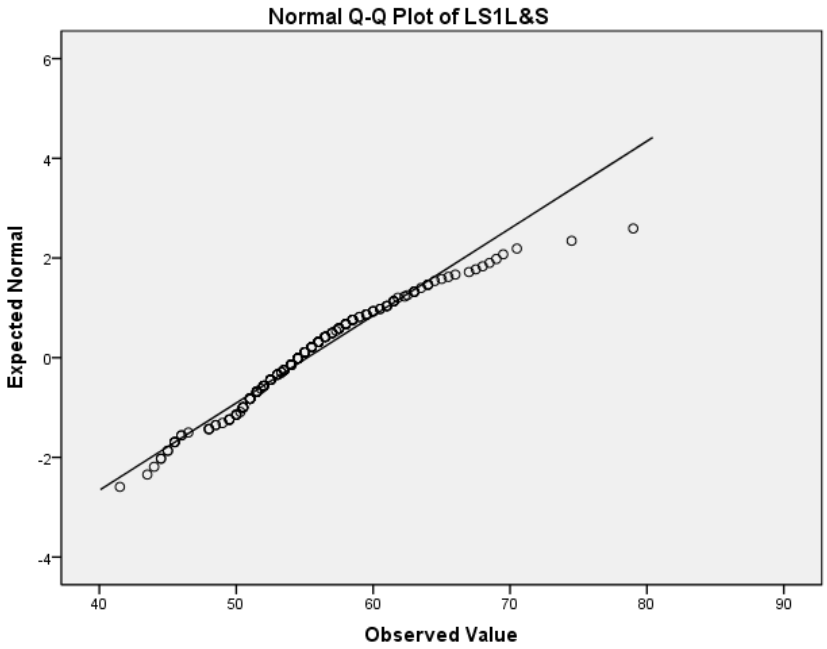

Figure 5. Normal Q-Q plot of LS1 L&S

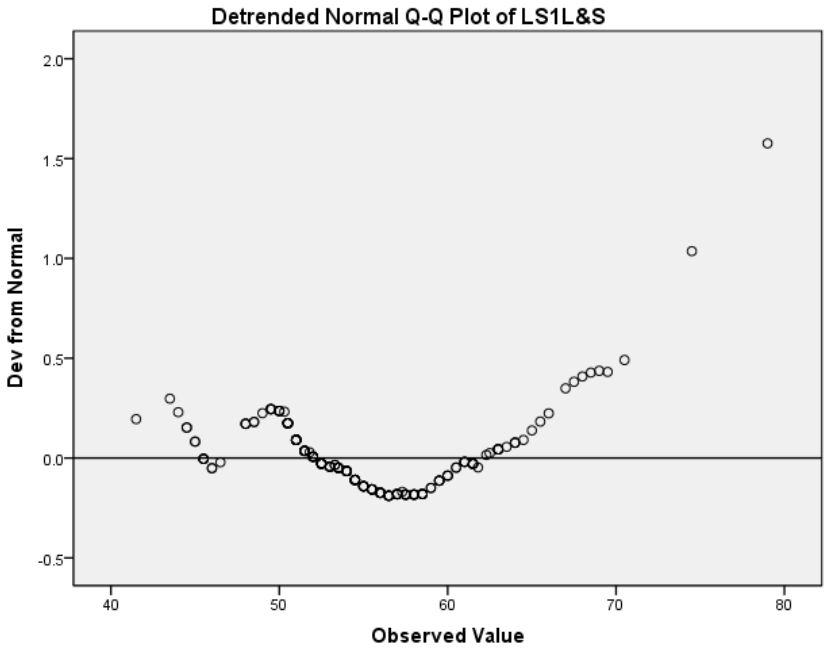

Figure 6. Detrended Q-Q plot of LS1 L&S

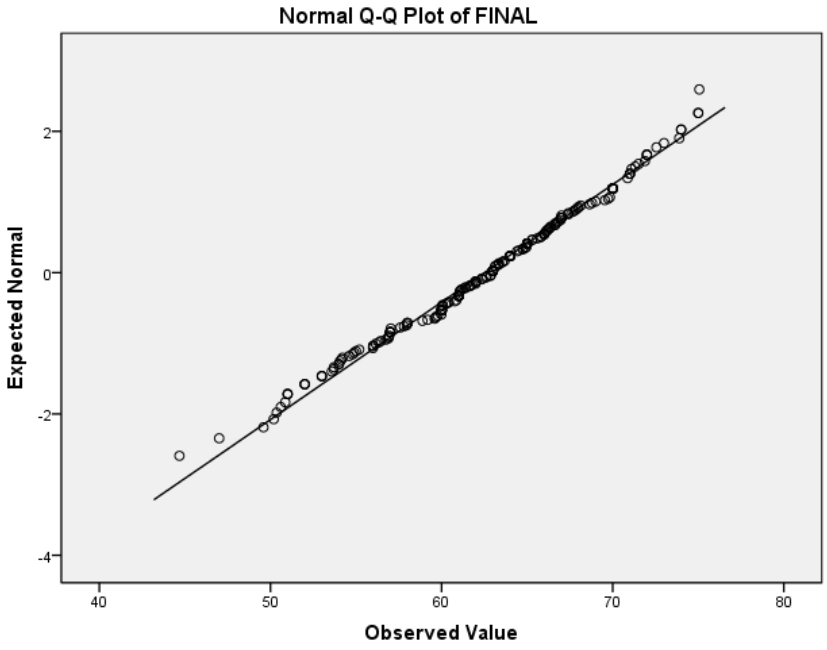

Figure 7. Normal Q-Q plot of Final Academic Score

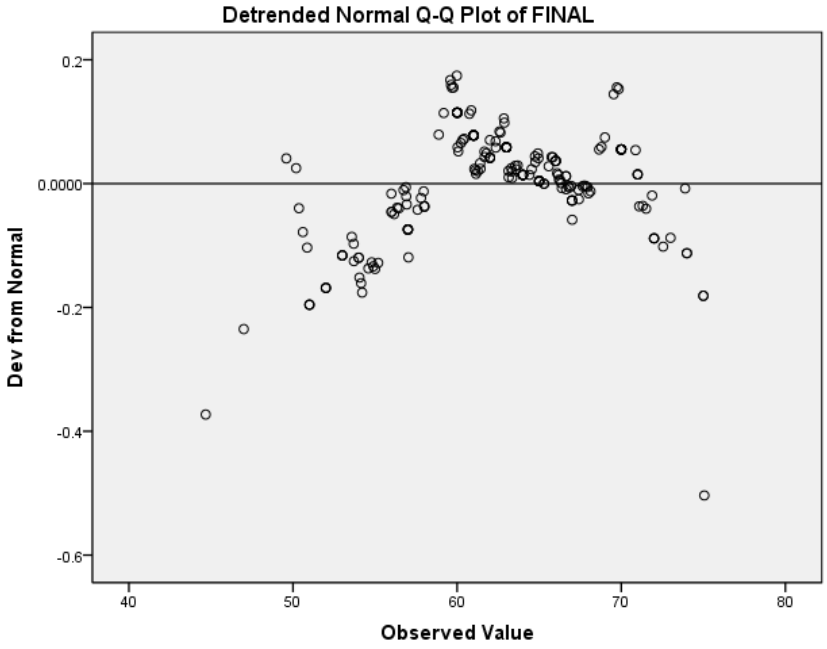

Figure 8. Detrended Q-Q plot of Final Academic Score

Appendix X: Box-and-whisker plots of frequency distribution output: SS1 Overall Score and Final minus SS1 (F-SS1)

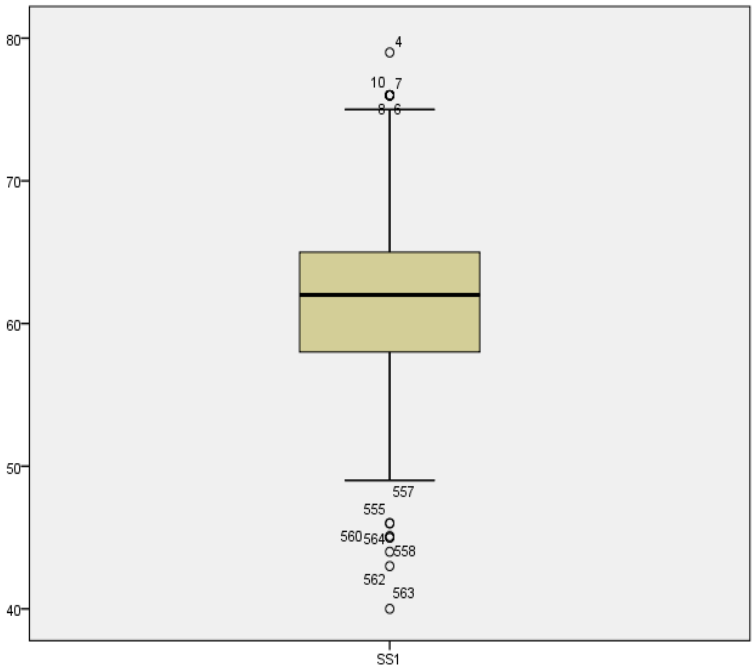

Figure 1. Box-and-whisker plot of frequency distribution output: SS1 Overall Score

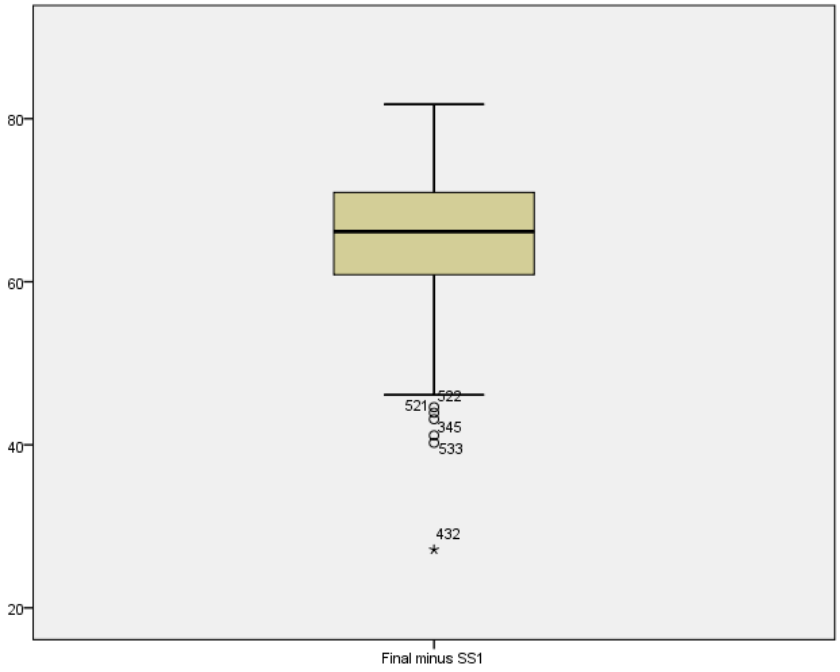

Figure 2. Box-and-whisker plot of frequency distribution output: Final minus SS1 (F-SS1)

Appendix XI: Normal and Detrended Q-Q plots of SS1 and Final minus SS1

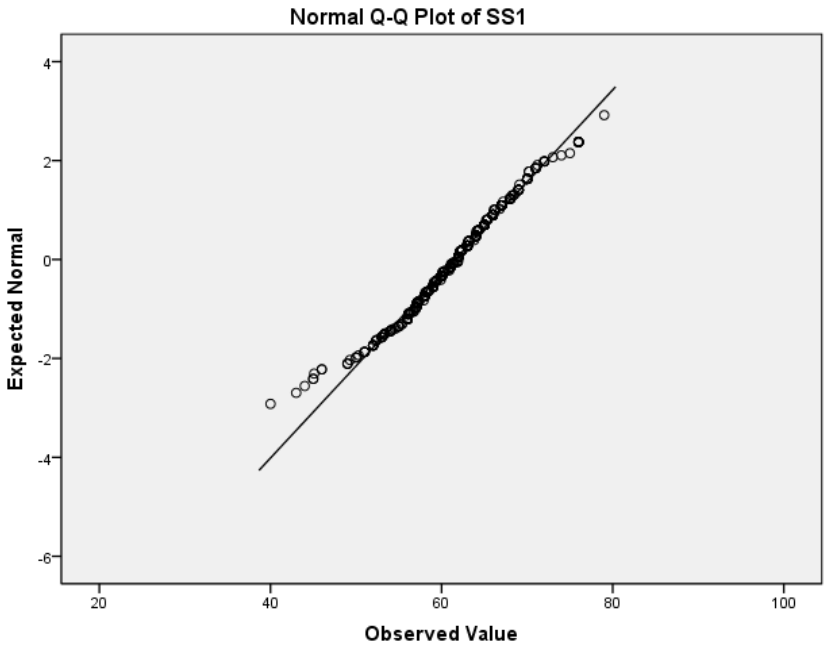

Figure 1. Normal Q-Q plot of SS1

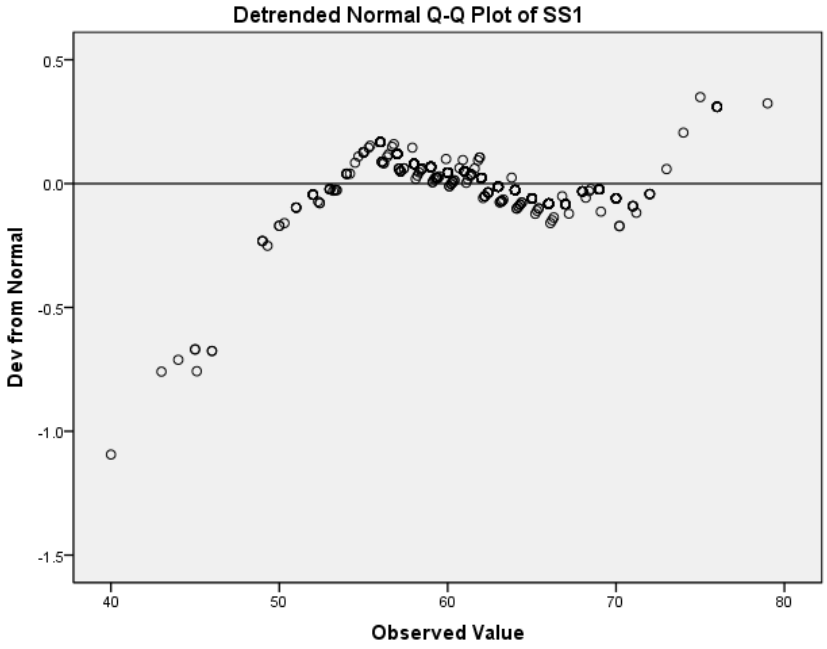

Figure 2. Detrended Q-Q plot of SS1

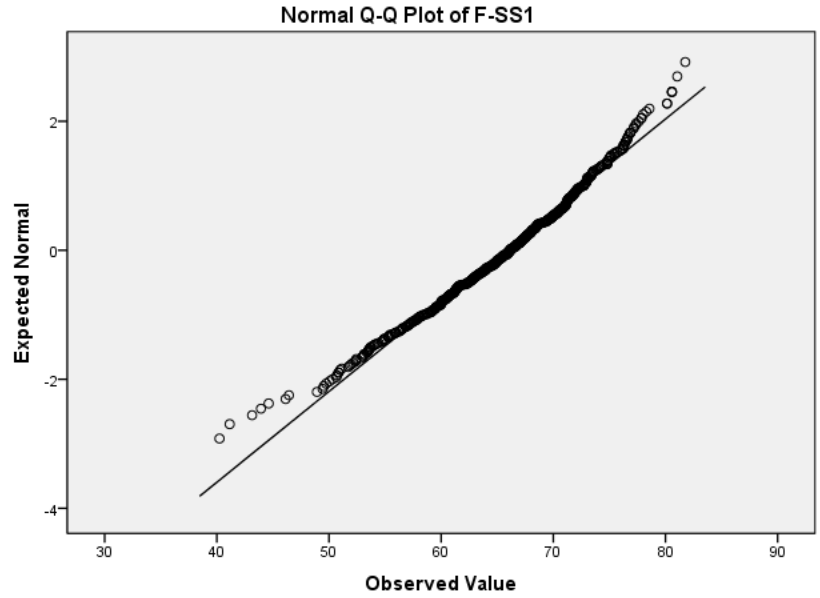

Figure 3. Normal Q-Q plot of Final minus SS1

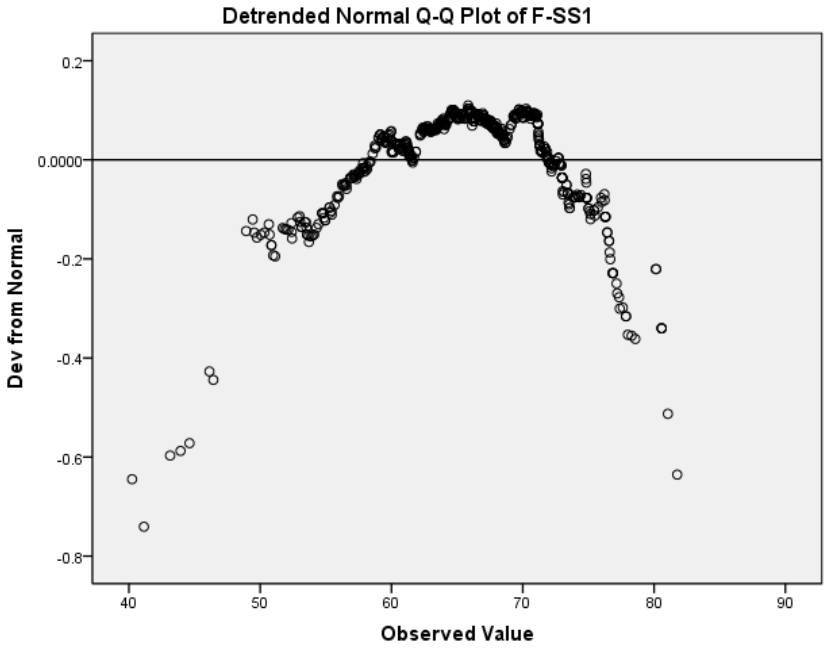

Figure 4. Detrended Q-Q plot of Final minus SS1