Centre for English Language Studies

Postgraduate programmes, Open Distance Learning

ESSAY COVER SHEET

Student ID number	906418
Module Number (1-6)	Module 6
Title of Degree Programme:	MA TESOL
Title of Module:	Testing
Assessment Task No.	TS/08/01
First submission or Resubmission	First
Date Submitted	March 2009
Name of tutor	Douglas Sewell

Words: 4314

TS/08/01 Discuss the opinions expressed by A and B on the IELTS test in Unit 1 of the ‘Testing’ course. You may alternatively do this question with reference to the TOEFL test; a similar conversation to that in Unit 1 could apply to TOEFL.

Table of Contents

1. Introduction

English, whilst being deemed to be the lingua franca by many, manages to employ thousands of people worldwide to teach, test and evaluate the language on a daily basis. With ever-evolving curricula with what are commonly perceived to be, by the writers at least, answers on how to learn English in an effective way, the actual testing of the language plays an important role in this continual cycle. However, Weir, in his studies (1990: 5), states that even though tests might integrate various language skills, only direct tests which contain and simulate authentic communication tasks can in turn claim to mirror actual authentic communicative interaction. Gilfert (1996) draws our attention to the fact that examinees often become experts in taking language tests but never actually learn *how* to use the language; so the necessity of a balanced, well-rounded test, covering all areas of the language, is essential.

Based on an imaginary conversation, detailed in Appendix 1, and drawing on my experience as an IELTS examiner, it will be the focus of this paper to look at whether or not the International English Language Testing System, or IELTS™ (hereafter 'the test' or 'IELTS'), is a actually a good yardstick for measuring students' abilities in English with conclusions being drawn about the effectiveness of IELTS being used on a global scale.

2. Background and Overview of IELTS

From its initial creation in 1980, when it was known as the English Language Testing Service (ELTS), it evolved into the International English Language Testing System in 1989 resulting from a validation survey conducted by Edinburgh University (Criper & Davies, 1988). It is now managed by the British Council, IDP (IELTS Australia) and the University of Cambridge ESOL Examinations, through more than 500 locations in roughly 120 countries. IELTS, as of 2008, has the highest recorded number of candidates internationally, with more than 1,700 universities in America alone and near to 6,000 organisations around the world recognising the test as a guideline for students' ability in English (IELTS Homepage, 2009). Whilst the Test of English for International Communication, TOEIC®, is aimed specifically at business English (Sharron, 1997; Chauncey Group International Ltd., 1999), the IELTS test claims to be a much broader test, which anyone can take (IELTS Homepage, 2009).

The test is comprised of four equally weighted sub-tests which are speaking, reading, writing and listening and is usually done over the course of one day in specified test centres with trained markers and examiners. The candidate's overall score is then worked out as the mean average of the four individual sub-tests. Examiners, for the speaking and writing modules, are monitored regularly and are re-standardised by an accredited IELTS trainer every two years; markers have to demonstrate that they are marking to set standards prior to marking listening and reading papers and are also re-tested every two years (IELTS Homepage, 2009).

3. The Need for Standardised Tests

Whilst some may argue that candidates' self-assessment is a more practical form of gauging personal levels of proficiency, and certainly a much more cost-effective option, Owen (1997: 4) highlights that such self-assessment is ultimately inadequate as it is subjective and non-systematic; hence the need for a common yardstick, in the form of a test, in order to make meaningful comparisons (Hughes, 1989: 4). Therefore, tests such as IELTS and TOIEC both claim that they offer this 'yardstick' (IELTS Homepage 2009; TOIEC Homepage, ETS, 2009).

3.1.1. Test Reliability

In order for a test to be deemed reliable, which Bachman and Palmer (1996: 19) define as "consistency of measurement", multiple administrations need to produce consistently similar results from an identical or near-identical test (Bachman, 1990; Bachman and Palmer, 1996; Weir, 1990). Hughes (1989:29) reinforces this by emphasising that, whilst it is impossible to produce tests which are 100% reliable, test writers generally strive to produce a test that results in similar scores between different administrations but with the same examinees. Another objective regarding reliability, that needs to be taken into consideration when devising tests, is to try to ensure that the test allows only for systematic errors, such as the actual skills of the test takers, rather than allowing unsystematic influences on test performances, such as distracting noises or simply lapses in candidates' concentration (Alderson, Clapham & Wall, 2005: 7), to affect measurement error. There will, however, be a certain amount of flexibility in these figures as it is nearly impossible to ignore the vast variation in different factors involved with a test taker's performances (Kluitmann, 2008).

Further theories related to test reliability are presented by Bachman (1990: 119; 164) including test method facets, personal attributes and random factors. Test method facets cover areas such as the testing environment, the test rubric, input, the expected response and finally the relationship between input and response. Personal attributes encompass age, gender, cognitive style and background; he then lists random factors such as tiredness, emotional condition and even more random differences in the testing environment. Taking all of these into consideration, it becomes evident that the actual measurement of test reliability is often seen to be a complicated task.

3.1.2. Reliability Measurement

Within the consistency of measurement (Bachman & Palmer, 1996:19) there are two main issues. First, would an examinee taking an identical test for a second time score equally well if all other variables involved remained constant? This estimate, expressed as a reliability coefficient falling between 0 and 1, is considered to be reliable if the number is higher within these parameters. Lado (*cited in* Hughes, 1989: 32) suggests that “good vocabulary, structure and reading tests are usually in the 0.9 to 0.99 range, while auditory comprehension tests are more often in the 0.8 to 0.89 range”.

This reliability coefficient helps us to compare the reliability of tests, but “it does not tell us directly how close an individual’s actual score is to what he or she might have scored on another occasion” (Hughes, 1989: 33). By factoring in the Classical True Score hypothesis (Bachman, 1990:167) that an examinee’s actual score is comprised of two components, namely the *true score,* which reflects the individual’s actual level of ability, and the *error score,* which is random due to external factors other than ability itself, it is possible to work out the Standard Error of Measurement, or SEM. The smaller this estimate, the closer test scores are to the *true score,* in turn making the test more reliable.

3.2. Test Validity

The validity of a test, or the degree to which a test actually measures what it is initially intended to measure (Hughes, 1989: 22; Brown, 2001: 387), is a complex criterion in the field of testing. With the ever growing list of recognised validities in the academic field of language testing, Owen *et al.,* (1997: 20) highlights that this list is often dispiriting due to the abundance of choice and approved validities. Amongst this selection, the most pertinent types of validity for the purpose of this paper are ‘construct’, ‘criterion-related’ (concurrent and predictive), ‘content’ and ‘face’ validities. The phenomenon of backwash will also be discussed.

3.2.1. Construct Validity

Construct validity is a term often central to theoretical testing literature. In simple terms, construct validity encompasses whether or not the test is actually testing the criteria it claims to test (Bachman, 1990; Hughes, 1989; Weir 1990). Hughes (1989: 26) explains this further by stating that “the word construct refers to any underlying ability which is hypothesised in a theory of language ability”, which Brown (2001: 389) highlights further by posing the question, “Does the test tap into the theoretical construct as it has been defined?”. Construct validity therefore reflects the area of target language ability being measured and very little else (Bachman & Palmer, 1996: 21). Hughes (1989: 27) goes on to say that construct validation is often viewed as a research activity where theories are tested and are either “confirmed, modified or abandoned”. Through conducting construct validation, empirical testing of hypothesized relationships between test scores and actual abilities takes place (Bachman, 1990: 256) In order for a test construct to be valid, it needs to be compared to a predefined rubric for each specific test; but it is here that “the (construct) theory itself is not called into question: it is taken for granted. The issue is whether the test is a successful operationalisation of the theory” (Alderson, Clapham & Wall, 2005: 183). Messick (1996) divides construct validity into two further sub-headings, namely construct under-representation and construct-irrelevant variance. The former suggests that a test is too narrow and omits essential target language and construct. The latter, conversely, describes the test as being too broad, with too many items that are not relevant to the construct.

3.2.2. Criterion-related Validity

Within this category, there are two commonly recognised types of criterion-related validity: concurrent validity and predictive validity. Hughes (1989: 23) states that “concurrent validity is established when the test and the criterion are administered at about the same time”, and Moritoshi (2001: 10) goes on to clarify this further saying that a test has “concurrent validity with another if the two measures yield consistently very similar results, expressed as a high positive correlation co-efficient”. So, if results from one test format are similar to those from a test with a different format, the tests are said to have concurrent validity.

Predictive validity of a test concerns itself with whether or not the test consistently and accurately predicts the candidates' future performance and behaviour. However, instead of "collecting the external measures at the same time as the administration of ... the test, the external measures will only be gathered some time after the test has been given" (Alderson, Clapham & Wall, 2005: 180). Taking this into consideration, Bachman (1990: 254) highlights the problems by stating that the target criterion behaviour that we want to predict is often exceptionally complex and is often dependent upon a large number of factors other than language abilities.

3.2.3. Content Validity

If a test is to be understood as having content validity its content must consist of a representative sample of language structures and skills with which it is meant to be concerned (Hughes, 1989: 22). Good content validity will represent genuine language use; an area which Brown (2001) demonstrates through a non-linguistic yet apt example to illustrate poor content validity in which a tennis competency test evaluates candidates through a 100-yard dash. Oller (1979: 51) highlights this further by asserting that content validity guarantees the candidates "perform tasks which are genuinely the same or fundamentally similar to tasks one normally performs in exhibiting the skill or ability the tests purports to measure".

Bachman (1990: 244) identifies two specific areas of content validity, namely content relevance and content coverage. Content relevance is applicable to not only the language ability being tested but also the test method itself, an area which is frequently ignored (Bachman, 1990: 244). Candidates have to face many different types of test method, including talking to a machine, both individually and surrounded by others, which is how the TOEFL-iBT test is administered (TOEFL Homepage, ETS, 2009), or talking to an examiner, as is the norm for the Cambridge Main Suite exams (Cambridge Centre, 2009). In tests where the examiner is directly involved, the candidate may well be influenced by how the examiner acts, thus affecting the overall performance and score gained.

Content coverage encompasses whether or not the tasks given in the test mirror tasks in the real world, which Bachman (1990: 245) says can be done by collecting multiple tasks from any given domain, which will "determine the extent to which different sets of tasks are equivalent". He does, however, highlight that boundaries of content domains in language testing are never clear-cut (Bachman, 1990: 245). In order to 'prove' content validity, 'experts' need to make judgement "in some systematic way" (Alderson, Clapham & Wall, 2005: 173). Kluitmann (2008: 25) then points out that this can then lead on to the issue of which ‘experts’ are chosen by the test developer and why they might be chosen; this might be because they have been noted though agreeing with each other in the past, or they might be chosen regardless of their opinion. Test developers need feedback and evidence of their validity as quickly as is possible, which in turn can then affect decisions made by the ‘experts’ (Alderson, Clapham & Wall, 2005: 175). However, the content validity of a test, whilst being a necessity, is not an effective way of evaluating the test, due to not actually being able to give any evaluative information about the interpretation of test scores (Bachman, 1990: 247).

3.2.4. Face Validity

Face validity is explained by Brown (1994: 256; 2001: 388, Hughes, 1989: 27) as whether the test, on the face of it, actually appears to test what it is designed to test, from the learner’s perspective. If test takers believe that results gained are accurate, then face validity can be associated more with acceptance than actual validity (Alderson, Clapham & Wall, 2005: 173). As this is more of a reflection of the opinion of non-experts and is a non-scientific method (Hughes, 1989), it is often dismissed by testers as being irrelevant (Alderson, Clapham & Wall, 2005: 172). However, if face validity is not high, then it can be assumed that the test itself will not be successful, and test takers themselves may well not perform as well as they might otherwise, making test validity an important consideration in test use (Bachman, 1990: 289). There are some though who believe that face validity is actually subordinate to other types of test validity (Jafarpur, 1987: 199).

3.2.5. Backwash

Backwash is also an important factor when considering language testing. Owen *et al* (1997: 26) explain that changing either a test or the marking system of that test can have ramifications on how the test subject is taught and how students might approach their learning. Alderson and Wall (1993: 117) describe backwash as being things that “teachers and learners do (that) they would not necessarily otherwise do because of the test”. Backwash can be both negative and positive (Hughes, 1989: 1; Bachman, 1990: 283), which can lead to students having higher motivation with positive backwash whilst negative backwash might lead to narrowing and distortion of the curriculum (Alderson & Wall, 1993) and potential test score pollution, which is defined as being an increase in test scores without an equal improvement in actual ability in the construct that is being tested (Haladnya, Nolan & Haas, 1991). It has been noted, however, that little empirical evidence has ever come to light in support of the theory that tests bear influence on teaching practice and whether or not backwash does in fact exist (Alderson & Wall, 1993).

If we are to believe Heisenberg's Uncertainty Principle (1926), then backwash certainly would exist within the majority of testing situations as it is here that *training* rather than general *education* or *learning* takes place, especially in countries where high-stakes tests are seen to be important (Amrein & Berliner, 2002). By typing 'IELTS' into a search engine, in Korea for example, there is a multitude of links to various sites proclaiming that they can provide the answers needed to obtain high scores within the test (Naver, 2009). They will often have lists of questions from the test which have been added by candidates from their own test experience, so they are not always 100% accurate, thus affecting how future students may prepare if they use these questions as a guideline. So, as long as candidates keep adding these questions to these sites and they are only partially accurate, negative backwash will indeed continue.

4. The Reliability and Validity of IELTS

A criticism which has been raised about language proficiency tests is whether they actually assess the communicative competence of the candidate (Brown, 2001: 387). With many Asian and South Asian countries relying on purely memorization and imitation (Ballard & Clanchey, 1991: 8), doubt begins to creep in as to the overall validity of the IELTS test, especially with some studies showing no correlation between scores gained through the IELTS test and general academic performance (Cotton & Conrow, 1998). Conversely, studies by people such as Bellingham (1993) and Ferguson & White (1993) have shown that there is a positive connection, albeit sometimes weak, between IELTS and students' grade point average (GPA). A small-scale survey about self-evaluation (Bayliss, 2006: 4) carried out in Australia questions whether or not we actually need testing, as the candidates appeared to have very accurate perceptions of their own language skills. The overall mean score of the self-rating was 6.43 compared to a mean IELTS rating of 6.45, and even though this was only small-scale, there could be global implications if candidates could regularly evaluate themselves to this kind of accuracy. However, clearly the purpose of most language exams is not to reaffirm a language learner's own perceptions, rather it is arguably more often for the practical purposes of gaining admission to academic institutions, emigrating or job applications.

4.1. IELTS and Reliability

As reported on the IELTS homepage (2009), test results from 2007 can be split into two groups, with the first being the reading and listening modules, as these are marked objectively; the second group contains the writing and speaking modules, as these are evaluated subjectively and, allegedly, “cannot be reported in the same manner” (IELTS Homepage, 2009). However, the reliability of the listening tests, in 2007, can be seen to be high as the coefficient, as reported on the homepage, stands at 0.89, a figure which Lado (*cited in* Hughes, 1989: 32) deems to be acceptable as a measurement of the consistency and reliability of a test. The reading module, in 2007, does not seem to have fared so well, as both the academic reading, with a coefficient of 0.86, and the general reading, with a coefficient of 0.89, have both fallen below Lado’s (1989: 32) target figure of 0.9 to 0.99. Whilst these figures don’t negate the IELTS reading module’s validity, it is however evident that a higher coefficient would further endorse the validity of the test.

Despite the emphasis being placed on certification, retraining and re-standardization of examiners for the speaking and writing modules, the IELTS homepage doesn’t actually offer any data for the reliability of these modules. They do, however, offer a ‘composite reliability estimate’ which they have based on a theory taken from Feldt & Brennan (1989) which, over the four modules, gives a ‘high’ coefficient of 0.95, in turn producing a ‘low’ SEM of 0.21. Nevertheless, until there is an actual method of objectively measuring the reliability coefficient for the speaking and writing modules, it is difficult to produce a valid coefficient figure that hasn’t potentially been manipulated for the purpose of propaganda and marketing.

4.2. IELTS and Construct Validity

The TOIEC test claims to assess overall communication skills by only testing listening and reading skills, which Messick (1996) would categorise as construct under-representation. Both the TOEFL-iBT and IELTS tests however, cover the four language skills, which, theoretically, would suggest high construct validity, but would not necessarily be considered as construct-irrelevant (Messick, 1996), obviously depending on what topics within the target construct are being tested. Hughes (2003: 31) goes so far as to suggest that defining construct validity is not necessary for direct tests of what are sometimes seen to be common-sense constructs, which he names as reading and writing.

With constant monitoring, evaluation and updating of materials for the IELTS test, (IELTS Homepage, 2009), answering the question posed by Brown (2001: 389) as to whether or not the test does actually tap into the theoretical construct as it has been defined would appear to be simple - the test *does* test the four aspects of English, thus suggesting that it is possible to demonstrate construct validity of IELTS.

4.3. IELTS and Criterion-related Validity

From figures published on the IELTS homepage (2009) detailing results from 2007, there is a high correlation between the reading and listening modules, standing at 0.89, and the reading at 0.88 respectively. These figures, by themselves would support the test's claims of concurrent validity but, Bachman warns (1990: 249), without evidence from an independent source supporting this interpretation of the criterion of the ability being tested, that there is no firm basis for interpreting this criterion as evidence of validity.

Whilst these figures are based solely on the listening and reading modules, making it difficult to comment on the concurrent validity of the test as a whole, studies were also carried out (IELTS Speaking Revision Project, 1998-2001) in order to find the coefficient of the subjectively marked speaking and writing modules. The speaking resulted in having a coefficient of 0.86, and the writing a coefficient of 0.85 – 0.93. With such a variance in the final figure for the writing, it immediately becomes clear that to find a precise coefficient, and with it a reliable measure of concurrent validity, could be a difficult task.

4.4. IELTS and Content Validity

Content validity for IELTS is regarded, by Bachman *et al.*(1995) at least, as being high. This opinion is mirrored by Weir (1990: 7-15) who states that IELTS is a variety of communicative tests because real-life tasks are presented to the candidates. This opinion, however, dates back to 1990, which possibly reduces its value, as the format of IELTS has been revised and updated in that time. However, studies by Farhady (2005) found that, in the listening module at least, candidates taking IELTS preferred being tested on real-life contexts, which again suggests good content validity for the test as a whole.

Initial research into the test as we know it today, called the IELTS Impact Study (IIS) conducted by Hawkey *et al.* (2001) through questionnaires sent to institutions both teaching and testing IELTS, commissioned by Cambridge ESOL and reported in 'Research Notes' (2004), also leans towards high content validity. Teachers and candidates alike thought that the content was relevant to target language activities, but some felt that the writing and some reading tasks were maybe too general. Whilst generating and collating information on content validity is deemed useful, it is however not necessarily a sufficient way of validating a test (O'Sullivan *et al.* 2002: 38), especially when the evidence presented is researched by those responsible for the construction and distribution of the test. However, clearly, this is not an accurate or objective indicator of face validity.

4.5. IELTS and Face Validity

As there is no evident way of quantifying face validity, it could be assumed, due to the popularity of IELTS being used around the world as a guideline to overall English ability , that IELTS has high face validity. Nevertheless, unlike the TOIEC test that only measures candidates' abilities in listening and reading which, in 2005, was dropped by 12 mid-ranking corporations as a requirement for jobs (Chosun Newspaper, 2005), IELTS has continued to be adopted by many companies and institutions, as detailed in section 2 of this paper, as a guideline of candidates' level of ability in English. Criper and Davies (1988: 99) state that face validity is high due to the modular approach of IELTS which is very popular among subject specialists. Despite this, the test itself has evolved considerably and further research into this area could well prove to be a fruitful exercise.

4.6. IELTS and Backwash

With teaching and learning apparently being affected by backwash (Hughes, 2003: 1) but with its existence also being questioned, (Alderson & Wall, 1993), it is difficult, without empirical evidence, to state whether or not backwash has a major influence on IELTS or those involved with the teaching or studying of it. Backwash does however become important when considering if IELTS ever becomes a major predictor of language ability in tests such as exit tests from universities, as detailed by Qian (2007: 33). His research showed that the ultimate goal in implementing an exit test is to demonstrate the importance of proficiency in English, which would result from enhanced teaching and learning activities, in turn resulting, largely, from positive backwash. He also touches on the negative backwash effect of IELTS as there are "a number of discrete-point item types, such as multiple choice and matching, which may cause negative impact on teaching and learning, as such formats allow for too much guessing".

Backwash is similar to face validity in that further research would be beneficial if effective, conclusive statements are to be drawn as, currently, there would appear to be very little empirical data on the subject.

5. Improving the Global Validity of IELTS

In order to establish overall validity, extensive further research would need to be done. However, with half of the modules of the test being marked objectively, namely the writing and listening, and the remainder being marked subjectively by trained examiners, namely the speaking and writing, it may well be difficult to come up with a satisfactory SEM for the test, which in turn may well hinder the search for conclusive evidence supporting the varying sub-headings for validity. Continual development and modification of the test has, over the years since its conception, made the test a strong contender in the field of language testing around the world (IELTS Homepage, 2009) and this presumably will be an ongoing process in the future. With more institutions and companies around the world specifying IELTS as being important in the recruitment process, more research also needs to be done to confirm the areas of validity which are presently difficult to quantify, such as predictive validity and face validity. In order to prevent exposure to criticisms of bias, any further research would need to be conducted by external bodies.

6. Conclusion

The International English Language Testing System would seem to be a relatively reliable measure of language proficiency. Unlike other tests, such as the TOEIC which only tests listening and reading, IELTS is a comprehensive test with high content validity which some might consider important if trying to assess real-life language proficiency. In order to establish greater reliability and validity, more independent research is needed if the test is to continue to effectively measure overall proficiency in English.

Testing of English, as a whole, is becoming an increasingly contentious issue around the world as, through globalisation, English is still seen by the majority to be the lingua franca. In this market, IELTS would, at face value at least, appear to be the most comprehensive test of an overall proficiency in English but, with the constant emergence of new tests in the same field (Arita, 2003), this may not continue to be the case without continual development and modification.

References

Alderson, C., Clapham, C. & **Wall, D.** (2005) *Language Test Construction and Evaluation.* 9th ed. Cambridge: Cambridge University Press.

Alderson, J.C. & **Wall, D.** (1993). Does washback exist? *Applied Linguistics, 14*, pp. 115-129.

Amrein, L. & **Berliner, D.** (2002) High-Stakes Testing, Uncertainty, and Student Learning. In *Education Policy Analysis Archives,* 10, 18.

Arita, E. (2003) New tests challenging TOEIC stronghold. In *Japan Times Online.* http://search.japantimes.co.jp/cgi-bin/nn20031025b3.html Accessed 6th January 2009

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

Bachman, L.F., Davidson, F., Ryan, K. & **Choi, I-C** (1995) *Studies in language testing 1: An investigation into the comparability of two tests of English as a foreign language.* Cambridge: Cambridge University Press.

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

Ballard, B. & **Clanchey, J.** (1991) Assessment by Misconception: Cultural Influences and Intellectual Traditions. In **Hamp-Lyons, L.** (Ed.) *Assessing Second Language Writing In Academic Contexts.* pp. 19-36. Norwood, N.J.: Ablex Publishing Corporation.

Bayliss, A. & **Ingram, D.** (2006) IELTS as a Predictor of Academic Language Performance. *Australian International Education Conference 2006.* http://www.aiec.idp.com/PDF/BaylissIngram%20(Paper)%20Wed%201630%20MR5.pdf Accessed 6th January 2009

Bellingham, L. (1993) The relationship of Language Proficiency to Academic Success for International Students. *New Zealand Journal of Educational Studies,* 30 (2), pp. 229-232.

Brown, H.D. (1994) *Principles of Language Learning.* 3rd ed. White Plains, New York: Longman.

Brown, H.D. (2001) *Principles of Language Learning.* 4th ed. White Plains, New York: Longman.

Cambridge Centre http://cambridgecentre.net/main_suite.html Accessed 6th February 2009.

Chauncey Group International Ltd. (1999) *TOEIC® User Guide.* Chauncey Group International

Chosun Newspaper (2005) The End of the Road for TOIEC. http://english.chosun.com/w21data/html/news/200512/200512040016.html Accessed 4th January 2009

Cotton, F. & **Conrow, F.** (1998) An Investigation of the Predictive Validity of IELTS amongst a Group of International Students studying at the University of Tasmania. *English Language Testing System Research Reports,* 1, pp.72-115.

Criper, C. & **Davies, A.** (1988) *ELTS Validation Project Report: Research Report 1(i).* The British Council/University of Cambridge Local Examinations Syndicate

ETS http://www.ets.org/portal/site/ets/menuitem Accessed 6th January 2009.

ETS http://www.ets.org/toefl Accessed 6th February 2009.

Farhady, H. (2005) The effect of coaching TOEFL type and task based tests. *Pazhuhesh-e Zaban-haye Khareji Journal,* 20, pp.1-10.

Feldt, L.S. & **Brennan, R.L.** (1989) Reliabilty. In **Linn** (Ed.), *Educational Measurement* 3rd ed. American Council on Education: Macmillan

Ferguson, G. & **White, E.** (1993) A small-scale study of predictive validity. *Melbourne Papers in Language Testing,* University of Edinburgh, pp. 15-63.

Gilfert, S. (1996) A Review of TOEIC. In *The Internet TESL Journal.* http://iteslj.org/Articles/Gilfert-TOIEC.html Accessed 6th January 2009.

Haladnya, T.M., **Nolan S.B**. & **Haas, N.S.** (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher 20,* pp.2-20.

Hawkey, R. (2004) An IELTS Impact Study: implementation and some early findings. 15. In *Research Notes.* http://wwwCambridgeESOL.org/rs_notes Accessed 6th January 2009.

Hughes, A. (1989) *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Heisenberg. **W.** (1930), *Physikalische Prinzipien der Quantentheorie* (Leipzig: Hirzel). English translation *The Physical Principles of Quantum Theory* . Chicago: University of Chicago Press.

Hughes, A. (2003) *Testing for Language Teachers* (2nd Ed.). Cambridge: Cambridge University Press.

IELTS http://www.ielts.org/general_pages/media_centre/ielts_tests_over_one_million.aspx Accessed 27th October 2008.

Jafarpur, A. (1987) 'The short-context technique: an alternative for testing reading comprehension.' *Language Testing* 4 (2), pp. 195-220.

Kluitmann, S. (2008) Testing English as a Foreign Language: Two EFL-Tests Used in Germany. MA Thesis, University of Albert-Ludwig.

Lado, R. (1961) *Language Testing.* New York: McGraw-Hill.

Messick, S. (1996) Validity and washback in language testing. *Language Testing, 13* (2), pp. 241-256.

Moritoshi, P. (2003) *The Test o f English for International Communication (TOEIC): necessity, proficiency levels, test score utilization and accuracy*. University of Birmingham. http://www.cels.bham.ac.uk/resources/Essays.htm Accessed 27th October 2008.

Naver http://search.naver.com/search.naver?where=nexearch&query=ielts&sm=top_hty&fbm=0 Accessed 26th January 2009.

Oller, J.W. (1979) *Language Tests at School.* London: Longman.

Owen, C. *et al* (1997) *Testing.* Centre for English Language Studies, Birmingham University.

Qian, D. (2007) Assessing University Students: Searching for an English Language Exit Test. In *RELC Journal*; 38; 18. pp. 18-37.

Saville, N. & **Hawkey, R.** (2003) A study of the impact of the International English Language Testing System, with special reference to its washback on classroom materials. In **Cheng, L., Curtis, A.** & **Watanabe, Y.** (Eds.) *Concept and method in washback studies: the influence of language testing on teaching and learning.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Saville, N. & **Hawkey, R.** (2004) The IELTS Impact Study: Investigating washback on teaching materials. In **Cheng, L., Curtis, A.** & **Watanabe, Y.** (Eds.) *Washback in language testing: Research contexts and methods.* Mahwah, N.J.: London: Lawrence Erlbaum

O'Sullivan, B., Saville, N. & **Weir, C.** (2002) Using observation checklists to validate speaking-test tasks. In *Language Testing,* 19, 33. pp. 33-56

Sharron, R.H. (1997) 'TOEIC® today and TOEIC® tomorrow.' In TOEIC® Steering Committee (ed.) *The 58th TOEIC® Seminar.* Educational Testing Service.

Vancouver English Centre http://secure.vec.bc.ca/toefl-equivalency-table.cfm Accessed 24th October 2008.

Weir, C. (1990) *Communicative Language Testing.* New York: Prentice Hall.

Appendices

Appendix 1

A. Quite frankly I have absolutely no confidence in the British Council. We still seem to get people who can't write a page of English without littering it with errors of every description. What's it called? Their test – you know – the thing they do before they get here, and get 6.5 on or whatever?

B. IELTS

A. That's the one. If you ask me, it's a total waste of time.

B. Well, I suppose you need a test of one sort, don't you, or you might end up with even more problems.

A. I've been thinking about that actually, and I'm not so sure. Look at it this way. Supposing you saw a course advertised in Germany, or Japan – I don't know – name the country of your choice – a course you really want to do because it isn't available here. Now, I don't know how good your German or Japanese is, but in my case, I know perfectly well that I would struggle a bit in German and wouldn't even get off the ground in Japanese; so I would probably have to go and improve my German for a few months, and wouldn't even be able to consider a course in Japan unless I was willing first to put in some really intensive language study for a year or two, possibly longer.

B. So?

A. Well the point is I know these things about myself without the Goethe Institut or the Japanese equivalent of the British Council telling me. And if I know them, I can't really see why people who want to come to Britain don't know them; they aren't stupid. Why do people need an elaborate test to tell them what they know already?

B. OK, but the university needs to know, or you'll end up with a whole load of incompetents clogging up the system.

A. But you won't get any more than you've got now. People just won't come if they know they can't do their course, any more than you or I would be foolish enough to sign up for an MA course in Japanese.

B. I’m afraid they would you know. I don’t think people do have such a good idea of whether their language proficiency is up to following a course. You may have, because you are an experienced lecturer, but the average overseas student probably hasn’t got a clue what he’s in for when he gets on the plane to Birmingham. You can’t just leave it to self-assessment, especially when you think it’s costing thousands of pounds to send these guys over here. As you said yourself, enough of them slip through the net as it is.

A. I reckon the number of people who apply for postgraduate courses at British universities, and who are turned down on flat grounds of inadequate English is probably very small. Either they get rejected on other grounds, or they just don’t apply. The test probably doesn’t alter decisions on acceptance in more than a tiny number of cases. What’s more, there are plenty of students who haven’t reached the official admissions requirement, 6.5 or whatever, and who get admitted anyway because the university can’t afford to turn them away.

B. Ah, there you may well have a point. But it doesn’t change the principle of the thing.